Written examination: 26. May 2019

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

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

Exercise	IX.1	IX.2	IX.3	IX.4	X.1	X.2	X.3	X.4	X.5	X.6
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	1	3	5	5	3	4	4	3	3	1

The exam paper contains 35 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 a cola tasting experiment there are 4 glasses with cola. Each glass contains either regular cola or light cola. You know that there are two glasses of each. A taster randomly chooses two glasses.

Question I.1 (1)

What is the probability that she gets regular cola in one of the glasses and light cola in the other?

1 □ 1/4

2 □ 1/3

3 □ 1/2

4* □ 2/3

5 □ 3/4

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

It is drawing without replacement, i.e. the hypergeometric distribution:

```
## We use
dhyper(x=1, m=2, n=2, k=2)

## [1] 0.6666667

## Or we can do it another way and calculate it directly. The probability of getting
(1/2*1/3)

## [1] 0.1666667

## thus the probability of getting two times cola zero is the same.
(1/2*1/3)

## [1] 0.1666667

## Hence not getting either 2 cola or 2 zeros (and thus getting one of each) is
1 - 2 * (1/2*1/3)

## [1] 0.6666667
```

------------------------------------ FACIT-END ------------------------------------

Question I.2 (2)

In another experiment, a glass of regular cola and a glass of light cola are given to each of 25 tasters. They are told to taste and answer if they think that there is a difference between the cola in the glasses. The answers are independent of each other.

From experience, one knows that it can be assumed that there is $p = 0.8$ probability that a taster can taste the difference between regular and light. Let X denote the number of the 25 tasters who say there is a difference. What will be the variance of X?

1 □ $\text{V}(X) = 5$

2* □ $\text{V}(X) = 4$

3 □ $\text{V}(X) = 3$

4 □ $\text{V}(X) = 2$

5 □ $\text{V}(X) = 1$

----------------------------------- FACIT-BEGIN -----------------------------------

In this setup it is “drawing” with replacement, hence X follows a binomial distribution

$$X \sim B(n = 25, p = 0.8)$$

and we can use Theorem 2.21 to find the variance

```
n <- 25
p <- 0.8
n*p*(1-p)

## [1] 4
```

------------------------------------ FACIT-END ------------------------------------

Continue on page 4

Exercise II

10 women measured their morning temperature on both July 1st and December 1st. From the measurements, one would like to investigate whether there is a difference in the morning temperature for women in the summer compared to the winter. It can be assumed that the summer measurements are normally distributed and that the winter measurements are normally distributed.

Question II.1 (3)

Which analysis will be most appropriate?

1 □ Test for the difference between two proportions

2 □ Regression analysis

3 □ (Un-paired) t-test

4* □ Paired t-test

5 □ Test in the binomial distribution

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

This is a paired setup, since for each individual there exists two observations of the same variable at two different times. Hence the two samples of each 10 temperatures can be paired via the women.

```
------------------------------------ FACIT-END ------------------------------------
```

Question II.2 (4)

When the test was carried out a p-value of 0.4 was obtained. This means that:

1 □ There is a 40% probability that there is a difference between the morning temperature in the summer compared to the winter.

2 □ There is a 0.4% probability that there is a difference between the morning temperature in the summer compared to the winter.

3 □ The hypothesis cannot be tested.

4 □ There is definitely a difference between the morning temperature in the summer compared to the winter.

5* □ Under the null hypothesis, the probability of obtaining a value of the test statistic which is less extreme, than the value obtained, is 0.6.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

This derived from the definition of the p-value in Definition 3.22.

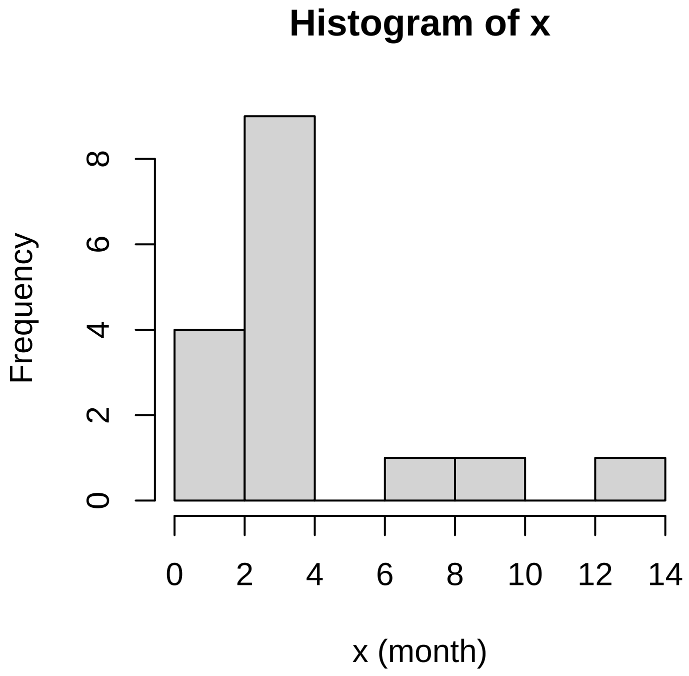
```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 6

Exercise III

A company has purchased a new 3D printer technology and they want to investigate whether it can be used to make components that are durable enough to be included in a specific product.

An experiment has been carried out where components, printed with the new technology, have been used in a batch of test products. These products have then been subjected to a test that determines their lifetime. It is assumed that the lifetime follows an exponential distribution, so let $X \sim Exp(\lambda)$ denote the lifetime in months. A sample has been collected for $n = 16$ products. A histogram of the sample is:



The observed life times has been saved in the vector `x` and the following R code is run:

