

Written examination: 20. August 2017

Course name and number: **Introduction to Statistics (02323 og 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 9 exercises. To answer the questions you need to fill in the prepared 30-question multiple choice form (on 6 separate 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).**

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

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

<b>Exercise</b>	VI.3	VI.4	VI.5	VII.1	VIII.1	VIII.2	VIII.3	IX.1	IX.2	IX.3
<b>Question</b>	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
<b>Answer</b>	5	5	4	2	3	4	3	3	2	3

The questionnaire contains 43 pages.

Continues on page 2

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

**Exercise I**

A swimming team goes on an weekly training camp with a focus on training the swimming stroke front crawl. A test is carried out where the time, for each swimmer swimming the same distance in front crawl, is measured. The test is carried out before and after the camp.

The measured times are stored (in the same order for the swimmers) in the following vectors in R: **before** holds the times before and **after** holds the times after the training camp.

The following hypothesis must be tested

$$\begin{aligned}\mu_{\text{after}} - \mu_{\text{before}} &= 0 \\ \mu_{\text{after}} - \mu_{\text{before}} &\neq 0\end{aligned}$$

where  $\mu_{\text{before}}$  and  $\mu_{\text{after}}$  denotes the mean times for the entire team before and after the camp.

**Question I.1 (1)**

Which of the following R-calls correctly calculates the  $p$ -value for a  $t$ -test of the hypothesis?

- 1  `t.test(after, before, mu=0)`
- 2  `t.test(after, before, mu=-10)`
- 3  `t.test(after, before, mu=10)`
- 4  `t.test(after, mu=10)`
- 5\*  `t.test(after-before, mu=0)`

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

See Section 3.2.3. Since there is a measurement for each swimmer before and after the camp the correct way to analyze the data is a paired  $t$ -test, and they are ordered such that the time for each swimmer is in the same place in **before** as in **after**. The paired analysis is carried out by using a single-sample  $t$ -test on the differences, this is done in Answer 5. Answer 1 to 3 assume independent samples (non-paired) and Answer 4 only test the speed after the camp.

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

**Question I.2 (2)**

The  $p$ -value of the test was calculated to 0.00287. Can the null hypothesis be rejected at significance level  $\alpha = 5\%$  (both conclusion and argument must be correct)?

- 1\*  Yes, since the  $p$ -value is below the significance level the null hypothesis is rejected
- 2  No, since the  $p$ -value is below the significance level the null hypothesis is accepted
- 3  Yes, since the  $p$ -value is over the significance level the null hypothesis is rejected
- 4  No, since the  $p$ -value is over the significance level the null hypothesis is accepted
- 5  More information is needed in order to decide against the null hypothesis

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

Since the  $p$ -value is less than the significance level ( $0.00287 < 0.05$ ), the null hypothesis is rejected (See Method 3.36)

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

**Question I.3 (3)**

Each day at the training camp, there is a random drawing about who should do the dishes. There must be 4 each day for doing the dishes and there are in total 35 participants. For each participant there is equally high probability of being drawn each day. Calculate the probability that a participant will not do the dishes at all during training camp, which includes 7 evenings with dish washing.

- 1   $1 - \binom{7}{0} \cdot 0.144^0 \cdot (1 - 0.144)^{7-0} = 0.57$
- 2   $\binom{5}{2} \cdot 0.144^2 \cdot (1 - 0.144)^{5-2} = 0.09$
- 3   $\binom{7}{7} \cdot 0.798^7 \cdot (1 - 0.798)^{7-7} = 0.21$
- 4\*   $\binom{7}{7} \cdot 0.886^7 \cdot (1 - 0.886)^{7-7} = 0.43$
- 5   $\binom{5}{2} \cdot 0.886^2 \cdot (1 - 0.886)^{5-2} = 0.01$

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

See Definition 2.20. The probability that a participant does not have to do the dishes at a specific day is  $1 - \frac{4}{35} = 0.886$ . Since the probability is the same every day, these must be

independent draws and the probability distribution for the number of times a participant have do the dishes is a binomial with  $n = 7$  and  $p = 0.886$ . Hence the probability can be calculated by

$$\binom{7}{7} \cdot 0.886^7 \cdot (1 - 0.886)^{7-7} = 0.43 \quad (1)$$

or in R by

```
## Sandsynligheden for ikke at bliver trukket alle 7 dage
dbinom(7, 7, 1-4/35)

## [1] 0.4276176
```

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

Continues on page 5

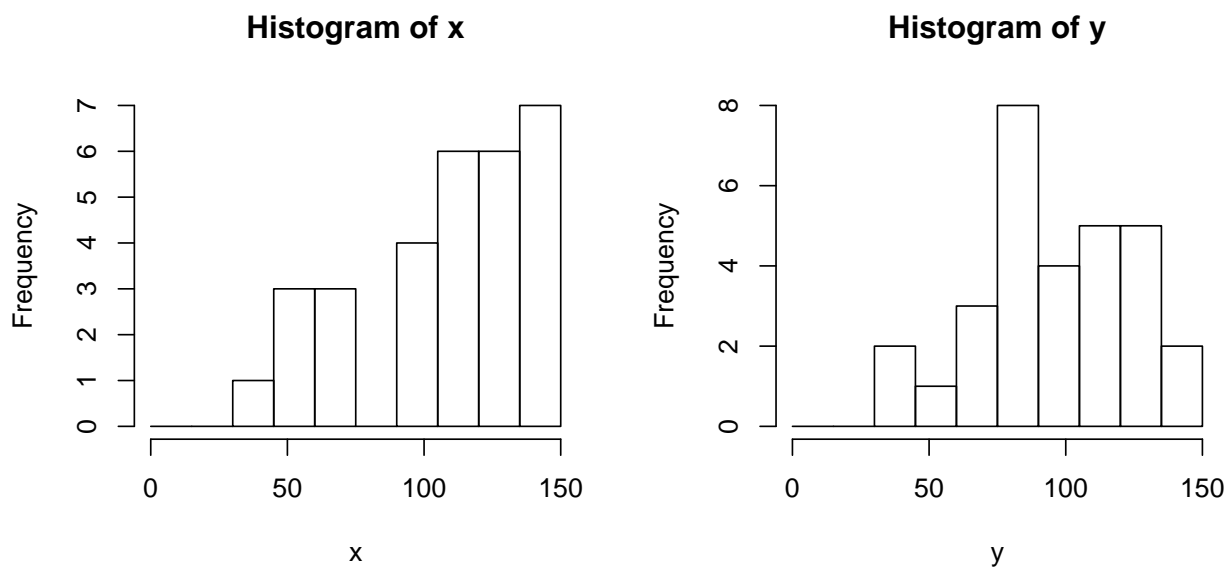
## Exercise II

In connection with the exam in Introduction to Statistics, it is desired to examine whether foreign students are doing well. The score of the exam is calculated as a number between -30 and 150, as there are 30 questions and a wrong answer gives -1 point and a correct answer gives 5 points. There can only be given one answer to each question.

Two random samples of the score has been taken: one for foreign students ( $x$ ) and one for Danish students ( $y$ ). Each sample has 30 observations.

### Question II.1 (4)

To assess the most appropriate analysis, a histogram is plotted of each sample:



What is the most appropriate statement based on the given information?

