Written examination: 17. December 2017

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

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

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

Exercise	XII.1	XIII.1	XIII.2	XIV.1	XV.1	XVI.1	XVI.2	XVII.1	XVII.2	XVIII.1
Question	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Answer	4	1	2	3	5	4	2	1	5	5

The questionnaire contains 45 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

We consider pairwise measurements of 2 stochastic variables, X og Y. Both variables can be assumed normally distributed.

x	38	35	47	38	42	41	48	35
y	25	21	26	23	28	27	29	18

Data can be loaded into R using the following command:

```
x <- c(38, 35, 47, 38, 42, 41, 48, 35)
y <- c(25, 21, 26, 23, 28, 27, 29, 18)
```

Question I.1 (1)

Provide an estimate for the correlation coefficient, ρ, between X and Y:

1 ☐ $\hat{\rho} = 0.12$

2 ☐ $\hat{\rho} = 0.22$

3 ☐ $\hat{\rho} = 0.64$

4 ☐ $\hat{\rho} = 0.73$

5* ☐ $\hat{\rho} = 0.82$

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

See Definition 1.19. The easiest way to do this is to use R. We copy into R to read in the data two vectors

```
x <- c(38, 35, 47, 38, 42, 41, 48, 35)
y <- c(25, 21, 26, 23, 28, 27, 29, 18)
```

and the we calculate the estimate of the correlation as the sample correlation

```r
cor(x,y)

## [1] 0.8237548
```

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

Continues on page 4

Exercise II

A biologist is evaluating the effect of 3 different diets on weight change in mice. In the experiment 4 different strains (genetically different types) of mice are included, as strain is expected to have an influence on weight change. Thus, 4 different strains of mice are exposed to 3 different diets, i.e. a total of 12 mice are included. The weight change is measured after 5 weeks for each diet. The weight change is denoted Y_{ij} (in grams). The weight change can be assumed normally distributed and thus the following model has been applied

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}.$$

In this model α_i denote the effect of diet i ($i = 1, 2, 3$) and β_j denotes the effect of mouse strain j ($j = 1, 2, 3, 4$). μ is the overall mean and ε_{ij} are the errors, assumed independent and normally distributed with mean 0 and constant standard deviation σ_ε.

Question II.1 (2)

State the critical value when you want to test whether the mean weight change is the same for the 3 diets and the significance level is $\alpha = 0.05$.

1 ☐ 12.20

2* ☐ 5.14

3 ☐ 1.96

4 ☐ 3.81

5 ☐ 4.35

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

The test we need to carry out is the F-test for a two-way ANOVA, in this case for the effect of diet which is the treatment following the book (Theorem 8.22). Hence, we have to find the two degrees of freedom and look up the $1 - \alpha$ quantile. The degrees of freedom are:

- $df_1 = k - 1 = 2$, since the number of levels for the treatment (number of diets) $k = 3$
- and $df_2 = (k - 1)(l - 1) = 2 \cdot 3 = 6$

The answer is then found by:

```
qf(p=0.95, df1=2, df2=6)

## [1] 5.143253
```

Note, that the results doesn't change if blocks and treatments are switched, i.e. if the diets are thought of as blocks and strains as treatments.

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

Question II.2 (3)

Assume that we have estimated the model parameters using R and concluded that both diet and type of mouse strain are statistically significant. Also assume that the model residuals, $\hat{\varepsilon}_{ij}$, are stored in the vector `resi`, and that we will use R to further analyze these. Which of the following claims is not correct?

1 ☐ The command `qqnorm(resi)` plots normal scores for $\hat{\varepsilon}_{ij}$

2 ☐ The command `plot(ecdf(resi))` plots the cumulative distribution for $\hat{\varepsilon}_{ij}$

3 ☐ The command `sum(resi*resi)/6` gives an estimate of the variance of ε_{ij}

4* ☐ The command `qnorm(resi)` gives a test for normality of ε_{ij}

5 ☐ The command `sum(resi)/length(resi)` gives an estimate for the mean of ε_{ij}

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

Lets go through the answers one by one:

1. TRUE statement. The residuals are sorted and, where $i \in (1, 2, \ldots, n)$ denotes the i'th element in sorted order, $\hat{\varepsilon}_i$ is plotted versus the $(i-0.5)/n$ quantile in the standard normal distribution
2. TRUE statement
3. TRUE statement. We know that MSE is an estimate of the error variance. And this can be found as $\frac{SSE}{(k-1)(l-1)}$.
4. FALSE statement. The command `qnorm(resi)`returns the quantiles in the standard normal distribution of the values in `resi`
5. TRUE statement. It is the sample mean, which is used as an estimate of the mean

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

Continues on page 6

Exercise III

In a study 605 test persons, all with a record of previous heart disease, were randomized to one of two possible diets (A or B), in order to study the effect of diet on health. After an observation period of 4 years the test persons were classified according to health status: (I) dead, (II) cancer, (III) other disease, (IV) well.

Health status

	I	II	III	IV	Total
Diet A	15	24	25	239	303
Diet B	7	14	8	273	302
Total	22	38	33	512	605

The null hypothesis in the study was that there is no association between diet and health.

Question III.1 (4)

State the distribution of the usual test statistics, when assuming that the null hypothesis is true:

1 ☐ The usual test statistics follows a χ^2-distribution with 8 degrees of freedom

2 ☐ The usual test statistics follows a F-distribution with (1, 603) degrees of freedom

3 ☐ The usual test statistics follows a t-distribution with 4 degrees of freedom

4 ☐ The usual test statistics follows a t-distribution with 302 degrees of freedom

5* ☐ The usual test statistics follows a χ^2-distribution with 3 degrees of freedom

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

The setup of the data is a multi-sample proportion setup (chapter 7.4). We must test the hypothesis, that the proportions in each group is equal

$$H_0 : P_1 = p_2 = p_3 = p_4.$$

and under this hypothesis the test statistic follows a χ^2-distribution with $c - 1$ degrees of freedom, and there are 4 groups, so 3 degrees of freedom (Method 7.20).

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

Question III.2 (5)

We now only consider the proportion of test persons who are healthy at the end of the 4 year period. We want to estimate at 95% confidence interval for the difference in proportions of test persons who are healthy for each of the 2 diets. Which of the suggestions below is the correct code in R to achieve this?

1 ☐ `prop.test(x=c(512), n=c(605), correct=FALSE)`

2 ☐ `prop.test(x=c(303,302), n=c(512,512), correct=FALSE)`

3* ☐ `prop.test(x=c(239,273), n=c(303,302), correct=FALSE)`

4 ☐ `prop.test(x=c(239,273), n=c(605,605), correct=FALSE)`

5 ☐ `prop.test(x=c(239,273), n=c(512,512), correct=FALSE)`

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

Here we are working with proportions in two populations as described in Chapter 7.3. We need the observed proportion which are well for each diet. So on Diet A 239 out of 303 are well and for Diet B 273 out of 302 are well, and these numbers are passed to `prop.test`, which then prints out the estimated confidence interval (same as Example 7.19).

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

Continues on page 8

Exercise IV

In the production of a consumer product 3 subprocesses are involved, denoted A, B and C. The time (in hours) it takes to complete each subprocess is represented with a random variable, which we denote X_A, X_B and X_C, respectively. It can be assumed, that X_A, X_B and X_C are all independent and normally distributed given by $X_A \sim N(12, 2^2)$, $X_B \sim N(25, 3^2)$ and $X_C \sim N(42, 4^2)$.