```
## Number of simulations
k <- 10000
nx <- length(x)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
## Calculate the sample mean
simmeans <- apply(simxsamples, 2, mean)
## Quantiles of the means
quantile(simmeans, c(0.005,0.995))

##  0.5% 99.5%
##  1.70  6.42

quantile(simmeans, c(0.025,0.975))
```

```
##  2.5% 97.5%
##  2.07  5.68

quantile(simmeans, c(0.05,0.95))

##   5%  95%
## 2.26 5.26
```

Question III.1 (5)

It was pre-planned to investigate whether it can be shown, at significance level $\alpha = 1\%$, that the average lifetime m_X is over 2 months for the components.

Can this be concluded on the basis of the collected sample and the calculations above (both conclusion and argument must be correct)?

1* □ Since 2 is contained in the calculated 99% confidence interval it cannot be concluded.

2 □ Since 2 is not contained in the calculated 99% confidence interval it can be concluded.

3 □ Since 2 is contained in the calculated 95% confidence interval it cannot be concluded.

4 □ Since 2 is not contained in the calculated 95% confidence interval it can be concluded.

5 □ With the given information it is not possible to answer this question.

------------------------------------ FACIT-BEGIN ------------------------------------

To find the 99% confidence interval we find the 0.005 and 0.995 quantiles of the simulated means (the first of the three quantile calculations above). Since 2 is contained in this interval which is [1.70, 6.42], we cannot conclude with $\alpha = 1\%$ that the true average lifetime is not 2.

------------------------------------- FACIT-END -------------------------------------

Question III.2 (6)

What is the sample mean of the collected sample?

1 □ $\bar{x} = 3.40$

2 □ $\bar{x} = 3.76$

3 □ $\bar{x} = 3.875$

4 □ $\bar{x} = 4.06$

5* □ With the given information it is not possible to answer this question.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

One approach is to take the value between the confidence intervals. But for this to be the mean, we need to assume that the means are normally distributed (so the quantiles of the means are symmetrical), and this is only true if nx is big enough so it satisfies the central limit theorem and we don't know if this is the case. Therefore we cannot find the mean of the sample.

```
------------------------------------ FACIT-END ------------------------------------
```

Question III.3 (7)

A new sample of lifetimes has been collected where a new material has been used to print the components. They are subsequently subjected to the same tests and the observed lifetimes are stored in the vector y. There are $n_Y = 17$ observations in the new sample.

The following R code is run afterwards:

```
## Number of simulations
k <- 10000
nx <- length(x)
ny <- length(y)
## Simulate k times
simxsamples <- replicate(k, rexp(nx, 1/mean(x)))
simysamples <- replicate(k, rexp(ny, 1/mean(y)))
## Calculate the simulated statistics
simdifmeans <- apply(simysamples, 2, mean) - apply(simxsamples, 2, mean)
simdifmedians <- apply(simysamples, 2, median) - apply(simxsamples, 2, median)
## Quantiles of the simulated statistics
quantile(simdifmeans, c(0.025,0.975))

##  2.5% 97.5%
## 0.733 9.443

quantile(simdifmeans, c(0.05,0.95))

##   5%  95%
## 1.30 8.59

quantile(simdifmedians, c(0.025,0.975))

##   2.5%  97.5%
## -0.428  8.265

quantile(simdifmedians, c(0.05,0.95))