- 1  Nothing indicates that the samples don't come from symmetrical distributed populations
- 2  The samples cannot be assumed to come from symmetrical distributed populations. This is supported by the histograms, in particular the distribution of  $x$  appears to be right-skewed
- 3\*  The samples cannot be assumed to come from symmetrical distributed population. This is supported by the histograms, in particular the distribution of  $x$  appear to be left-skewed
- 4  The populations from which the samples are taken can both be assumed to be exponentially distributed
- 5  The populations from which the samples are taken can both be assumed to be normally distributed

- The histogram of  $x$  show that the empirical distribution is highly skewed, hence Answer 1 is wrong
- The  $x$  data is left skewed (since the mean is smaller than the median) hence Answer 2 is wrong
- Answer 3 is correct (see the arguments for 1 and 2 being wrong)
- Exponential data is right-skewed, hence Answer 4 is wrong
- Normally distributed data is symmetrical hence 5 is wrong

### Question II.2 (5)

It is decided that the best analysis is included in the following R code:

```
## Number of simulations
k <- 10000
## Simulate each sample k times
simxsamples <- replicate(k, sample(x, replace=TRUE))
simysamples <- replicate(k, sample(y, replace=TRUE))
## Calculate the sample mean differences
simmeandifs <- apply(simxsamples,2,mean) - apply(simysamples,2,mean)
## Quantiles of the differences gives the CI
quantile(simmeandifs, c(0.005,0.995))

## 0.5% 99.5%
## -9.23 31.63

quantile(simmeandifs, c(0.025,0.975))

## 2.5% 97.5%
## -4.125 26.106

## CI for the median differences
simmediandifs <- apply(simxsamples,2,median) - apply(simysamples,2,median)
quantile(simmediandifs, c(0.005,0.995))

## 0.5% 99.5%
## -10.42 43.05

quantile(simmediandifs, c(0.025,0.975))

## 2.5% 97.5%
## -3.975 39.525
```

Which of the following statements is correct?

- 1\*  Non-parametric bootstrap confidence intervals have been calculated for differences between two populations
- 2  Parametric bootstrap confidence intervals have been calculated for differences between two populations
- 3  Confidence intervals for differences between two populations have been calculated under the assumption of normal distributions
- 4  Confidence intervals for differences between two populations have been calculated under the assumption of exponential distributions
- 5  Confidence intervals for differences between two populations have been calculated under the assumption of Poisson distributions

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

The R code calculates non-parametric bootstrap confidence intervals for the differences in scores between the two populations (DK and foreign students). It is done both for the mean (with levels 99% and 95%), and the same confidence intervals for the median. It is non-parametric because no assumption about the distribution is made, which is carried out by sampling directly from the observations with the `sample()` function (instead of e.g. using `rnorm()` which would be under assumption of normal distribution).

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

### Question II.3 (6)

The following hypothesis should be tested at significance level  $\alpha = 5\%$

$$H_0 : q_{0.5,x} = q_{0.5,y}$$

$$H_1 : q_{0.5,x} \neq q_{0.5,y}$$

where  $q_{0.5,x}$  denotes the 50% quantile for foreign students and  $q_{0.5,y}$  denotes the 50% quantile for Danish students.

Which of the following statements is correct (not all of the statements are necessarily meaningful)?

- 1   $H_0$  is rejected and it can be concluded that Danish students perform significantly better than foreign students at the indicated level of significance
- 2   $H_0$  is rejected and it can be concluded that foreign students perform significantly better than Danish students at the indicated level of significance

- 3\*   $H_0$  is not rejected and it cannot be concluded that Danish students perform significantly different than foreign students at the indicated level of significance
- 4   $H_0$  is not rejected and it can be concluded that Danish students perform significantly different than foreign students at the indicated level of significance
- 5  None of the above statements are correct

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

The R code for the previous question gives the 95% confidence interval for the median as  $[-3.975; 39.525]$ , hence the null hypothesis is not rejected at level  $\alpha = 0.05$  since  $(0 \in [-3.975; 39.525])$ . It's close to the boundary, so maybe with a slightly larger sample we would have concluded a significant difference, but we have not and must accept that.

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

Continues on page 9

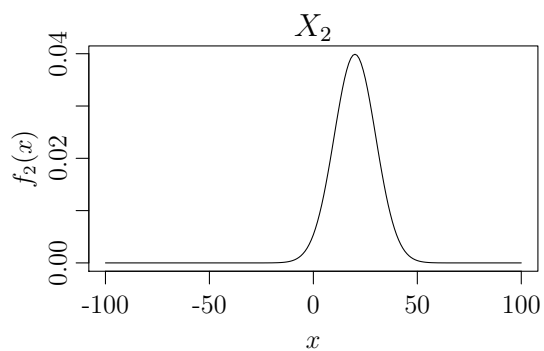
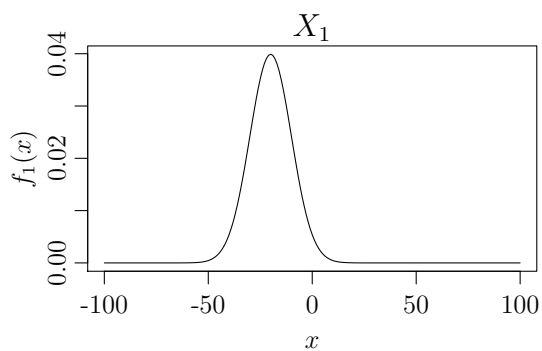


**Exercise III**

Let two independent random variables be given by

$$X_1 \sim N(-20, 10^2) \quad \text{and} \quad X_2 \sim N(20, 10^2).$$

Their probability density functions (pdfs) are then:

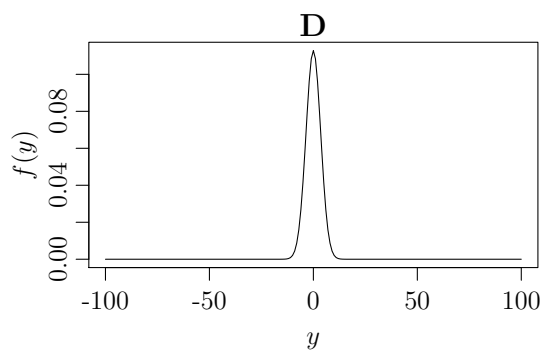
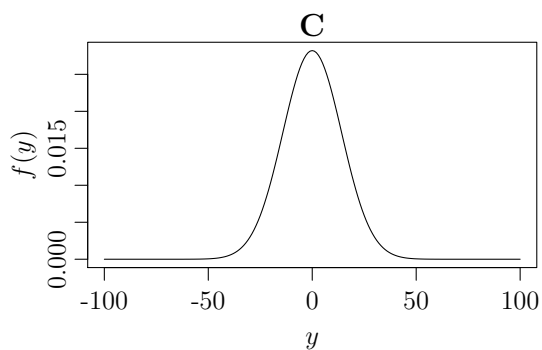
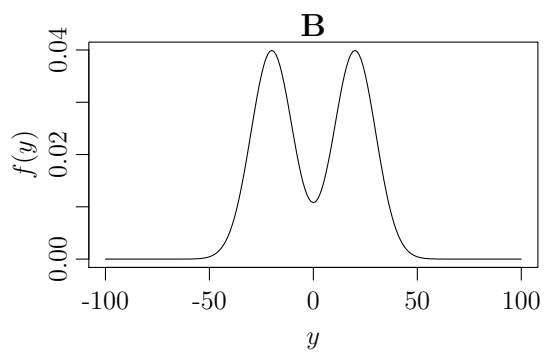
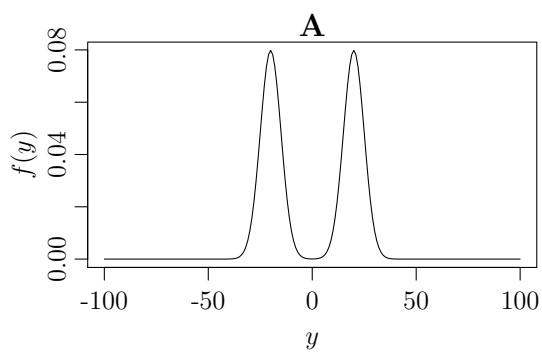


**Question III.1 (7)**

Now a new random variable is defined by

$$Y = X_1 + X_2.$$

Which of the following plots is then the pdf for  $Y$ ?