The total production time, Y, is now defined by

$$Y = X_A + X_B + X_C.$$

Question IV.1 (6)

State the probability that the total production time, Y, exceeds 85 hours:

1 ☐ 0.0081

2 ☐ 0.1080

3* ☐ 0.1326

4 ☐ 0.4180

5 ☐ 0.6301

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

We need to find the mean and variance of Y, which we know is normal distributed, since a linear function of normal distributed random variables is also normal distributed (Theorem 2.56).

We use the identities in Theorem 2.56 to get

$$\mu_Y = \mathrm{E}(Y) = \mathrm{E}(X_A + X_B + X_C) = \mathrm{E}(X_A) + \mathrm{E}(X_B) + \mathrm{E}(X_C) = 12 + 25 + 42 = 79,$$

and

$$\sigma_Y^2 = \mathrm{V}(Y) = \mathrm{V}(X_A + X_B + X_C) = \mathrm{V}(X_A) + \mathrm{V}(X_B) + \mathrm{V}(X_C) = 4 + 9 + 16 = 29.$$

Alternatively we could also have simulated the variance in R.

```
k <- 1000000
X_a <- rnorm(k, 12, 2)
X_b <- rnorm(k, 25, 3)
X_c <- rnorm(k, 42, 4)
Y <-X_a + X_b + X_c
var(Y)
```

```
## [1] 29.01567
```

This we use to look up the probability $P(Y > 85) = 1 - P(Y \leq 85)$ in R by:

```
1 - pnorm(q=85, mean=79, sd=sqrt(29))
## [1] 0.1326027
```

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

Question IV.2 (7)

An engineer is now able to perform some optimization of the process, so that the improved process time Y^*, becomes

$$Y^* = 0.9 \cdot X_A + 0.8 \cdot X_B + X_C,$$

where X_A, X_B and X_C are defined as in the previous question.

State the variance of Y^*:

1 ☐ $\mathrm{V}(Y^*) = (0.9 + 0.8 + 1) \cdot (2^2 + 3^2 + 4^2)$

2 ☐ $\mathrm{V}(Y^*) = (0.9^2 + 0.8^2 + 1^2) \cdot (2^2 + 3^2 + 4^2)$

3 ☐ $\mathrm{V}(Y^*) = 0.9 \cdot 2^2 + 0.8 \cdot 3^2 + 1 \cdot 4^2$

4* ☐ $\mathrm{V}(Y^*) = 0.9^2 \cdot 2^2 + 0.8^2 \cdot 3^2 + 1^2 \cdot 4^2$

5 ☐ $\mathrm{V}(Y^*) = 2^2 + 3^2 + 4^2$

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

Again the identities in Theorem 2.56 to get

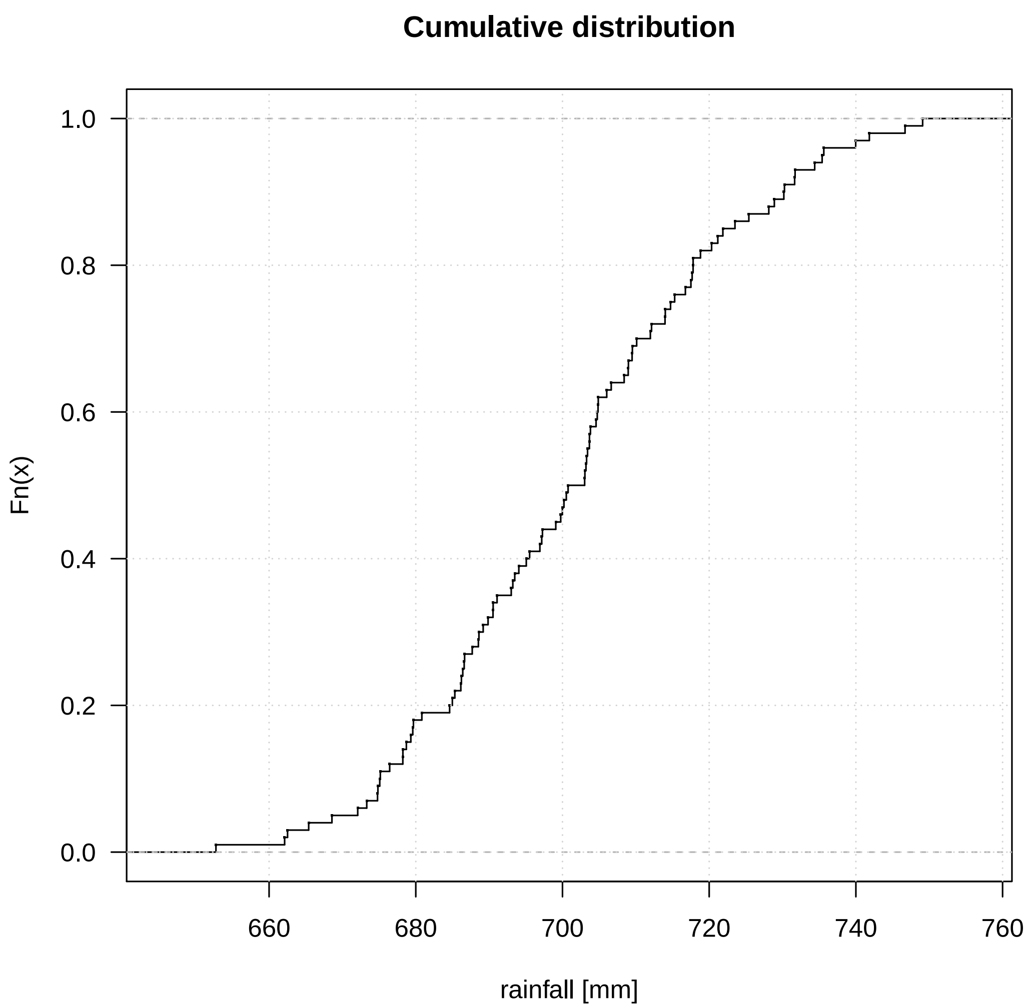
$$\begin{aligned}
\sigma^2_{Y^*} = \mathrm{V}(Y^*) &= \mathrm{V}(0.9 \cdot X_A + 0.8 \cdot X_B + X_C,) \\
&= 0.9^2 \, \mathrm{V}(X_A) + 0.8^2 \, \mathrm{V}(X_B) + \mathrm{V}(X_C) \\
&= 0.9^2 \cdot 2^2 + 0.8^2 \cdot 3^2 + 1^2 \cdot 4^2
\end{aligned}$$

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

Continues on page 10

Exercise V

The yearly rainfall has been registered within a region for the last 100 years. It can be assumed that the rainfall is independent from year to year. The cumulative distribution for the yearly rainfall is shown in the figure below:



The following summary of the data has been conducted by the use of R, where the yearly rainfall measurements are stored in the variable `rainfall`:

```
> var(rainfall)
[1] 412.7042
> summary(rainfall)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  652.8   686.6   701.9   701.3   714.9   749.1
```

Continues on page 11

Question V.1 (8)

Which of the following statements is not correct?

1 ☐ The estimate of the standard deviation of the mean $\hat{\sigma}_{\bar{X}}$, becomes $\frac{\sqrt{412.7042}}{10}$ mm

2 ☐ The 50% quantile for the 100 observations is 701.9 mm

3 ☐ The standard deviation of the sample, s, for the 100 measurements is $\sqrt{412.7042}$ mm

4 ☐ 50% of the 100 observations are between 686.6 and 714.9 mm

5* ☐ The estimated coefficient of variation for the 100 observations becomes $\frac{412.7042}{701.9}$

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

Lets go through them one by one:

1. TRUE statement. The formula for the estimate is $\frac{s}{\sqrt{n}}$ (also called the standard error of the mean). See Definition 3.7
2. TRUE statement. Seen from the `summary()` call
3. TRUE statement. Standard deviation is the square root of the variance
4. TRUE statement. 686.6 is the first quartile (25% quantile) and 714.9 is the third quartile (75% quantile), and certainly 50% of the observations lies between the 25% and 75% quantile
5. FALSE statement. The estimated coefficient of variation is $\hat{V} = \frac{s}{\bar{x}} = \frac{\sqrt{412.7042}}{701.3}$. See Definition 1.12.