##     5%    95%
## 0.0837 7.3868
```

Which of the following conclusions can be drawn on the basis of these calculations?

1 □ At $\alpha = 5\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

2* □ At $\alpha = 10\%$ significance level it can be concluded that the 50% quantile of the product lifetime is higher with components of the new material.

3 □ At $\alpha = 5\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

4 □ At $\alpha = 10\%$ significance level it can be concluded that there is at least 50% probability that the product lifetime is higher with components of the new material.

5 □ With the given information no conclusions can be drawn.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The 50% quantile is the same as the median. Therefore we are using the simdifmedians.From the two last calculations in the R-code it can be seen that the 95% confidence overlaps with 0, but the 90% confidence interval does not. Therefore, with $\alpha = 5\%$, we cannot conclude that there is a difference between the medians, but at $\alpha = 10\%$ we can, so answer 2 is correct.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 10

Exercise IV

Assume that X is normally distributed with mean 10 and variance 4, Y is normally distributed with mean 20 and variance 25, and X and Y are independent.

Question IV.1 (8)

Then $2Y - 2X + 4$ has the variance:

1 □ 36

2 □ 58

3 □ 84

4* □ 116

5 □ None of the values above.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

Use the variance identities in Theorem 2.54 and 2.56 to get

$$\mathrm{V}(2Y - 2X + 4) = 4\,\mathrm{V}(Y) + 4\,\mathrm{V}(X) = 4 \cdot 4 + 4 \cdot 25 = 116.$$

Or you can simulate it:

```
k <- 100000
x <- rnorm(k, 10, sqrt(4))
y <- rnorm(k, 20, sqrt(25))
z <- 2*y - 2*x + 4
var(z)

## [1] 116.0609
```

```
------------------------------------ FACIT-END ------------------------------------
```

Question IV.2 (9)

What is the standard deviation of $f(X,Y) = 2Y^2 + X^3/3$ (tip: if you solve this using simulation, remember to have many repetitions and choose the answer with the result being approx. ± 10 from the stated number in the answer)?

1 □ $\sigma_{f(X,Y)} \approx 100$

2 □ $\sigma_{f(X,Y)} \approx 250$

3 □ $\sigma_{f(X,Y)} \approx 350$

4* □ $\sigma_{f(X,Y)} \approx 450$

5 □ $\sigma_{f(X,Y)} \approx 5 \cdot 10^4$

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

Solve it using simulation (as presented in beginning of Chapter 4)

```
k <- 1000000
x <- rnorm(k, 10, sqrt(4))
y <- rnorm(k, 20, sqrt(25))
sd(2*y^2 + x^3/3)

## [1] 460.3031
```

or use the linear approximation with the error propagation formula in Method 4.3

$$\begin{aligned}\sigma^2_{f(X,Y)} &= (\frac{\partial f}{\partial x})^2\sigma^2_x + (\frac{\partial f}{\partial y})^2\sigma^2_y \\ &= (3X^2/3)^2 \cdot 4 + (4Y)^2 \cdot 25 \\ &= X^4 \cdot 4 + 16Y^2 \cdot 25 \\ &= 10^4 \cdot 4 + 16 \cdot 20^2 \cdot 25\end{aligned}$$

```
  sqrt(10^4 * 4 + 16 * 20^2 * 25)

## [1] 447.2136
```

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 12

Exercise V

The association between pressure (p) and depth (h) in an open liquid container may be described theoretically by the equation

$$p = p_0 + \rho g h\,,$$

where p_0 is atmospheric pressure, ρ is the density of the liquid, and g is the acceleration due to gravity. An experiment was conducted with the purpose of determining the density of a special liquid. 10 measurements of depth (in m) and pressure (in Pa) were conducted in this liquid, and the results were assigned to two vectors in R, `depth` and `pressure`, respectively. Furthermore, the following R code was run:

```r
model1 <- lm(pressure ~ depth)
summary(model1)

##
## Call:
## lm(formula = pressure ~ depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119166  -73422   30513   53635  124689
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.019e+08  5.867e+04 1737.529  < 2e-16 ***
## depth       5.031e+03  9.455e+02    5.321 0.000711 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85880 on 8 degrees of freedom
## Multiple R-squared:  0.7797,Adjusted R-squared:  0.7521
## F-statistic: 28.31 on 1 and 8 DF,  p-value: 0.0007105
```

Question V.1 (10)

Give the estimate of the atmospheric pressure during the experiment:

1 □ $5.031 \cdot 10^3$ Pa

2 □ $5.867 \cdot 10^4$ Pa

3 □ $9.455 \cdot 10^7$ Pa

4* □ $1.019 \cdot 10^8$ Pa

5 □ $1.025 \cdot 10^8$ Pa

```
------------------------------------ FACIT-BEGIN -----------------------------------
```

Estimated atmospheric pressure corresponds to the model intercept. It can be seen from the theoretical equation that the atmospheric pressure is the bias term corresponding to our normal β_0.