- 1  Plot A
- 2  Plot B
- 3\*  Plot C
- 4  Plot D
- 5  None of the shown plots can be close to the pdf of  $Y$

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

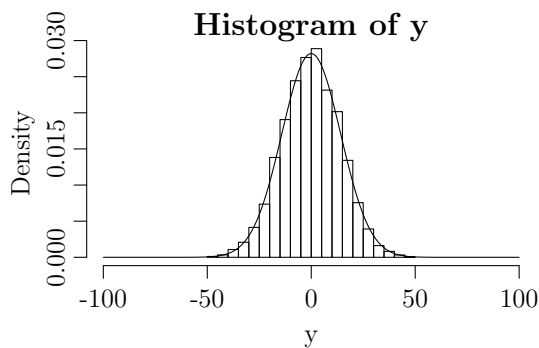
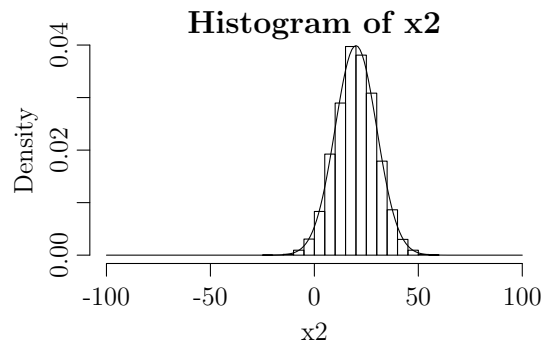
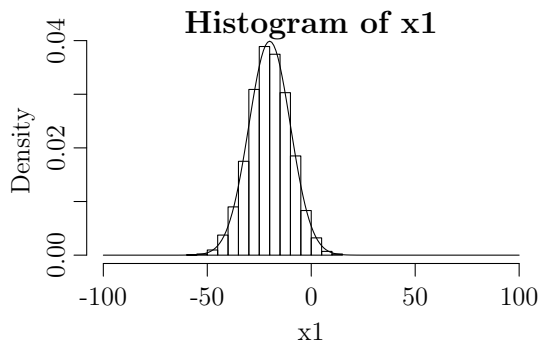
First of all, as stated in Theorem 2.40, sums of independent normal random variables are normal random variables (this exclude plots A and B). Also from Theorem 2.56 we have that

$$E[Y] = E[X_1] + E[X_2] = 0,$$

$$V[Y] = V[X_1] + V[X_2] = 200.$$

The variance of  $Y$  is greater than the variance of  $X_1$  and  $X_2$  hence we can exclude Plot D, and the only remaining option is Plot C that have a larger variance the  $X_1$  and  $X_2$ , hence plot C is correct. This can be confirmed with simulation:

```
## (x1+x2) = var(x1) + var(x2) = 100 + 100 = 200
x1 <- rnorm(n, mean=-20, sd=10)
x2 <- rnorm(n, mean=20, sd=10)
y <- x1 + x2
par(mfrow=c(2,2), mgp=c(1.6,0.5,0), mar=c(4,3,1,1), tcl=-0.4)
hist(x1, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=-20,sd=10), type="l")
hist(x2, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=20,sd=10), type="l")
hist(y, xlim=c(xmin,xmax), prob=TRUE)
lines(xseq, dnorm(xseq,mean=0,sd=sqrt(200)), type="l")
```



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

**Question III.2 (8)**

Assuming  $X_1$  and  $X_2$  each represent a population and the test for difference in mean value with the commonly used non-paired  $t$ -test should be carried out. What is the smallest sample size  $n = n_1 = n_2$  that must be taken from each population, at significance level  $\alpha = 5\%$ , in order to achieve a power of the test of at least 99%?

- 1\*   $n = 4$  observations in each sample
- 2   $n = 12$  observations in each sample
- 3   $n = 38$  observations in each sample
- 4   $n = 69$  observations in each sample
- 5   $n = 248$  observations in each sample

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

The difference in mean of the two distributions is 40, and the standard deviation in each of the two groups is 10, hence we can find the the number of observation needed with:

```
power.t.test(delta=40, sd=10, sig.level=0.05, power=0.99)

##
##      Two-sample t test power calculation
##
##              n = 3.644287
##             delta = 40
##              sd = 10
##      sig.level = 0.05
##              power = 0.99
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

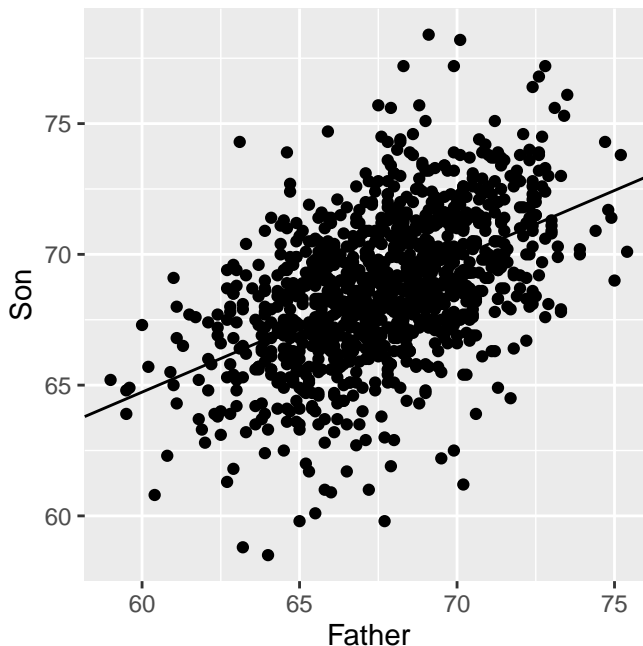
hence the correct answer is  $n = 4$ , since we must round up to nearest integer. See more in Section 3.3

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

Continues on page 13

### Exercise IV

The figure below shows the relation between the height of about 1000 fathers and their sons measured in inches:



The shown regression line describes the fit of the following model

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.},$$

where  $Y_i$  is the height of the  $i$ 'th son and  $x_i$  is the height of the  $i$ 'th father.

#### Question IV.1 (9)

Which of the following statements is a correct description of the regression line?

- 1  The line describes an estimate of the mean height of the sons as a function of their fathers mean height
- 2  The line describes an estimate of the linear correlation between the average height of father and son
- 3  The line describes an estimate of the fathers mean height as a function of the height of their sons
- 4\*  The line describes an estimate of the sons mean height as a function of the height of their fathers
- 5  The line describes the height of a son as a function of the height of the father

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

- The regression line show an estimate of the mean (or expected) height of sons as a function fathers height (not mean height) hence Answer 1 is wrong
- The estimation of correlation cannot be directly derived from the regression line, hence Answer 2 is wrong
- In Answer 3 the relation is reversed (hence it is wrong)
- Answer 4 correctly states that the line describe the mean (or expected) height of sons as a function of fathers heights
- The line only describe an estimate of the sons' mean height (hence the points are scattered around the line), not the actual height. So answer 5 is wrong.

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

### Question IV.2 (10)

It is chosen to analyze the data with the following R code, where `fs` is a data frame with the columns `Son` and `Father` holding the observed heights:

```
summary(fit <- lm(Son ~ Father, data=fs))
```

Which gives the following result where two numbers are replaced by letters:

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min       1Q   Median       3Q      Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280          A   18.49  <2e-16 ***
Father       0.51401          B   19.00  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512, Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

How large a proportion of the variation in the height of sons is not explained by the height of the fathers?

- 1  Approximately 25%
- 2  Approximately 50%
- 3\*  Approximately 75%
- 4  Approximately 86.5%
- 5  Approximately 66%

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

The proportion of explained variation is the multiple  $R^2$  value, which can be read as 0.2512 or approximately 25% from the R output, hence the variation not explained will is approximately 75%.

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

Continues on page 16

### Question IV.3 (11)

What is the estimate of the standard deviation of the coefficient for **Father**?

- 1   $\hat{\sigma}_{\beta_1} = 0.514/2.438 = 0.211$
- 2   $\hat{\sigma}_{\beta_1} = 2.438/1076 = 0.00227$
- 3   $\hat{\sigma}_{\beta_1} = 0.514 \cdot 19.00 = 9.77$
- 4   $\hat{\sigma}_{\beta_1} = 33.89/18.49 = 1.83$
- 5\*   $\hat{\sigma}_{\beta_1} = 0.514/19.0 = 0.027$

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

From Theorem 5.12 we have the formula for the test statistic

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

and we know that the default  $t$ -test printed out by `summary()` is with the null hypothesis that the slope is 0:

$$H_0 : \beta_{0,1} = 0$$

so we can write

$$t_{\beta_1, \text{obs}} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}},$$

where  $t_{\beta_1, \text{obs}}$  is the observed  $T_{\beta_1}$  which is seen from the R output to be 19.0.

Rearranged it gives and values found in the result

$$\hat{\sigma}_{\beta_1} = \frac{\hat{\beta}_1 - 0}{t_{\beta_1, \text{obs}}} = \frac{0.514}{19.0} = 0.027.$$

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

### Question IV.4 (12)

Given the following calculations in R, what is a 95% confidence interval for the mean height of sons of fathers who are 75 inches tall?



```

mean(fs$Father); var(fs$Father)

## [1] 67.68683
## [1] 7.539566

mean(fs$Son); var(fs$Son)

## [1] 68.68423
## [1] 7.930949

```

$$\begin{aligned}
1^* \square & 33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}} \\
2 \square & 33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540}} \\
3 \square & 33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}} \\
4 \square & 33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}} \\
5 \square & 33.893 + 0.514 \cdot 75 \pm 1.65 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}}
\end{aligned}$$

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

The general formula for the confidence interval for a point on the line is given in Method 5.18

$$\beta_0 + \beta_1 x_{\text{new}} \pm t_{\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$$

we can find  $\beta_0 = 33.89$ ,  $\beta_1 = 0.514$ ,  $\sigma = 2.438$ , and  $n = 1076 + 2 = 1078$  in the summary output above. In addition we got  $\bar{x} = 67.69$  in the calculation above, note also that  $S_{xx} = (n - 1)s_x^2$  where  $s_x^2$  is the empirical variance of farthers height, hence  $S_{xx} = 1077 \cdot 7.540$ . The degrees of freedom for the  $t$ -distribution is  $n - 2 = 1076$  which gives  $t_{0.975} = 1.96$  Inserting the numbers we get Answer 1 as the correct answer. This can also be checked by the following R calculations.

```

n <- 1078
sxx <- 7.540 * (n-1)
xnew=75
xbar <- 67.687
sigma <- 2.438
beta0 <- 33.893
beta1 <- 0.514

round(beta0 + beta1*xnew + c(-1, 1) * qt(0.975, df=n-2) * sigma *
  sqrt((1/n) + ((xnew - xbar)^2/sxx)), 2)

```

```
## [1] 72.03 72.86

round(predict(fit, data.frame("Father"=75), interval="conf"), 2) # Check

##      fit   lwr   upr
## 1 72.44 72.03 72.86
```

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

### Question IV.5 (13)

Now information about each family's monthly income is obtained and the following model is setup

$$Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.,}$$

where  $Y_i$  is the height of the  $i$ 'th son,  $x_{1,i}$  is height of the  $i$ 'th father, and  $x_{2,i}$  is the income for the  $i$ 'th family.

Under the following two assumptions:

- Rich families eat better and a better diet has a significant positive effect, which gives the sons of the family a higher growth
- There is independence between the father's height and the family's income

what is the consequence of adding the income into the model (not all answers are necessarily meaningful)?

- 1\*  Inclusion of income in the model will contribute to reducing the residual variance ( $\hat{\sigma}^2$ ) and the uncertainty of the regression coefficient for the father's height ( $\beta_1$ ) will be reduced
- 2  Inclusion of income in the model will contribute to reducing the residual variance ( $\hat{\sigma}^2$ ), but this will not affect the uncertainty of the regression coefficient for the father's height ( $\beta_1$ )
- 3  As the fathers height is independent of the fathers income, the inclusion of income in the model will not affect the estimate of  $\beta_1$  or the uncertainty of it
- 4  Inclusion of income in the model will use one more degree of freedom, such that a confidence interval for  $\beta_1$  may be expected to be wider than if incomes were not included in the model
- 5  One must expect a high degree of multicollinearity between the estimates of  $\beta_1$  and  $\beta_2$ , so the model must be reduced to a simple linear regression model

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

Lets go through the possibilities one by one:

- Answer 1: Including an effect that has a significant effect will reduce the residual variation, and with independence between fathers height and family income the uncertainty for  $\beta_1$  will be reduced. Hence 1 is correct

- Answer 2: The residual variation has a direct effect on the uncertainty of the parameters, hence 2 cannot be correct
- Answer 3: With the argument in ans 2, this cannot be correct either
- Answer 4: The effect of using a degree of freedom is very small, hence the effect of reducing the variance will dominate, hence 4 is not correct
- Answer 5: Since we assume that these are independent we will not expect multicollinearity

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

Continues on page 21

### Exercise V

In humans there are a variety of different genetic determined blood type systems. The most well-known are probably the ABO- and Rhesus-systems. Another blood type system is the so-called MN blood type system, which is determined by a single gene Glycophorin A (GPA). In the GPA-gene there are two alleles M and N, such that a human may have the genotype (blood type) MM, MN, or NN.

The distribution of blood types in the MN blood type system is now sought estimated from a sample of volunteer students of two different Philippine universities. One university, University of the Philippines-Diliman, here shortened UPD, is the country's largest university where students come from all over the country. The second university, Isabela State University, here abbreviated ISU, is a small university where the students primarily come from the local area. The following table lists the distribution of genotypes among the students in the samples from the two universities:

Bloodtype	UDP	ISU
MM	19	43
MN	15	7
NN	17	9

#### Question V.1 (14)

State the  $\chi^2$  test statistic and the conclusion of the test in which the MN blood type distribution in the two universities are compared (both test size and conclusion must be correct).

- 1  The test statistic is  $\chi^2 = 14.15$ , its distribution has 2 degrees of freedom and the test shows that there is some evidence for a difference in the MN blood type distribution at the two universities
- 2\*  The test statistic is  $\chi^2 = 14.15$ , its distribution has 2 degrees of freedom and the test shows that there is very strong evidence for a difference in the MN blood type distribution at the two universities
- 3  The test statistic is  $\chi^2 = 3.76$ , its distribution has 2 degrees of freedom and the test shows that there is not found any evidence for a difference in the MN blood type distribution at the two universities
- 4  The test statistic is  $\chi^2 = 4.57$ , its distribution has 1 degree of freedom and the test shows that there is evidence for a difference in the MN blood type distribution at the two universities
- 5  The test statistic is  $\chi^2 = 3.76$ , its distribution has 1 degree of freedom and the test shows that there is weak evidence for a difference in the MN blood type distribution at the two universities

The easiest way to solve this is by using `chisq.test`:

```
## Define the table
Bloodtype <- matrix(c(19,43,15,7,17,9),nrow=3,byrow=T)
colnames(Bloodtype) <- c("UDP", "ISU")
rownames(Bloodtype) <- c("MM", "MN", "NN")
## Answer:
chisq.test(Bloodtype)

##
## Pearson's Chi-squared test
##
## data: Bloodtype
## X-squared = 14.154, df = 2, p-value = 0.0008443
```

We can see that the test-statistics is 14.15 and the degrees of freedom is 2, the  $p$ -value is 0.00084, and hence very strong evidence against the null hypothesis. The null hypothesis being that there is no difference in the distribution of blood cells between the universities. This is Answer 2.

We could also solve this by “hand”-calculations

```
## By hand:
mat <- Bloodtype
Exp <- rowSums(mat) %o% colSums(mat) / sum(mat)
Chisq.val <- sum((mat - Exp)^2 / Exp)
df <- prod(dim(mat)) - 1
pchisq(Chisq.val, df, lower=FALSE)

## [1] 0.0008443032
```

### Question V.2 (15)

A biological population is said to be in Hardy-Weinberg (HW) equilibrium if the proportion of genotypes can be written as

$$\begin{aligned} p_{MM} &= p^2, \\ p_{MN} &= 2pq, \\ p_{NN} &= q^2. \end{aligned}$$

Where  $p$  and  $q$  are the allele frequencies for M and N, respectively. They are calculated by

$$p = \frac{2 \cdot X_{MM} + X_{MN}}{2n},$$

$$q = \frac{2 \cdot X_{NN} + X_{MN}}{2n},$$

where  $X_{\text{bloodtype}}$  is the observed number of the blood type and  $n$  is the sample size. Thus for UDP

$$p_{MN} = 2 \cdot \frac{2 \cdot X_{MM} + X_{MN}}{2n} \cdot \frac{2 \cdot X_{NN} + X_{MN}}{2n} = 0.4992,$$

is set as the proportion of MN blood type under HW-equilibrium.

A simple test to decide whether the population on UDP is not in HW-equilibrium can therefore be of the hypothesis

$$H_0 : p_{MN,UDP} = 0.4992$$

$$H_1 : p_{MN,UDP} \neq 0.4992$$

i.e. if the observed proportion of MN blood type on UDP  $p_{MN,UDP}$  is equal to the proportion under HW-equilibrium.

We want to test whether it can be rejected that the genotypes on UDP are in HW-equilibrium. What is the usually applied test statistic for this test?

- 1  The test statistic is  $\chi^2 = 2(1.99 + 4.30 + 2.32) = 17.2$
- 2\*  The test statistic is  $z_{\text{obs}} = \frac{15 - 25.46}{\sqrt{25.46 \cdot (1 - \frac{25.46}{51})}} = -2.93$
- 3  The test statistic is  $z_{\text{obs}} = \frac{15 - 51}{\sqrt{51 \cdot (1 - \frac{15}{51})}} = -6.00$
- 4  The test statistic is  $\chi^2 = (1.99^2 + 4.30^2 + 2.32^2)/2 = 13.9$
- 5  The test statistic is  $\chi^2 = 1.99 + 4.30 + 2.32 = 8.6$

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

This is a large sample test for a single proportion so we use the standard normal distribution (as stated in Theorem 7.10)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

and the values are found for UDP and inserted

$$np_0 = 0.4992 \cdot 51 = 25.46,$$

$$z_{\text{obs}} = \frac{15 - 25.46}{\sqrt{25.46(1 - \frac{25.46}{51})}} = \frac{15 - 25.46}{\sqrt{25.46(1 - \frac{25.46}{51})}} = -2.93.$$

**Question V.3 (16)**

For another type of test for HW-equilibrium the test statistic is found to  $\chi^2 = 24.52$  and under the null hypothesis it will follow a  $\chi^2$ -distribution with 1 degree of freedom. What is the  $p$ -value and conclusion of the test using a significance level of 0.001?

- 1   $p$ -value is `pchisq(24.52, df=1)  $\approx$  1` and the hypothesis of HW-equilibrium cannot be rejected
- 2   $p$ -value is `1 - pchisq(24.52, df=1) < 0.001` and the hypothesis of HW-equilibrium cannot be rejected
- 3\*   $p$ -value is `1 - pchisq(24.52, df=1) < 0.001` and the hypothesis of HW-equilibrium is rejected
- 4   $p$ -value is `1 - pnorm(sqrt(24.52)) < 0.001` and the hypothesis of HW-equilibrium cannot be rejected
- 5   $p$ -value is `1 - pnorm(sqrt(24.52)) < 0.001` and the hypothesis of HW-equilibrium is rejected

The solution is to find that the correct  $p$ -value is calculated in R by:

```
## Either
pchisq(24.52, df=1, lower.tail=FALSE)

## [1] 7.354249e-07

## Or
1 - pchisq(24.52, df=1)

## [1] 7.354249e-07
```

so the  $p$ -value is below 0.001 and thus the conclusion is that the null hypothesis is rejected.

**Question V.4 (17)**



For theoretical reasons it has been suggested that the frequencies of genotypes MM and NN in the underlying population are the same and it is now of interest to investigate this on the basis of the observations from UDP. Assuming that the proportions for MM and NN are independent a 90% confidence interval for the difference in the proportion of MM and NN ( $p_{MM,UDP} - p_{NN,UDP}$ ) is given by:

$$1^* \square 2/51 \pm 1.64 \sqrt{\frac{19 \cdot 32}{51^3} + \frac{17 \cdot 34}{51^3}}$$

$$2 \square 2/51 \pm 1.96 \sqrt{\frac{19 \cdot 32}{51^3} + \frac{17 \cdot 34}{51^3}}$$

$$3 \square 2/51 \pm 1.64 \sqrt{\frac{19 \cdot 32}{51^2} + \frac{17 \cdot 34}{51^2}}$$

$$4 \square 2/51 \pm 1.96 \sqrt{\frac{19 \cdot 32}{51^2} + \frac{17 \cdot 34}{51^2}}$$

$$5 \square 2/51 \pm 1.68 \sqrt{\frac{19 \cdot 32}{51^3} + \frac{17 \cdot 34}{51^3}}$$

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

We use Theorem 7.15 and insert the values:

```
## Answer ("manually"):
sd.pid <- sqrt(19*32/51^3 + 17*34/51^3)
CI <- (19/51 - 17/51) + c(-1, 1) * qnorm(0.95) * sd.pid
round(CI, 3)

## [1] -0.116  0.195
```

Notice that  $\sqrt{\frac{19 \cdot 32}{51^3} + \frac{17 \cdot 34}{51^3}} = \sqrt{\frac{19}{51} \left(1 - \frac{19}{51}\right) + \frac{17}{51} \left(1 - \frac{17}{51}\right)}$

Alternative we use the `prop.test` in R as shown in example 7.19

```
## Answer using prop.test
event <- c(19,17)
n <- c(51,51)
prop.test(event, n, conf.level=0.90, correct=FALSE)

##
## 2-sample test for equality of proportions without continuity
## correction
##
## data:  event out of n
## X-squared = 0.17172, df = 1, p-value = 0.6786
```

```
## alternative hypothesis: two.sided
## 90 percent confidence interval:
## -0.1163144  0.1947457
## sample estimates:
##   prop 1   prop 2
## 0.3725490 0.3333333
```

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

Continues on page 27

### Question V.5 (18)

What is the usual test statistic for the test that the proportions of MM and NN are equal at UDP, i.e.  $H_0 : p_{MM,UDP} = p_{NN,UDP}$  ?

1   $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{19 \cdot 32}{51^2} + \frac{17 \cdot 34}{51^2}}}$

2   $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19 \cdot 32}{51^3} + \frac{17 \cdot 34}{51^3}}}$

3\*   $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{6}{17} \frac{11}{17} \frac{2}{51}}}$

4   $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{6}{17} \frac{6}{19} \frac{2}{51}}}$

5   $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19 \cdot 34}{51^3} + \frac{17 \cdot 32}{51^3}}}$

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

Use Method 7.18 and insert the values:

```
x1 <- 17;
x2 <- 19;
n1 <- n2 <- 15+19+17
p1 <- 17/n1
p2 <- 19/n2
(delta.p <- p2 - p1)

## [1] 0.03921569

(phat <- (17+19)/(51+51))

## [1] 0.3529412

(zobs <- delta.p / sqrt(phat*(1-phat)*(1/n1 + 1/n2)))

## [1] 0.4143877

## The formula from the answer give the same answer
2/51/sqrt((6/17)*(11/17)*(2/51))

## [1] 0.4143877
```

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

Continues on page 28

## Exercise VI

A sample with the following 10 observations is taken:

```
x <- c(-1.63, -1.37, -1.21, -0.60, -0.36, -0.26, -0.18, 0.02, 0.29, 0.39)
```

Notice that the observations have been sorted in the code above.

The sample mean and sample standard deviation are calculated:

```
mean(x)
## [1] -0.491

sd(x)
## [1] 0.7003
```

### Question VI.1 (19)

What is the sample variance?

- 1   $s^2 = 0.21$
- 2\*   $s^2 = 0.49$
- 3   $s^2 = 1.46$
- 4   $s^2 = 1.70$
- 5   $s^2 = 2.36$

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

The sample variance is simply the squared standard deviation (Definition 1.11)  $s^2 = 0.07^2 = 0.49$ , or in R:

```
var(x)
## [1] 0.4904
```

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

### Question VI.2 (20)

What is the first quartile of the sample?

- 1   $Q_1 = -1.37$
- 2   $Q_1 = -1.29$
- 3\*   $Q_1 = -1.21$
- 4   $Q_1 = -0.91$
- 5   $Q_1 = -0.60$

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

See Definition 1.7. With  $n=10$  observations we get  $pn = 0.25 \cdot 10 = 2.5$  and hence the first quartile is  $Q_1 = x_{(3)} = -1.21$ .

In R we can get this by:

```
quantile(x, type=2,prob=0.25)
## 25%
## -1.21
```

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

### Question VI.3 (21)

Which of the following is a correct 95% confidence interval for the mean of the population from which the sample is taken?

- 1   $-0.491 \pm t_{0.975} \frac{0.490}{\sqrt{10}} = [-0.84, -0.14]$  where  $t_{0.975} = 2.26$  is a quantile in  $t$ -distribution with 9 degrees of freedom
- 2   $-0.491 \pm t_{0.95} \frac{0.700}{\sqrt{9}} = [-0.92, -0.64]$  where  $t_{0.95} = 1.83$  is a quantile in  $t$ -distribution with 9 degrees of freedom
- 3   $-0.491 \pm t_{0.95} \frac{0.490}{\sqrt{9}} = [-0.79, -0.19]$  where  $t_{0.95} = 1.83$  is a quantile in  $t$ -distribution with 9 degrees of freedom
- 4   $-0.491 \pm t_{0.975} \frac{0.700}{10} = [-0.65, -0.33]$  where  $t_{0.975} = 2.26$  is a quantile in  $t$ -distribution with 9 degrees of freedom

5\*   $-0.491 \pm t_{0.975} \frac{0.700}{\sqrt{10}} = [-0.99, 0.01]$  where  $t_{0.975} = 2.26$  is a quantile in  $t$ -distribution with 9 degrees of freedom

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

As stated in Method 3.9, the 95% confidence interval is given by

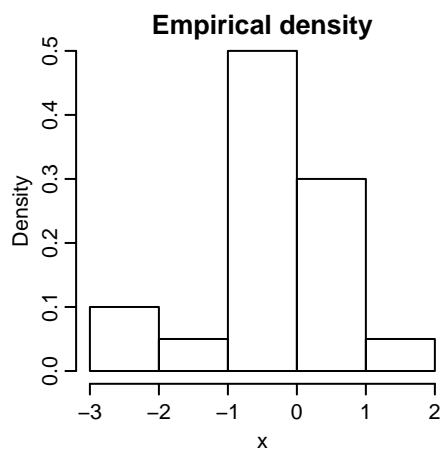
$$\bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

with the numbers given above the correct is Answer 5.

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

### Question VI.4 (22)

Another sample is taken and its empirical density is:



What is the size of the sample, i.e. how many observations  $n$  are in the sample?

1  20

2  30

3  100

4  300

5\*  This question cannot be answered with the given information

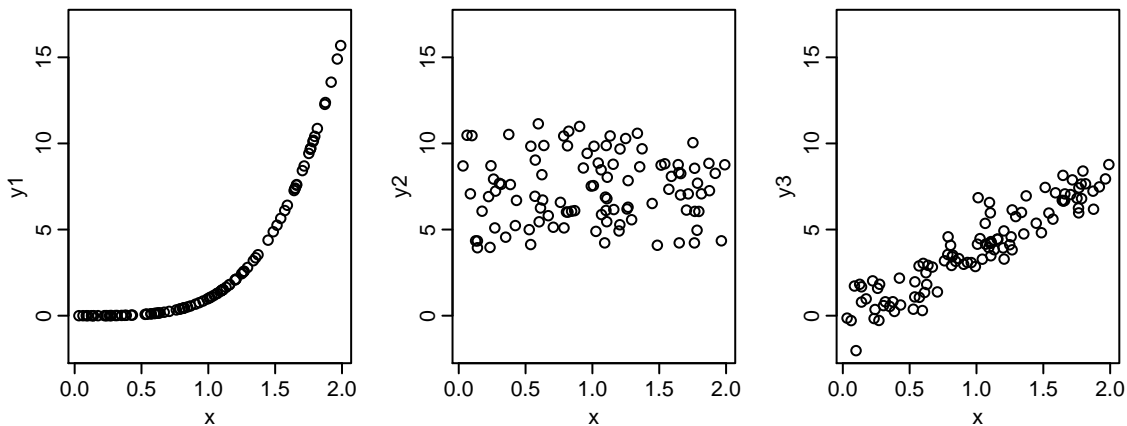
----- FACIT-BEGIN -----

Since we only have the relative frequencies (the empirical density), we cannot know how many observations are in the sample.

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

**Question VI.5 (23)**

The following three plots are of coherent values of  $x$  and  $y$  for samples from three different populations:



The following statements are about the correlations of the three populations from which the samples were taken. Which of the statements is not very unlikely?

- 1   $\rho_{XY_1} = 0, \rho_{XY_2} = 0$  and  $\rho_{XY_3} = 0.33$
- 2   $\rho_{XY_1} = 0, \rho_{XY_2} = 0$  and  $\rho_{XY_3} = -0.89$
- 3   $\rho_{XY_1} = 0, \rho_{XY_2} = 0.61$  and  $\rho_{XY_3} = 0.91$
- 4\*   $\rho_{XY_1} = 0.87, \rho_{XY_2} = 0$  and  $\rho_{XY_3} = 0.92$
- 5   $\rho_{XY_1} = 0.22, \rho_{XY_2} = 0$  and  $\rho_{XY_3} = -0.34$

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

From the plots we can see that  $\rho_{XY_1} > 0, \rho_{XY_2} \approx 0, \rho_{XY_3} > 0$ , since there is a positive correlation between  $x$  and  $y_1$  and  $x$  and  $y_3$ , but no visible correlation between  $x$  and  $y_2$ , hence the only plausible is Answer 4.

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

Continues on page 32

**Exercise VII**

In a finite population of  $N$  units with mean  $E[Y] = \mu$  and variance  $V[Y] = \sigma^2$  we are considering a sample with  $n$  units  $Y_i, i = 1, \dots, n$ . If the sample is taken randomly and without replacement, then the sample mean is  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  and the variance is  $V(\bar{Y}) = \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n}$ . The interest is now on the sum of the sample  $\tau = \sum_{i=1}^n Y_i$ , which can be estimated by  $\hat{\tau} = \frac{N}{n} \sum_{i=1}^n Y_i$ .

**Question VII.1 (24)**

What is the variance of the estimator  $\hat{\tau}$  i.e.  $V(\hat{\tau})$ ?

1   $V(\hat{\tau}) = \frac{N^2}{n} \sigma^2$

2\*   $V(\hat{\tau}) = N(N - n) \frac{\sigma^2}{n}$

3   $V(\hat{\tau}) = \frac{N^2}{n^3} \sigma^2$

4   $V(\hat{\tau}) = N^2(1 - n) \sigma^2$

5   $V(\hat{\tau}) = \frac{N}{n} \sigma^2$

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

Since

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y},$$

where  $\bar{y}$  symbolizes the sample mean of  $y$  we can bring out the  $N$  and square it (as described in Theorem 2.54)

$$\begin{aligned} V[\hat{\tau}] &= V[N\bar{y}] \\ &= N^2 V[\bar{y}] \\ &= N^2 \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n} \\ &= N(N-n) \frac{\sigma^2}{n}. \end{aligned}$$

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

Continues on page 33



## Exercise VIII

Up until the 1970s in Finland, it was only allowed to sell and serve alcoholic beverages in towns and not in rural areas. When it was wanted to ease the restrictions on alcohol sale in rural areas it raised concerns if this would lead to an increased rate of road accidents. Ahead of easing the restrictions a project was carried out in which: four rural municipalities were granted extraordinary permission to sell alcohol in shops, and four other municipalities were granted permission to, besides selling alcohol in shops, serve alcohol in restaurants and others serving places. Finally, four other rural municipalities without extraordinary permits acted as control. Data on the number of traffic accidents from the 12 selected municipalities over the year the project ran is presented in the following table:

Name	Control	Sale	SaleAndServing
	177	226	226
	225	196	229
	167	198	215
	176	206	188
Sum	745	826	858

and the chosen analyses is an ANOVA. The result is:

```
## Analysis of Variance Table
##
## Response: Accidents
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  A 1696.2  848.08      C      D
## Residuals  B 3670.7  407.86
```

Where **Treatment** is a factor dividing the municipalities into the three groups and **Accidents** is the number of accidents.

### Question VIII.1 (25)

To investigate whether the permission to sell alcohol has an effect on the rate of traffic accidents, the average number of traffic accidents in the 3 groups are compared. Assuming that the variance in the number of traffic accidents is constant between the groups, what is then the result of the test for a difference in the mean number of traffic accidents between the 3 groups on significance level  $\alpha = 0.05$ ?

- The test statistic  $F_{\text{obs}} = 1.232$  which under  $H_0$  follows an  $F$ -distribution with 3 and 8 degrees of freedom, gives a  $p$ -value of 0.360 and the study therefore gives no reason to believe that an easing of alcohol restrictions will increase number of traffic accidents
- The test statistic  $F_{\text{obs}} = 2.079$  which under  $H_0$  follows an  $F$ -distribution with 2 and 9 degrees of freedom, gives a  $p$ -value of 0.181 and the study therefore shows that easing of alcohol restrictions will certainly lead to an increase in the number of traffic accidents

- 3\*  The test statistic  $F_{\text{obs}} = 2.079$  which under  $H_0$  follows an  $F$ -distribution with 2 and 9 degrees of freedom, gives a  $p$ -value of 0.181 and the study therefore gives no reason to believe that an easing of alcohol restrictions will increase number of traffic accidents
- 4  The test statistic  $F_{\text{obs}} = 4.324$  which under  $H_0$  follows an  $F$ -distribution with 2 and 9 degrees of freedom, gives a  $p$ -value of 0.0434 and the study therefore shows that easing of alcohol restrictions will lead to a change of the number of traffic accidents
- 5  The test statistic  $F_{\text{obs}} = 4.324$  which under  $H_0$  follows an  $F$ -distribution with 3 and 8 degrees of freedom, gives a  $p$ -value of 0.0434 and the study therefore shows that easing of alcohol restrictions will lead to a change of the number of traffic accidents

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

This is a one-way ANOVA. First note that the degrees of freedom are  $A = 3 - 1 = 2$ , and  $B = 4 \cdot 3 - 3 = 9$ , these are the degrees of freedom needed to calculate the  $p$ -value. Now let's calculate the numbers C and D as described in Method 8.6