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

Question V.2 (9)

Provide a 95% confidence interval for the variance of the rainfall based on the 100 observations, still assumed to be normally distributed:

1 ☐ $\left[\frac{20.31512^2 \cdot 134.6416}{99}; \frac{20.31512^2 \cdot 69.22989}{99}\right]$

2 ☐ $\left[\frac{20.31512^2 \cdot 99}{134.6416}; \frac{20.31512^2 \cdot 99}{69.22989}\right]$

3* ☐ $\left[\frac{412.7042 \cdot 99}{128.422}; \frac{412.7042 \cdot 99}{73.36108}\right]$

4 ☐ $\left[\frac{412.7042 \cdot 99}{123.2252}; \frac{412.7042 \cdot 99}{77.04633}\right]$

5 ☐ $\left[\frac{20.31512 \cdot 99}{123.2252}; \frac{20.31512 \cdot 99}{77.04633}\right]$

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

We find the formula for a $1-\alpha$ confidence interval for the variance of a normal distributed population in Method 3.19 and insert the values

$$\left[\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}, \frac{s^2(n-1)}{\chi^2_{\alpha/2}}\right]$$

The chi-square quantiles are found in R as

```
qchisq(c(0.025, 0.975), 99)

## [1]  73.36108 128.42199
```

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

Continues on page 13

Question V.3 (10)

We continue with the exercise from the previous page. The following code in R has now been run:

```
k = 10^5
Q5 = function(x){ quantile(x, 0.95) }
samples = replicate(k, sample(rainfall, replace = TRUE))
simvalues = apply(samples, 2, Q5)
interval = quantile(simvalues, c(0.025,0.975))
```

which gives the result:

```
> interval
    2.5%    97.5%
728.9515 742.0814
```

What has been calculated in the vector `interval`?

1 ☐ A 95% confidence interval for the mean of the yearly rainfall (parametric bootstrap)

2 ☐ A 95% confidence interval for the 5% quantile of the yearly rainfall (parametric bootstrap)

3* ☐ A 95% confidence interval for the 95% quantile of the yearly rainfall (non-parametric bootstrap)

4 ☐ A 95% confidence interval for the 2.5% and 97.5% quantile of the yearly rainfall (non-parametric bootstrap)

5 ☐ A 95% confidence interval for the 2.5% and 97.5% quantile of the yearly rainfall (parametric bootstrap)

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

We look at the R code and see that it is a bootstrapping is carried out by simulating the sample 100000 times, and not assuming any distribution (since the `sample` function is used), therefore it is non-parametric.

The statistic calculated for each simulated sample is the 95% quantile and since the quantiles taken for these values are the 2.5% and the 97.5%, then the results is a 95% confidence interval for the 95% quantile.

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

Continues on page 14

Exercise VI

We consider an experiment that can result in one of two possible outcomes, here denoted A or B. The probability of outcome A is denoted $P(A)$. By defintion we get the probability of outcome B as $P(B) = 1 - P(A)$.

Question VI.1 (11)

Assume that we observe a random variable, X, which counts the number of times that we observe the outcome A out of $n = 300$ independent trials of the experiment. If we assume that $P(A) = 0.40$ in a single trial, what is then the expected number $\mathrm{E}(X)$ and variance $\mathrm{V}(X)$?

1 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot (1 - 0.4)$ and $\mathrm{V}(X) = 300^2 \cdot 0.4$

2 ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ and $\mathrm{V}(X) = 300^2 \cdot 0.4 \cdot 0.6$

3* ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ and $\mathrm{V}(X) = 300 \cdot 0.4 \cdot 0.6$

4 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ and $\mathrm{V}(X) = 300^2 \cdot 0.4^2 \cdot 0.6^2$

5 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ and $\mathrm{V}(X) = 300 \cdot 0.4^2 \cdot 0.6^2$

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

X follows a Binomial distribution with $p = 0.4$ and we have a formula for the mean and variance defined in Theorem 2.21, which we use to get

$$\mu = \mathrm{E}(X) = np = 300 \cdot 0.4,$$
$$\sigma^2 = \mathrm{V}(X) = np(1-p) = 300 \cdot 0.4 \cdot 0.6.$$

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

Question VI.2 (12)

Regardless of your answer to the previous question we now want to estimate the probability $P(A)$ based on the $n = 300$ trials. From the $n = 300$ trials we count that in 120 of these the outcome was A and in the remaining 180 trials the outcome was B. Provide a 95% confidence interval for the probability $P(A)$:

1 ☐ [0.33, 0.48]

2* ☐ [0.35, 0.46]

3 ☐ [0.35, 0.42]

4 ☐ [0.31, 0.53]

5 ☐ [0.29, 0.54]

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

Using the inbuilt function in R the results is

```
prop.test(120, 300, correct=FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  120 out of 300, null probability 0.5
## X-squared = 12, df = 1, p-value = 0.000532
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3461652 0.4563634
## sample estimates:
##   p
## 0.4
```

whereas using the formula in method 7.3 gives a slightly different result is obtained

```
n <- 300
x <- 120
phat <- x/n
phat + c(-1,1) * qnorm(p=0.975) * sqrt(phat*(1-phat)/n)

## [1] 0.3445638 0.4554362
```

This is due to a numerical rounding by R and can occur sometimes. The answer is in any case closest to the answer marked correct [0.35, 0.46].

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

Continues on page 16

Exercise VII

An engineer is examining the quality in a batch of raw materials. The quality demand is that the purity of the raw material is at least 90%. The engineer takes a sample of 10 independent measurements from the batch and saves the measured values (in %) of the purity in a vector `x`.

He then runs the following code in R

```
> x <- c(90.6, 90.3, 88.9, 87.5, 87.6, 88.1, 87.5, 88, 88, 89.6)
> n <- length(x)
> tobs <- (mean(x) - 90) / (sd(x) / sqrt(n))
> pt(tobs, df=n-1)
```

Which yields the following output

```
[1] 0.002279236
```

Question VII.1 (13)

Based on the calculations listed above, and assuming that the measurements of the purity are normally distributed and applying a significance level of $\alpha = 0.05$, what can the engineer conclude?

1 ☐ The engineer can conclude that the purity of the raw material is at least 88.6%

2 ☐ The engineer can conclude that the mean purity of the raw material is at most 88.6%

3 ☐ The engineer has with probability 99.7% shown that the mean purity of the raw material is 90%

4 ☐ The engineer can assume that the mean purity of the raw material is 90%

5* ☐ The engineer can reject that the mean purity of the raw material is 90%

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

We can see from the way that `tobs` is calculated that the null hypothesis is that the $\mu = 90$ (See Method 3.23) Since the p-value is `2*pt(tobs, df=n-1)=0.0046` and thus much lower than $\alpha = 0.05$. This leads to the conclusion that the null hypothesis, that the mean purity is 90%, must be rejected.