```
------------------------------------- FACIT-END ------------------------------------
```

Question V.2 (11)

One would like to test the hypothesis that the expected atmospheric pressure is $1.005 \cdot 10^8$ Pa under the experimental conditions. Give the usual test statistic used to test this hypothesis:

1 □ $t_{\text{obs}} = 1738$

2 □ $t_{\text{obs}} = 5.321$

3* □ $t_{\text{obs}} = 23.86$

4 □ $t_{\text{obs}} = 28.31$

5 □ $t_{\text{obs}} = 0.000711$

```
------------------------------------ FACIT-BEGIN -----------------------------------
```

Method 5.14 for the intercept parameter β_0 with $\hat{\beta}_0 = 1.019 \cdot 10^8$, $\beta_{0,0} = 1.005 \cdot 10^8$ and $\hat{\sigma}_{\beta_0} = 5.867 \cdot 10^4$. Then

$$t_{\text{obs}} = \frac{1.019 \cdot 10^8 - 1.005 \cdot 10^8}{5.867 \cdot 10^4} = 23.86$$

------------------------------------ FACIT-END ------------------------------------

Question V.3 (12)

Give a 95% confidence interval for the parameter which describes the association between depth and pressure:

1 □ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880/(10-2)$

2 □ $1.019 \cdot 10^8 \pm 2.306 \cdot 85880$

3 □ $5031 \pm 2.306 \cdot 85880$

4 □ $1.019 \cdot 10^8 \pm 2.306 \cdot 5.867 \cdot 10^4$

5* □ $5031 \pm 2.306 \cdot 945.5$

----------------------------------- FACIT-BEGIN -----------------------------------

Method 5.15 with $\hat{\beta}_1 = 5031$, $\hat{\sigma}_{\beta_1} = 945.5$ (both from the R-output) and $t_{0.975}$ found using

```
qt(0.975, df = 10-2)

## [1] 2.306004
```

------------------------------------ FACIT-END ------------------------------------

Question V.4 (13)

Give an estimate of the density of the liquid during the experiment, when the acceleration due to gravity, g, is 9.82 N/kg:

1* □ 512 kg/m^3

2 □ 1004 kg/m^3

3 □ 307 kg/m^3

4 □ 802 kg/m^3

5 □ 610 kg/m^3

----------------------------------- FACIT-BEGIN -----------------------------------

The model slope is $\beta_1 = \rho g$, so

$$\hat{\rho} = \frac{\hat{\beta}_1}{g} = \frac{5031}{9.82} = 512$$

----------------------------------- FACIT-END -----------------------------------

Continue on page 16

Exercise VI

A sample was taken with independent observations from a normally distributed population. One would like to test the hypothesis that the mean is zero against the alternative, that it is different from zero. The test statistic for the test follows a t-distribution. A p-value of 0.001 was obtained.

Question VI.1 (14)

What is then known about the 99% confidence interval for the mean?

1 □ It contains zero.

2* □ It does not contain zero.

3 □ It contains zero, but not the estimate of the mean.

4 □ There is not enough information to know anything specific about the confidence interval.

5 □ It contains 0.01.

```
------------------------------------ FACIT-BEGIN -----------------------------------
```

It is a one-sample t-test. We use Theorem 3.33 and we can see that testing the null hypothesis $H_0 : \mu = 0$ and rejecting it on a significance level $\alpha = 0.01$ means that the 0 is not contained in the 99% confidence interval.

```
------------------------------------ FACIT-END -------------------------------------
```

Question VI.2 (15)

If there were $n = 20$ observations in the sample, what do we then know about the observed test statistic?

1 □ $t_{\text{obs}} = -1.33$ or $t_{\text{obs}} = 1.33$

2 □ $t_{\text{obs}} = -1.73$ or $t_{\text{obs}} = 1.73$

3 □ $t_{\text{obs}} = -3.55$ or $t_{\text{obs}} = 3.55$

4 □ $t_{\text{obs}} = -3.58$ or $t_{\text{obs}} = 3.58$

5* □ $t_{\text{obs}} = -3.88$ or $t_{\text{obs}} = 3.88$

----------------------------------- FACIT-BEGIN -----------------------------------

It is a one-sample t-test and since had have a p-value of 0.001, hence

$$\begin{aligned}2 \cdot P(T > t_{\text{obs}}) &= 0.001 \Leftrightarrow \\ P(T > t_{\text{obs}}) &= 0.001/2 = 0.0005\end{aligned}$$

we need to find the 0.05% or 99.95% quantile. So

```
qt(0.001/2, df=19)

## [1] -3.883406
```

or

```
qt(1 - 0.001/2, df=19)

## [1] 3.883406
```

------------------------------------ FACIT-END ------------------------------------

Continue on page 18

Exercise VII

The Danish Veterinary and Food Administration wants to reduce the proportion of resistant bacteria in pigs intestinal flora, as they pose a human risk. qPCR is one microbiological method to count the number of specific genes in a faeces sample. Below is the count of three genes: 16S, which is a reference gene, and two genes that encode resistance to tetracycline (tetO and tetM). Four samples were taken at different times (first Sample 1, then 2, 3 and finally 4) on the same farm and the researchers want to investigate whether changes have occurred.

	16S	tetO	tetM	Sum
Sample 1	4675	171	76	4922
Sample 2	2222	95	1	2318
Sample 3	2750	49	2	2801
Sample 4	2040	47	1	2088
Sum	11687	362	80	12129

A χ^2-test should be carried out to determine if the proportion of resistant genes has changed over time.

Question VII.1 (16)

The degrees of freedom in this test is:

1 □ 8

2 □ 12

3* □ 6

4 □ 9

5 □ It doesn't make sense to do a χ^2-test, when two of the observations are 1.

----------------------------------- FACIT-BEGIN -----------------------------------

It is the χ^2-test in Method 7.22 which is used. The degrees of freedom is:

$$(r-1)(c-1) = 3 * 2 = 6$$

------------------------------------ FACIT-END ------------------------------------

Question VII.2 (17)

Under the null hypothesis what is the expected number of tetM copies in Sample 4?

1 □ 20

2 □ 1

3* □ 13.77

4 □ 26.10

5 □ 696

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The expected number in the cell is found as in equation 7-53 by

$$\frac{columntotal * rowtotal}{grandtotal} = 80 * 2088/12129 = 13.77$$

```
------------------------------------ FACIT-END ------------------------------------
```

Question VII.3 (18)

The test statistic turns out to be 132.3. The relevant p-value is found using which of the following calls in R?

1 □ `1 - dchisq(132.3, df=6)`

2* □ `1 - pchisq(132.3, df=6)`

3 □ `qchisq(132.3, df=6)`

4 □ `pchisq(132.3, df=6)`

5 □ `qchisq(1/132.3, df=6)`

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

See Method 7.22. We need to find $P(\chi^2 > \chi^2_{\text{obs}})$, which can by done either by

```
1 - pchisq(132.3, df=6)

## [1] 0
```

which, since the value is below the machine precision is zero, actually by

```
pchisq(132.3, df=6, lower.tail = FALSE)

## [1] 4.212873e-26
```

it can be calculated, but in practice it so small that is doesn't make much difference, but maybe nicer to be led to think that it is exactly zero.

`------------------------------------ FACIT-END ------------------------------------`

Question VII.4 (19)

It has previously been planned to investigate whether the occurrence of tetO has changed between the first sample and fourth sample. For reasons not explained here, the observations of tetM should not be considered in this test. The following code has been run with the associated code output:

```
prop.test(x=c(171, 47), n=c(4675+171, 2040+47), correct=FALSE, conf.level=0.95)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(171, 47) out of c(4675 + 171, 2040 + 47)
## X-squared = 7.8067, df = 1, p-value = 0.005205
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.004550394 0.020982546
## sample estimates:
##     prop 1     prop 2
## 0.03528683 0.02252036
```

The usual $\alpha = 0.05$ significance level is used. What is the conclusion (both the conclusion and the argumentation must be correct)?

1 □ No significant change has been detected, since $0.02098 < 0.02252$.

2 □ A significant change has been detected, since $0.0052 < 0.95$, but it is not possible to conclude if the occurrence has increased or decreased.

3 □ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has increased.

4 □ A significant change has been detected, since $0.0052 < 0.95$, and the occurrence of tetO has increased.

5* □ A significant change has been detected, since $0.0052 < 0.05$, and the occurrence of tetO has decreased.

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

We compare the p-value (0.0052) with the significance level of 0.05. Since $P - value < \alpha$ there is a significant change. If we look at the sample estimates we see that the estimate has changed from 0.035 in the first sample to 0.023 in the fourth sample. So answer is correct.

```
------------------------------------ FACIT-END ------------------------------------
```

Continue on page 22

Exercise VIII

The IQ of a randomly selected individual is modeled by a normally distributed random variable. 50% of the population have an IQ over 100 (and 50% have an IQ below 100). Suppose 68% of the population have an IQ in the range of 85-115.

Question VIII.1 (20)

What percentage of the population have an IQ of at least 140 and is thus considered geniuses according to this model?

1 □ 0.01%

2 □ 1%

3 □ 4%

4* □ 0.4%

5 □ 0.06%

------------------------------------ FACIT-BEGIN ------------------------------------

Use need to use the standardized normal distribution. Theorem 2.43. First we find out how many percent have $IQ < 85$:

```
## The quantile at 85
50-68/2

## [1] 16
```

We find out that this is 16% of the population. We now need to find the 0.16 quantile in the std. normal distribution:

```
## The 0.16 quantile in the std. norm.
qnorm(0.16)