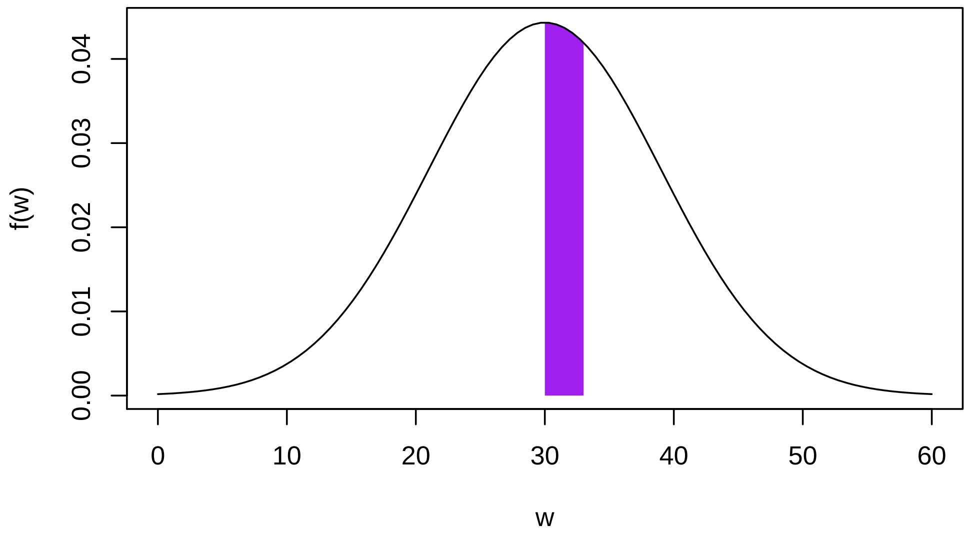
------------------------------------- FACIT-END -------------------------------------

Continues on page 17

Exercise VIII

We consider a random variable W with density function $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$.

The density function is shown in the figure below, where the probability $P(30 < W < 33)$ is shown as the shaded area.



Question VIII.1 (14)

Calculate the probability $P(30 < W < 33)$:

1 ☐ 0.09

2* ☐ 0.13

3 ☐ 0.24

4 ☐ 0.34

5 ☐ 0.84

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

The answer is obtained from recognizing the the formula for the probability density function (pdf) for the normal distribution in definition 2.37

$$f(w) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(w-\mu)^2}{2\cdot\sigma^2}}$$

and thus to find the mean $\mu = 30$ and variance $\sigma = 9$. These are then used to obtain

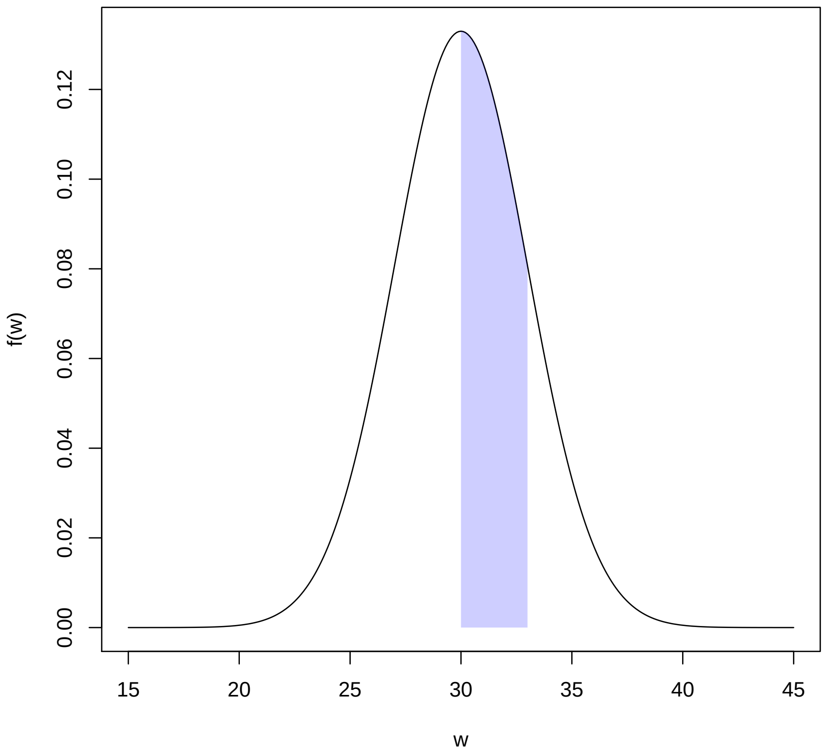
$$P(30 < W < 33) = P(X < 33) - P(X < 30)$$

in R

```
pnorm(33, mean=30, sd=9) - pnorm(30, mean=30, sd=9)

## [1] 0.1305587
```

SINCE in the original exam the plot was which indeed was wrong, it was of the normal distri-



bution with mean $\mu = 30$ and variance $\sigma = 3$

```
pnorm(33, mean=30, sd=3) - pnorm(30, mean=30, sd=3)

## [1] 0.3413447
```

then the Answer 4 is also counted as correct!

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

Continues on page 19

Question VIII.2 (15)

We consider a situation where we take 3 different samples denoted A, B, and C. All three samples are from the population characterized by the density $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$ as in the previous question.

Sample A is of size $n_{\mathrm{A}} = 10$ and the estimated mean is denoted $\hat{\mu}_{\mathrm{A}}$. Sample B is of size $n_{\mathrm{B}} = 30$ and the estimated mean is denoted $\hat{\mu}_{\mathrm{B}}$. Sample C is of size $n_{\mathrm{C}} = 100$ and the estimated mean is denoted $\hat{\mu}_{\mathrm{C}}$.

The question is now whether the sample mean will exceed the value 33, even when the population mean is equal to 30.

Which statement is correct?

1* ☐ $P(\hat{\mu}_{\mathrm{A}} \geq 33) > P(\hat{\mu}_{\mathrm{B}} \geq 33)$

2 ☐ $P(\hat{\mu}_{\mathrm{C}} \geq 33) > P(\hat{\mu}_{\mathrm{A}} \geq 33)$

3 ☐ $P(\hat{\mu}_{\mathrm{C}} \geq 33) = P(\hat{\mu}_{\mathrm{B}} \geq 33)$

4 ☐ $P(\hat{\mu}_{\mathrm{A}} \geq 33) = P(\hat{\mu}_{\mathrm{B}} \geq 33) \cdot P(\hat{\mu}_{\mathrm{C}} \geq 33)$

5 ☐ $P(\hat{\mu}_{\mathrm{A}} \geq 33) = \frac{1}{2} P(\hat{\mu}_{\mathrm{B}} \geq 33)$

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

Lets go through the statements:

1. TRUE statement. The $\hat{\mu}$ is the sample mean, which we know follow the distribution $\hat{\mu} \sim N(\mu, \sigma^2/n)$ (Theorem 3.3), so we get the following

$$\hat{\mu}_{\mathrm{A}} \sim N(30, 81/10)$$
$$\hat{\mu}_{\mathrm{B}} \sim N(30, 81/30)$$
$$\hat{\mu}_{\mathrm{C}} \sim N(30, 81/100)$$

and we can actually then realize, that the probability of getting a an outcome above the same value, must be higher for X_{A} than the two others, since its pdf has higher variance than the others. In R we can check it by:

```r
## P(X_A >= 33)
(1-pnorm(q=33, mean=30, sd=sqrt(81/10)))

## [1] 0.1459203

## P(X_B >= 33)
(1-pnorm(q=33, mean=30, sd=sqrt(81/30)))

## [1] 0.03394458
```

2. FALSE statement. Following same argument as above
3. FALSE statement. Since the variance is different, then they are not equal
4. FALSE statement. Be sure by checking the product in R:

```
(1-pnorm(q=33, mean=30, sd=sqrt(81/30))) *
  (1-pnorm(q=33, mean=30, sd=sqrt(81/100)))

## [1] 1.456427e-05
```

5. FALSE statement. Be sure by checking the product in R:

```
0.5 * (1-pnorm(q=33, mean=30, sd=sqrt(81/30)))

## [1] 0.01697229
```

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

Continues on page 22

Exercise IX

The yield from a chemical process, Y_i, is assumed to depend linearly on the temperature, t_i, measured in degrees. In order to achieve insight about this relation, an experiment has been conducted where $n = 50$ pairwise measurements of Y_i and t_i has been taken. It is assumed that the following model can give a reasonable description of the relation

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \varepsilon_i.$$

The residuals in this model are assumed independent and normally distributed with constant variance, i.e. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Relevant output from the analysis in R is given below:

```
Call:
lm(formula = y ~ t)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0816 -1.4994 -0.2493  1.5175  4.8506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.4919     2.7757  23.595   <2e-16 ***
t             0.1637     0.1103   1.485    0.144
---


Residual standard error: 2.296 on 48 degrees of freedom
Multiple R-squared:  0.04392,Adjusted R-squared:  0.024
F-statistic: 2.205 on 1 and 48 DF,  p-value: 0.1441
```

Question IX.1 (16)

Which of the following statements is correct when the significance level $\alpha = 0.05$ is applied?