## [1] -0.9944579
```

We can then use this in the equation to find the standard deviation of the IQ distribution:

$$Z = \frac{X - \mu}{\sigma} \iff \sigma = \frac{X - \mu}{Z} = \frac{85 - 100}{-0.994}$$

```
## The standard deviation in the IQ distribution
sigma <- (85-100) / qnorm(0.16)
sigma

## [1] 15.0836
```

And we now have the mean and the standard deviation of the IQ distribution and like usual we can find the proportion of people with $IQ > 140$ as:

```
## The proportion of geniueses
1 - pnorm(140, mean=100, sd=sigma)

## [1] 0.004002158
```

------------------------------------ FACIT-END ------------------------------------

Continue on page 25

Exercise IX

The data below have been collected from two groups:

Group 1: 10.5, 9.3, 10.7, 10.8, 11.2

Group 2: 8.9, 9.5, 10.2, 9.8, 10.3

All measurements are assumed to be taken independent. The Group 1 measurements are believed to originate from a normal distribution, and the measurements in Group 2 are assumed to originate from a normal distribution. In addition, it is assumed that the variances in the two normal distributions are identical.

Question IX.1 (21)

What is the sample mean of the Group 2 sample?

1* □ 9.74

2 □ 9.8

3 □ 10.2

4 □ 10.31

5 □ 48.5

----------------------------------- FACIT-BEGIN -----------------------------------

Simply calculate the sample mean of the values from Group 2 in R:

```
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
mean(y)

## [1] 9.74
```

------------------------------------ FACIT-END ------------------------------------

Question IX.2 (22)

What will be the numerical value of the test statistic for the usual test of the hypothesis that there is no difference in mean of the two groups?

1 □ 0.8

2 □ 1.04

3* □ 1.86

4 □ 2.19

5 □ 2.55

------------------------------------ FACIT-BEGIN ------------------------------------

This is a two-sample t-test, hence either use the formulas in Method 3.47 or maybe easier do it in R by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
## Same result of tobs if the variance is pooled
t.test(x, y)

## 
##  Welch Two Sample t-test
## 
## data:  x and y
## t = 1.8564, df = 7.601, p-value = 0.1024
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1927507  1.7127507
## sample estimates:
## mean of x mean of y 
##     10.50      9.74

# Or we can specify the variances as being equal (in which case the pooled variance w
t.test(x, y, var.equal = TRUE)

## 
##  Two Sample t-test
## 
## data:  x and y
## t = 1.8564, df = 8, p-value = 0.1005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1840545  1.7040545
## sample estimates:
## mean of x mean of y 
##     10.50      9.74
```

------------------------------------- FACIT-END -------------------------------------

Question IX.3 (23)

What is the 90% confidence interval for the mean in Group 1?

1 □ [9.61, 11.39]

2 □ [9.32, 11.68]

3 □ [8.92, 12.03]

4 □ [9.87, 12.03]

5* □ None of the intervals above are correct.

------------------------------------ FACIT-BEGIN ------------------------------------

Use the formula in Method 3.9 or do the calculation in R by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
t.test(x, conf.level=0.9)

##
##  One Sample t-test
##
## data:  x
## t = 32.717, df = 4, p-value = 5.204e-06
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##   9.815813 11.184187
## sample estimates:
## mean of x
##      10.5
```

------------------------------------ FACIT-END ------------------------------------

Question IX.4 (24)

A new experiment must be designed in order to achieve a greater power of the statistical test for the mean values. There is still an equal number of observations in each group. The researchers want to have 99% power to discover a difference in mean of at least 1 between the two groups, at significance level 1%. As a guess of the variance, the pooled variance estimate from the two samples are used.

What is the minimum number of observations needed from each group in order for the above requirements to be fulfilled?

1 □ At least 4

2 □ At least 6

3 □ At least 12

4 □ At least 18

5* □ At least 22

----------------------------------- FACIT-BEGIN -----------------------------------

First use Method 3.52 as done in Example 2.85 to calculate the pooled variance estimate by:

```
x <- c(10.5, 9.3, 10.7, 10.8, 11.2)
y <- c(8.9, 9.5, 10.2, 9.8, 10.3)
n1 <- length(x)
n2 <- length(y)
varp <- ((n1-1)*var(x)+(n2-1)*var(y)) / (n1+n2-2)
```

Then use this to calculate the needed sample size by inserting the 4 out of 5 needed values and the R calculate the sample size, see Section 3.3.3:

```
power.t.test(delta=1, sd=sqrt(varp), sig.level=0.01, power=0.99)

##
##      Two-sample t test power calculation
##
##               n = 21.87928
##           delta = 1
##              sd = 0.6473021
##       sig.level = 0.01
##           power = 0.99
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