1 ☐ The yield increases by 16.37% when the temperature increase one degree

2* ☐ There is no significant linear relation between temperature and yield

3 ☐ The test statistics for no effect of temperature on yield (i.e. the null hypothesis $H_0 : \beta_1 = 0$) is 23.595

4 ☐ A 95% confidence interval for the effect of temperature, β_1, is [-0.132027, 0.4594821]

5 ☐ The correlation between temperature and yield is 0.04392

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

Lets go through the answers one by one:

1. FALSE statement. The yield is estimated to increase 0.1637 units (we are not informed about the units) per degree, which is not the same as 16.37% (increasing some proportion per degree, would also lead to an exponential relation, not linear)

2. TRUE statement. The test of the null hypothesis

$$H_0 : \beta_1 = 0$$

leads to a p-value of 0.144, which is not below the significance level $\alpha = 0.05$ and since this is equivalent to testing for correlation equal to zero

$$H_0 : \rho = 0$$

there is not found a significant linear relation between the yield and the temperature

3. FALSE statement. Since, the test statistic for no effect is 1.485

4. FALSE statement. The lower limit of the CI is $0.1637 - 1.96 * 0.1103 = -0.052$ and the upper is $0.1637 + 1.96 * 0.1103 = 0.380$

5. FALSE statement. The correlation is $\sqrt{r^2} = \sqrt{0.04392} = 0.21$

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

Continues on page 24

Question IX.2 (17)

We continue with the exercise from the previous page. It turns out that the pH of the process may influence the yield, and since pH has been measured, it is decided to include it into the model, which in its extended form becomes:

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Estimation of the model parameters gives the following output in R:

```
Call:
lm(formula = y ~ t + pH)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7253 -1.2818 -0.2978  1.0724  4.4488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.46756    4.09799  12.071 5.25e-16 ***
t            0.24113    0.09315   2.589   0.0128 *
pH           2.37090    0.50097   4.733 2.06e-05 ***
---
Signif. codes:  0  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.91 on 47 degrees of freedom
Multiple R-squared:  0.3525,Adjusted R-squared:  0.3249
F-statistic: 12.79 on 2 and 47 DF,  p-value: 3.667e-05
```

Give estimates for the model parameters, i.e. β_0, β_1, β_2 and σ_ε^2

1 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.24113,\ 2.37090, 0.3525)$

2* ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090, 1.91^2)$

3 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090, 1.91 \cdot 47)$

4 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.09315,\ 0.50097, 1.91 \cdot 47)$

5 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (2.37090,\ 0.50097,\ 4.733\ , 1.91)$

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

The estimates are read directly from the printed output. See Example 6.3

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

Continues on page 25

Question IX.3 (18)

We continue with the exercise from the previous page and the model

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Provide a 95% confidence interval for the effect on yield when pH increases one unit:

1 ☐ $0.24113 \pm 2.01174 \cdot 0.09315$

2* ☐ $2.37090 \pm 2.01174 \cdot 0.50097$

3 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 2.01174 \cdot (4.09799 + 0.09315 + 0.50097)$

4 ☐ $2.37090 \pm 0.509920 \cdot 0.50097$

5 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 0.509920 \cdot 0.50097$

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

See Method 6.5. The confidence interval for the effect of pH is found inserting the printed values into

$$\hat{\beta}_2 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_2}$$

using the t-distribution with $n - (p+1) = 47$ degrees of freedom to find the quantile $t_{1-\alpha/2}$:

```
qt(p=0.975, df=47)

## [1] 2.011741
```

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

Exercise X

Assume there exists a dice with 10 sides and where the probability for each of the 10 outcomes, $1, 2, \ldots, 10$, is the same. Consider the discrete random variable X with density $f(x) = 0.1$ for $x \in (1, 2, \ldots, 10)$.

Question X.1 (19)

Give the mean value of X:

1 ☐ $\frac{1}{(10-1)}\sum_{i=1}^{10} x_i = 6.11$

2 ☐ $\frac{1}{(10-6.11)}\sum_{i=1}^{10} |x_i - 6.11| = 6.48$

3 ☐ $\frac{1}{(10)}\sum_{i=1}^{10}(x_i - 6.11)^2 = 8.62$

4 ☐ $\sum_{i=1}^{10} \frac{10-1}{10} x_i \cdot 0.1 = 4.95$

5* ☐ $\sum_{i=1}^{10} x_i \cdot 0.1 = 5.50$

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

See Definition 2.13. We use the formula for calculating the mean value of a discrete random variable

$$\sum_{i=1}^{n} x_i f(x_i)$$

and insert the values. In R:

```
sum(1:10*0.1)

## [1] 5.5
```

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

Continues on page 28

Exercise XI

The yield of a process is $\mu = 60$ mg/l. Certain changes to the process are being planed and it is desirable to be able to prove an effect on the mean yield if the change is at least 5 mg/l (i.e. a two-sided test).

An engineer is now going to plan an experiment to evaluate the effect of the process changes. He wants to decide how large a sample is needed. The sample size has to be large enough to detect the relevant effect (5 mg/l) with a power of 0.8 when applying a significance level of $\alpha = 0.05$. It can be assumed that the standard deviation is $\sigma = 10$ mg/l.

Question XI.1 (20)

Based on the information above, and by applying the function `power.t.test` in R, one concludes that, if an equal number of measurements are taken, then the minimum number of measurements n needed becomes:

1 ☐ $n \simeq 256$ measurements

2 ☐ $n \simeq 128$ measurements

3 ☐ $n \simeq 64$ measurements

4* ☐ $n \simeq 34$ measurements

5 ☐ $n \simeq 27$ measurements

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

Based on the given information the planned test is a one-sample test, since it is not stated that a sample should be taken before the change, only that the yield before is $\mu = 60$ mg/l. See Example 3.67.

```
power.t.test(delta=5, sd=10, sig.level=0.05, power=0.8, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 33.3672
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Rounding up to $n \simeq 34$ measurements.

Since, it is not completely clear, that the it should not be a two-sample setup – one could argue that a nothing in the information given prevents it from being a two-sample test – then Answer 3 is also taken as correct, since:

```
power.t.test(delta=5, sd=10, sig.level=0.05, power=0.8, type="two.sample")

##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 5
##              sd = 10
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Further, since it is also not specified that n is the number of measurements is in each group (and not the total), then Answer 2 is also taken as correct.