------------------------------------ FACIT-END ------------------------------------

Continue on page 29

Exercise X

How much clothes a person wears (the clothing level) has a large influence on the level of comfort in offices. In the table below the average clothing level (on a scale 0 to 1) for men and women at different levels of outdoor temperature is given:

Temp.	1	2	3	4	5	6	7	8	Average	sd
Male	0.52	0.53	0.54	0.53	0.50	0.46	0.44	0.31	0.479	0.077
Female	0.76	0.71	0.68	0.64	0.50	0.51	0.43	0.31	0.568	0.155
Average	0.640	0.620	0.610	0.585	0.500	0.485	0.435	0.310	0.523	

As an initial analysis, a two-way analysis of variance, with temperature level and gender as explanatory factors is done. The result is shown in the R output below (where significant codes have been removed and some numbers are replaced by letters):

```
## Analysis of Variance Table
##
## Response: clo
##           Df   Sum Sq   Mean Sq F value  Pr(>F)
## temp       7 0.179194 0.0255991       A  0.01643
## sex        1 0.031506 0.0315062       B  0.03141
## Residuals  7 0.030644 0.0043777
## ---
```

Question X.1 (25)

What are A and B?

1 □ $A = 7.20$ og $B = 5.85$

2 □ $A = 5.85$ og $B = 50.39$

3* □ $A = 5.85$ og $B = 7.20$

4 □ $A = 5.85$ og $B = 1.03$

5 □ $A = 0.813$ og $B = 7.20$

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

See Theorem 8.22. A and B are the F-test statistics for the treatment and block respectively.These are given by

$$F_{Tr} = \frac{MS(Tr)}{MSE} \tag{1}$$

and

$$F_{Bl} = \frac{MS(Bl)}{MSE} \tag{2}$$

These number can be seen directly from the table and the F-test statistics can be calculated by

```
0.0255991/0.004377

## [1] 5.848549

0.0315062/0.004377

## [1] 7.198127
```

which is answer no. 3.

------------------------------------ FACIT-END ------------------------------------

Question X.2 (26)

What is the conclusion (at significance level $\alpha = 0.05$) about the effect of temperature level and gender (both the conclusion and the argument must be correct)?

1 □ The temperature effect is not significant since $0.18 > 0.05$, while there is significant difference between men and women since $0.032 < 0.05$

2 □ The temperature effect is significant since $0.18 > 0.05$, while there is no significant difference between men and women since $0.032 < 0.05$

3 □ Both are significant since $0.0044 < 0.05$.

4* □ Both are significant since $0.016 < 0.05$ and $0.031 < 0.05$.

5 □ Both are significant since $0.031 < 0.05$

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

The p-value for temperature is 0.01643 and the p-value for gender is 0.03141 both are below 0.05 and hence significant (answer no. 4).

```
------------------------------------ FACIT-END ------------------------------------
```

Question X.3 (27)

What is the expected clothing level for men at a temperature level of 1?

1 □ 0.520

2 □ 0.523

3 □ 0.560

4* □ 0.596

5 □ 0.640

```
----------------------------------- FACIT-BEGIN -----------------------------------
```

We have the two-way ANOVA model in Equation 8-34. In order to get the expected value of cell (1,1)

$$\mathrm{E}(Y_{11}) = E(\mu + \alpha_1 + \beta_1 + \varepsilon_{11}) = \mu + \alpha_1 + \beta_1.$$

We find the estimates from Equations 8-35 to 8-37. From the table above find the values we need

$$\begin{aligned}\mu &= 0.523\\ \alpha_1 &= 0.479 - 0.523 = -0.044\\ \beta_1 &= 0.640 - 0.523 = 0.117\end{aligned}$$

So the result is

$$0.523 - 0.044 + 0.117 = 0.596$$

```
0.523+(0.64-0.523)+(0.479-0.523)

## [1] 0.596

(0.48+0.64)/2

## [1] 0.56
```

------------------------------------ FACIT-END ------------------------------------

Question X.4 (28)

What is a pre-planned 95% confidence interval for the difference in mean value for men and women (i.e. it was planned to make this confidence interval only, before the sample was collected)?

1 □ $[-0.224, 0.047]$

2 □ $[-0.245, 0.067]$

3* □ $[-0.167, -0.011]$

4 □ $[-0.220, 0.043]$

5 □ $[-0.144, -0.034]$

----------------------------------- FACIT-BEGIN -----------------------------------

We use Method 8.9 to calculate the single pre-planned post-hoc 95% confidence interval

```
0.479-0.568+c(-1,1)*qt(0.975,df=7)*sqrt(0.0043777*2/8)

## [1] -0.16722672 -0.01077328
```

The degrees of freedom $n - k = 7$ and the $\frac{SSE}{n-k} = MSE = 0.0043777$ we get from the ANOVA table printed in the R result.