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

Continues on page 30

Exercise XII

In a study the aim is to investigate the possible cholesterol lowering effect of a product. 9 test persons had their cholesterol level measured (denoted `x1`). After 3 months, while using the product, the same 9 test persons had their cholesterol level measured again (denoted `x2`). Data is shown in the table below:

Person	1	2	3	4	5	6	7	8	9
x1	63.5	66.7	59.2	57.4	63.9	63.2	60.7	62.6	63.3
x2	51.3	51.9	57.8	50.2	54.6	43.3	51.2	40.4	52.2

The following code is now run in R, in order to test whether the change over time can be assumed to be zero ($H_0 : \delta = 0$):

```
x1 <- c(63.5, 66.7, 59.2, 57.4, 63.9, 63.2, 60.7, 62.6, 63.3)
x2 <- c(51.3, 51.9, 57.8, 50.2, 54.6, 43.3, 51.2, 40.4, 52.2)
```

The output from the standard statistical analysis is given below. Please note that some numbers in the standard output have been replaced by the letters A, B and C.

```
t = -5.6354, df = A, p-value = B
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.847799  C
sample estimates:
mean of the differences
-11.95556
```

Question XII.1 (21)

What conclusion can be made when applying a significance level of $\alpha = 0.05$?

1 ☐ We can show an effect since $\mu_D = -11.95556$

2 ☐ We can not show an effect since the upper limit of the confidence interval is 7.063312

3 ☐ We can not show an effect since the lower limit of the confidence interval is -7.063312

4* ☐ We can show an effect since the p-value is $4.897 \cdot 10^{-4}$

5 ☐ We can show an effect since the p-value is $2.394 \cdot 10^{-4}$

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

The standard statistical test for this setup is a paired two-sample t-test. The R output is from `t.test()`, and the easiest way to solve this is by copying and running

```
x1 <- c(63.5, 66.7, 59.2, 57.4, 63.9, 63.2, 60.7, 62.6, 63.3)
x2 <- c(51.3, 51.9, 57.8, 50.2, 54.6, 43.3, 51.2, 40.4, 52.2)
## The call is then either "t.test(x2, x1, paired=TRUE)" or
t.test(x2-x1)

##
##  One Sample t-test
##
## data:  x2 - x1
## t = -5.6354, df = 8, p-value = 0.0004897
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -16.847799  -7.063312
## sample estimates:
## mean of x
## -11.95556
```

and from the p-value we can find the correct answer. See section 3.1.7 for more examples.

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

Continues on page 32

Exercise XIII

A biologist is interested in examining the effect of 4 different growth inhibitors, denoted V_1, V_2, V_3 og V_4. The 4 growth inhibitors are added to samples from the same cell line and growth after one week is measured Y_{ij} (number of cells per cm^2). 8 replicates are made for each growth inhibitor, i.e. we have a total of 32 measurements. As the measurements can be assumed normally distributed, it is chosen to apply the following analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

In this model α_i denotes the effect of growth inhibitor i ($i = 1, 2, 3, 4$), μ is the overall average ε_{ij} are the errors, assumed independent and normally distributed with mean zero and standard deviation σ_ε.

An analysis of variance is performed for the above model and the output is given below. Please note that the output is incomplete as some numbers are replaced by the symbols A, B and C.

```
Analysis of Variance Table

Response: growth
          Df Sum Sq Mean Sq F value    Pr(>F)
treatment  A 281.07       B       C 0.0001409 ***
Residuals 28 268.46   9.588
```

Question XIII.1 (22)

Provide the usual test statistics (denoted by `C`) in order to test for equal mean effect of the 4 growth inhibitors

1* ☐ 9.77

2 ☐ 7.23

3 ☐ 2.95

4 ☐ 4.57

5 ☐ 16.11

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

As stated in Theorem 8.6, we can calculate the observed test statistic by

$$F_{\text{obs}} = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{281.07/(4-1)}{268.46/(32-4)} = 9.77,$$

where

- $SS(Tr)$ is the variance explained by the effect of the treatment
- SSE is the variance remaining after the model (sum of squared error)
- n is the total number of observations
- k is the number of groups

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

Continues on page 34

Question XIII.2 (23)

We now want to calculate a post hoc 95% confidence interval for a difference in mean between growth inhibitor V_1 and V_2, here denoted $I_{0.95}(V_1 - V_2)$. From the experiment it is known that the estimated mean difference between V_1 and V_2 is 4.5. State the interval $I_{0.95}(V_1 - V_2)$:

1 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \frac{9.588}{12} \cdot \sqrt{28}$

2* ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \sqrt{9.588} \cdot \sqrt{2/8}$

3 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot \frac{\sqrt{9.588}}{\sqrt{12}}$

4 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot 9.588^2 \cdot \sqrt{1/8}$

5 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 1.960 \cdot \frac{9.588}{\sqrt{8}}$

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

See method 8.9. The post hoc confidence interval for the difference is

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

So we use the t-distribution with $n - k = 32 - 4 = 28$ degrees of freedom

```
qt(p=0.975, df=28)

## [1] 2.048407
```

and insert the values

$$4.5 \pm 2.048 \cdot \sqrt{\frac{268.46}{28}\left(\frac{1}{8} + \frac{1}{8}\right)},$$

which we cannot directly find among the answers, so we shorten it

$$4.5 \pm 2.048 \cdot \sqrt{9.588\left(\frac{2}{8}\right)},$$

and finally find the answer

$$4.5 \pm 2.048 \cdot \sqrt{9.588} \cdot \sqrt{2/8}.$$

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

Exercise XIV

We consider a continuous random variable random, where the well-known cumulative distribution function $F(x)$ is given by $P(X \leq x) = 1 - e^{-x/2}$, where $x > 0$.

Question XIV.1 (24)

Provide the mean of X:

1 ☐ $\frac{1}{2}$

2 ☐ 1

3* ☐ 2

4 ☐ $\frac{3}{2}$

5 ☐ 4

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

It is recognized as the cdf of the exponential distribution (Definition 2.48), which is verified by

$$\int_0^x \lambda e^{\lambda y} dy = \left[-e^{-\lambda y} + c\right]_0^x = -e^{-\lambda x} + e^0 = 1 - e^{-\lambda x}$$

and it can be seen that $\lambda = \frac{1}{2}$. Using the formula for the mean of an exponential distribution (Theorem 2.49)

$$\mu = \frac{1}{\lambda} = 2.$$

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

Continues on page 36

Exercise XV

A biologist is examining the bio-diversity within an area and has measured the number of different type of plants per 10 m^2 in different places in the area. She has obtained a total of 30 independent measurements, y_i, and these are in in the vector `Yobs` in R.

Question XV.1 (25)

The biologist would like to estimate a 95% confidence interval for the coefficient of variation for the bio-diversity (number of different type of plants per 10 m^2) by applying the non-parametric bootstrap. Which of the following suggestions in R is most suitable to achieve this?