------------------------------------ FACIT-END ------------------------------------

Question X.5 (29)

As an aid for this question, the following R code for reading in data, is given:

```
tab <- data.frame(clo=c(0.52,0.53,0.54,0.53,0.50,0.46,0.44,0.31,
                        0.76,0.71,0.68,0.64,0.50,0.51,0.43,0.31))
tab$temp <- as.factor(c(1,2,3,4,5,6,7,8,1,2,3,4,5,6,7,8))
```

What would the p-value for the effect of temperature be if you did not take gender into account?

1 □ 0.031

2 □ 0.016

3* □ 0.058

4 □ 0.00011

5 □ 0.069

----------------------------------- FACIT-BEGIN -----------------------------------

We have to do the one-way ANOVA analysis only including the `clo` variable. One way would be to read in the data using the R code and then the result would simply be found using the call

```
anova(lm(clo ~ temp, data = tab))

## Analysis of Variance Table
##
## Response: clo
##           Df  Sum Sq   Mean Sq F value  Pr(>F)
## temp       7 0.17919 0.0255991  3.2951 0.05826 .
## Residuals  8 0.06215 0.0077688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

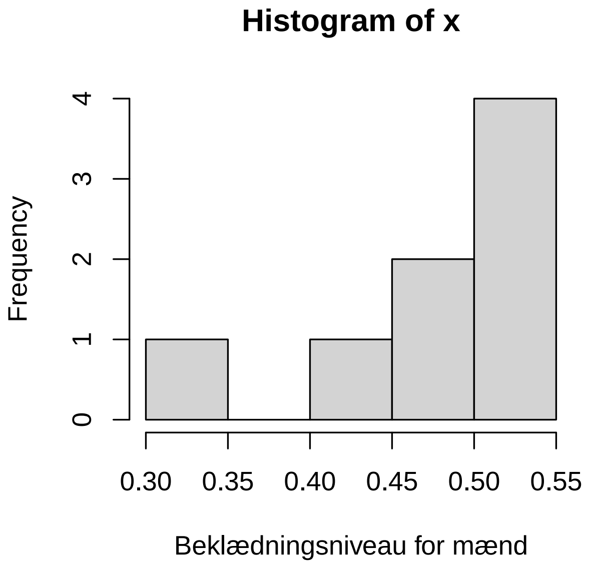
However, we could calculate it “by-hand” using Method 8.6 taking values only from the table and result given in the beginning of the exercise

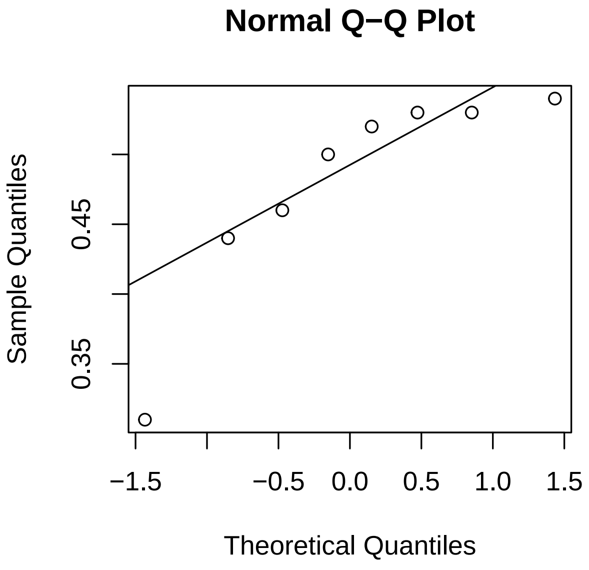
```
sse <- 0.031506+0.030644
mse <- sse/8
F <- 0.0255991/mse
1-pf(F,df1=7,df2=8)

## [1] 0.05826369
```

```
-------------------------------------- FACIT-END --------------------------------------
```

Question X.6 (30)

The following histogram and normal qq-plot are of the observed clothing level for men:





What can rightly be judged based on the observed distribution?

1* □ That the distribution of clothing level is probably left-skewed.

2 □ That the distribution of clothing level is probably right-skewed.

3 □ That the distribution of clothing level is certainly symmetrical.

4 □ That the distribution of clothing level is certainly normally distributed.

5 □ That the distribution of clothing level is certainly exponentially distributed.

----------------------------------- FACIT-BEGIN -----------------------------------

It is clear from the first plot that the mean is lower than the median (since there is a high density in the right-most bar), which is typical for left-skewed distributions. Furthermore it is seen from the QQ-plot that the points do not follow the line very well, and therefore the distribution does not follow a normal distribution.

------------------------------------ FACIT-END ------------------------------------

The exam is finished. Have a great summer!