1 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

2 ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,var)/apply(samples,2,sd)
quantile(results, c(0.025,0.975))
```

3 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,var)/apply(samples,2,median)
quantile(results, c(0.025,0.975))
```

4 ☐
```
samples = replicate(10000,sample(Yobs,replace=FALSE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

5* ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

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

In the code in Answer 1 and 3 the samples are simulated using `rnorm()`, hence a normal distribution is assumed and it is not non-parametric bootstrapping (but parametric).

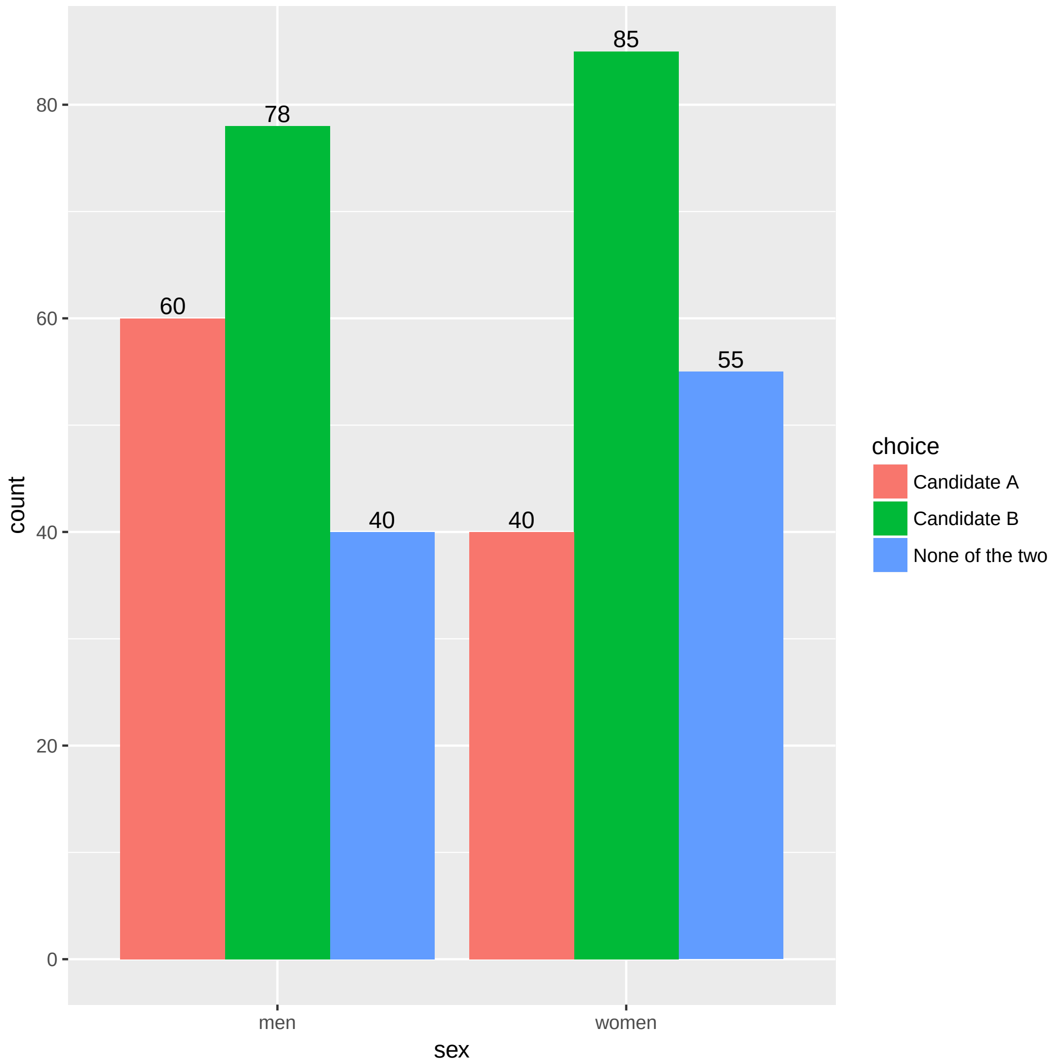
In Answer 2 it is not the coefficient of variation which is calculated by
`apply(samples,2,var)/apply(samples,2,sd)`,
which it is in Answer 4 and 5 by
`apply(samples,2,sd)/apply(samples,2,mean)`.
The difference between 4 and 5 is that in Answer 4 the samples are drawn without replacement
`sample(Yobs,replace=FALSE)`,
which is wrong, where in Answer 5 the samples are drawn correctly with replacement
`sample(Yobs,replace=TRUE)`. See Chapter 4.3 for more on non-parametric bootstrap.

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

Continues on page 37

Exercise XVI

In a study 178 men and 180 women were asked to answer whom of 2 political candidates, A or B, they preferred. Alternatively, they could answer "none of the two". The distribution of the answers is shown in the figure below.



Continues on page 38

Question XVI.1 (26)

It is seen from the figure that we observe that 85 out of the 180 women prefer Candidate B. If we can assume the same distribution of answers by gender, how many women out of the 180 would we expect to prefer Candidate B?

1 ☐ $\frac{163}{358} \cdot \frac{95}{358} \cdot 358$

2 ☐ $\frac{100}{358} \cdot \frac{223}{358} \cdot 358$

3 ☐ $\frac{95}{358} \cdot \frac{190}{358} \cdot 358$

4* ☐ $\frac{163}{358} \cdot \frac{180}{358} \cdot 358$

5 ☐ $\frac{95}{358} \cdot \frac{180}{358} \cdot 358$

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

See chapter 7.2. The total number of respondents are $n = 180 + 178 = 358$ and if we assume the same distribution of answers by gender, i.e. the under the hypothesis that the proportion of men and women prefering B is equal

$$H_0 : p_{\text{men},B} = p_{\text{women},B} = p,$$

then

$$p = \frac{\text{"Total number for B"}}{\text{"Total number"}} = \frac{78 + 85}{358} = \frac{163}{358}.$$

It is then simply this fraction we expect out of the total number of women

$$\frac{163}{358} \cdot 180,$$

which is then expressed a little longer by

$$\frac{163}{358} \cdot \frac{180}{358} \cdot 358.$$

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

Question XVI.2 (27)

Provide the usual test statistics when you want to conduct the test of whether the distribution of answers is the same for men and women:

1 ☐ $\chi^2_{\text{obs}} = 5.9915$

2* ☐ $\chi^2_{\text{obs}} = 6.6581$

3 ☐ $\chi^2_{\text{obs}} = 16.212$

4 ☐ $\chi^2_{\text{obs}} = 8.3836$

5 ☐ $\chi^2_{\text{obs}} = 4.5067$

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

Maybe the easiest is to copy example 7.21 from the book of testing multiple proportions

```
prop <- matrix(c(60, 78, 40, 40, 85, 55), ncol = 3, byrow = TRUE)
rownames(prop) <- c("Men", "Women")
colnames(prop) <- c("A", "B", "None")
prop

##        A  B None
## Men   60 78   40
## Women 40 85   55

chisq.test(prop, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  prop
## X-squared = 6.6581, df = 2, p-value = 0.03583
```

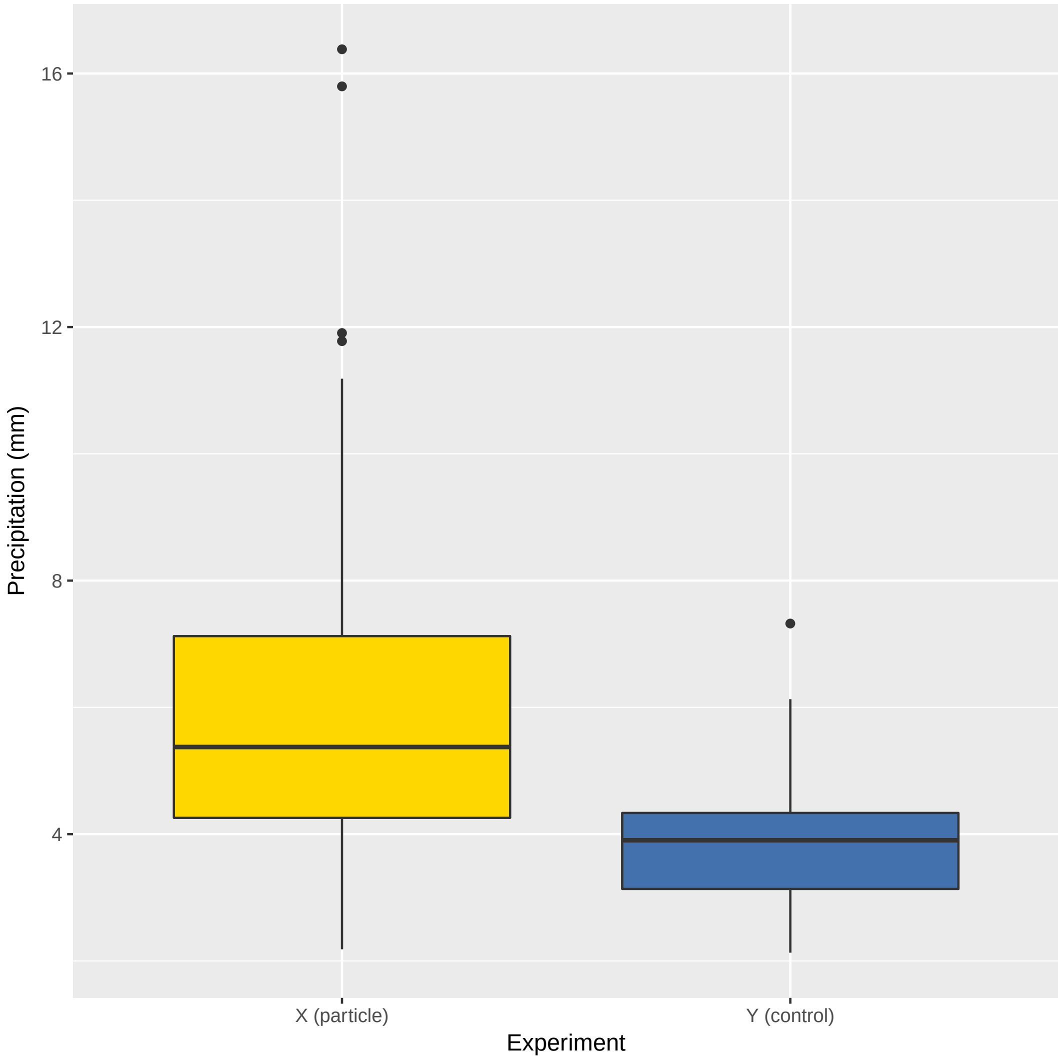
------------------------------------ FACIT-END ------------------------------------

Continues on page 40

Exercise XVII

Cloud seeding is a form of weather modification that can be used to increase the amount of precipitation that falls from the clouds, by dispersing substances (small particles) e.g. aluminiumoxid into the clouds to modify their development.

In an experiment the aim was to study the effect of cloud seeding by using a new type of particles. The amount of precipitation (mm precipitation per day) for 35 days with cloud seeding using the new particles is denoted X_i, $(i = 1, 2, \ldots, 35)$. This was compared to the amount of precipitation on 30 days without cloud seeding, denoted Y_j, $(j = 1, 2, \ldots, 30)$. Measurements were only taken on days where there was sufficient humidity in the air to make the experiment relevant. Data from the experiment is shown in the figure below.



Continues on page 41

We now want to analyze the data described on the previous page using R. Data x_i is stored in the vector `x` and data y_j is stored in the vector `y`, and the following code has been run:

```
k <- 10^4
resultX <- replicate(k, sample(x, replace = TRUE))
resultY <- replicate(k, sample(y, replace = TRUE))
result <- apply(resultX, 2, median) - apply(resultY, 2, median)
quantile(result, c(0.5, 0.025,0.975))
```

Which gives the result

```
      50%      2.5%     97.5%
1.6283069 0.2843492 2.4233546
```

Question XVII.1 (28)

If we apply a significance level of $\alpha = 0.05$ what can then be concluded?

1* ☐ The median for X is significantly higher than the median for Y

2 ☐ The median for X is 62.8% higher than the median for Y

3 ☐ Precipitation for X is between 28.4% and 142.3% higher than precipitation for Y

4 ☐ The mean precipitation can be assumed equal for the two methods

5 ☐ The medianen for Y is $[0.28; 2.42]$ higher than the median for X

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

In the R code a 95% non-parametric bootstrap confidence interval for the difference in median is calculated, and since 0 is not contained in the interval, then the hypothesis

$$H_0 : q_{0.5,X} = q_{0.5,Y}$$

must be rejected on significance level $\alpha = 0.05$, thus concluded that

$$H_1 : q_{0.5,X} \neq q_{0.5,Y}$$

and further, since $X - Y$ was calculated and the interval is on the positive side, then it can be concluded that $q_{0.5,X} > q_{0.5,Y}$.

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

Continues on page 42

Question XVII.2 (29)

In a different experiment using cloud seeding a different kind of particles were examined. Also in this experiment the amount of precipitation was compared when the particles were used to a situation with no use of particles. In this study, however, it was decided to log transform (the natural logarithm) the data before comparing the groups. By transforming the data it can be assumed that data in the two groups follows a normal distribution. The data is summarized in the table below (unit is log mm precipitation).

	Particles, X (log mm precipitation)	Control, Y (log mm precipitation)
Estimated mean	$\hat{\mu}_X = 1.573$	$\hat{\mu}_Y = 1.314$
Estimated variance	$\hat{\sigma}_X^2 = 0.333$	$\hat{\sigma}_Y^2 = 0.171$
Number of observations	$n_X = 35$	$n_Y = 30$

We now want to test whether the means of the 2 groups can be assumed equal, i.e.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

It is given that the usual test statistics assuming the null hypothesis becomes 2.0958 with 61.19 degrees of freedom. State the p-value and conclusion when a significance level of $\alpha = 0.05$ is applied:

1 ☐ p-value $\simeq 0.82$ i.e. H_0 is accepted

2 ☐ p-value $\simeq 0.41$ i.e. H_0 is rejected

3 ☐ p-value $\simeq 0.21$ i.e. H_0 is accepted

4 ☐ p-value $\simeq 0.10$ i.e. H_0 is rejected

5* ☐ p-value < 0.05 i.e. H_0 is rejected

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

This is a two-sample t-test and we get the information we need from $t_{\text{obs}} = 2.0958$ and degrees of freedom is 61.19, so the p-value is calculated by

```
2 * (1-pt(abs(2.0958), df=61.19))

## [1] 0.04024393
```

which is lower than 0.05, so we reject the null hypothesis.

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

Continues on page 43

Exercise XVIII

At a Christmas marked there is a lottery. 24 balls are placed in bowl. On each of 4 balls there is a picture of a star. On each of the remaining 20 balls there is a picture of an elf. The lottery is now played so that 2 balls are drawn without replacement from the bowl. If both balls show a picture of a star then you have won a prize!

Question XVIII.1 (30)

You participate in the game once. Provide the probability of winning a prize:

1 ☐ $\frac{80}{276}$

2 ☐ $\frac{56}{276}$

3 ☐ $\frac{40}{276}$

4 ☐ $\frac{16}{276}$

5* ☐ $\frac{6}{276}$

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

This is drawing without replacement, hence we must use the hypergeometric distribution (Chapter 2.3.2). However, to get most easily to the answer in the presented form, we can use the basic definition of probability

$$P(\text{success}) = \frac{x}{n},$$

where x is the number of successes in a population of size n. We need possible successful combinations, where a ball with a star is drawn. In the first draw one out of the four must be drawn and in the second draw one out of the three remaining must be drawn, thus

$$x = 4 \cdot 3 = 12.$$

The number of elements in the population (of possible draws) is

$$n = 24 \cdot 23 = 552,$$

since in the first draw there are 24 balls and in the second there are one less. Put together this gives

$$\frac{12}{552} = \frac{6}{276}.$$

Alternatively, the x number of successful combinations could be calculated by

```r
dhyper(x=2, m=4, n=20, k=2)

## [1] 0.02173913
```

which multiplied with the population size gives x

```r
dhyper(x=2, m=4, n=20, k=2) * (24*23)

## [1] 12
```

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

The exam is finished. Have a great Christmas vacation!