```
(F <- 848.08 / 407.86)
## [1] 2.079341
(p.value <- 1 - pf(F,df1=2,df2=9))
## [1] 0.1809821
```

Hence the only possible correct answers are 2 and 3, but in Answer 2 a wrong conclusion is drawn (with the given  $p$ -value=0.181 > 0.05) while Answer 3 correctly states that there is not evidence against the null-hypothesis (that the number of accident will not increase).

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

### Question VIII.2 (26)

What is the estimate of the standard deviation of the errors?

- 1   $\hat{\sigma} = 1696.2/(12 - 1) = 154$
- 2   $\hat{\sigma} = \sqrt{1696.2/(3 - 1)} = 29.1$
- 3   $\hat{\sigma} = \sqrt{1696.2/(12 - 1)} = 11.0$
- 4\*   $\hat{\sigma} = \sqrt{3670.7/(12 - 3)} = 20.2$
- 5   $\hat{\sigma} = 5367.1/(12 - 3)^2 = 66.3$

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

See the end of Chapter 8.2.2. Since  $MSE = s^2$ , the number can be calculate directly from the ANOVA table as  $\hat{\sigma} = \sqrt{407.86} = 20.2$ .

Or more in detail R:

```
(SSE <- (4-1) * (687.58 + 187.666667 + 348.3333)) # Eq 8-6
## [1] 3670.74

(MSE <- SSE/(n-k))
## [1] 407.86

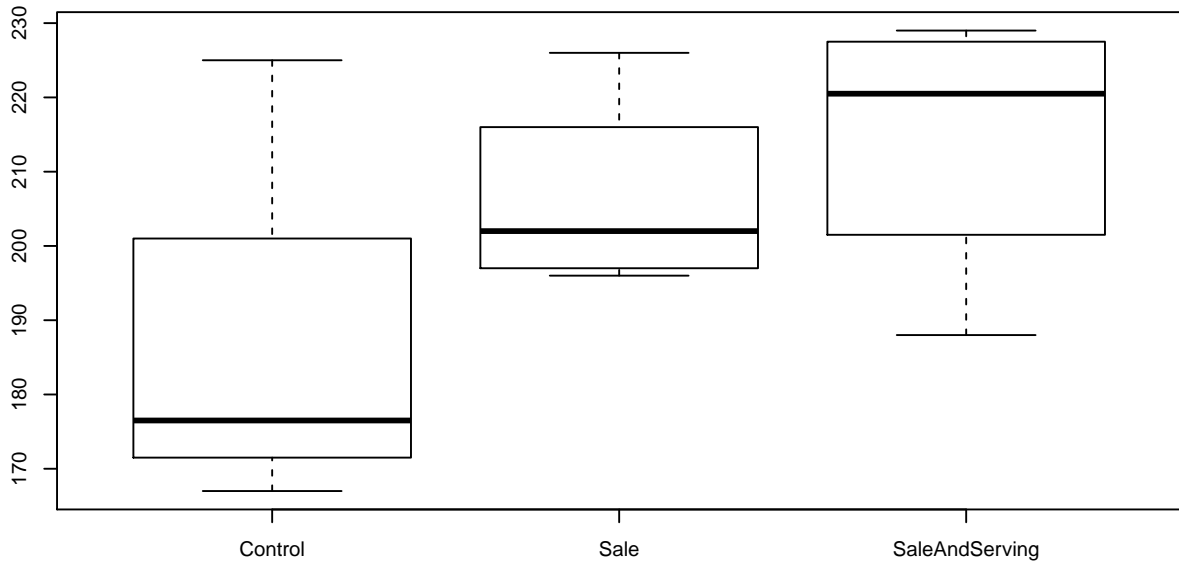
## The answer
sqrt(MSE)
## [1] 20.19554

sqrt(3670.7/(12-3))
## [1] 20.19543
```

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

### Question VIII.3 (27)

The assumption of homogeneous variance is validated with the following box plots:



Which of the following statements is the most correct conclusion based on this plot and the informations given (not all the statements are necessarily meaningful)?

- 1  Taking the high number of observation into account there is no evidence that the assumption of homogeneous variance is not fulfilled
- 2  Taking the high number of observation into account there is evidence that the assumption of homogeneous variance is not fulfilled
- 3\*  Taking the low number of observation into account there is no evidence that the assumption of homogeneous variance is not fulfilled
- 4  Taking the low number of observation into account there is evidence that the assumption of homogeneous variance is not fulfilled
- 5  Based on the information provided there cannot be drawn any conclusions about the assumption of homogeneous variance

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

The number of observations for each box-plot is 4, which is a small number (hence we can exclude answer 1 and 2). With a low number of observations we will have to accept some differences between the box plots and hence there is no evidence against the hypothesis of homogeneous variance.

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

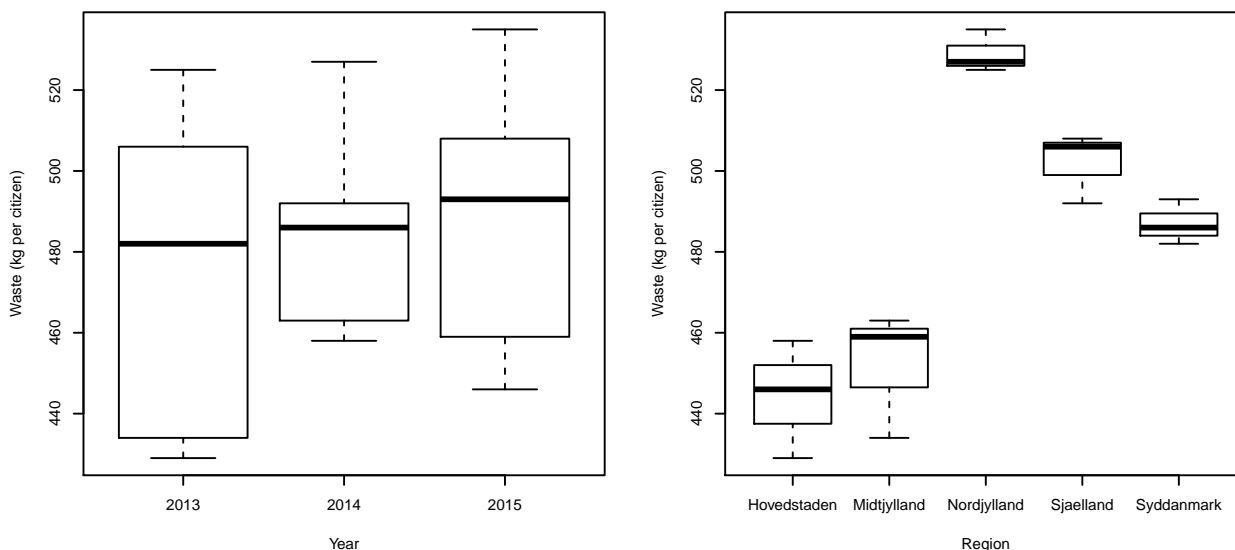
Continues on page 37

## Exercise IX

The Environmental and Food Agency collects data on waste in Denmark every year and publishes a report with data and analyses. The report is named “Affaldsstatistik 2015”<sup>(1)</sup> and in it one can find the amount of waste (kg) per citizen for the years 2013 to 2015 grouped on regions:

	Hovedstaden	Midtjylland	Nordjylland	Sjælland	Syddanmark
2013	429	434	525	506	482
2014	458	463	527	492	486
2015	446	459	535	508	493

The following box plot shows waste per citizen grouped on year and on region:



A 2-way ANOVA is carried out and the result is:

```
## Analysis of Variance Table
##
## Response: Waste
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Year       2   463.3   231.7  2.5551    0.1386
## Region     4 14847.1  3711.8 40.9386 2.266e-05 ***
## Residuals  8   725.3    90.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<sup>1</sup><http://www2.mst.dk/Udgiv/publikationer/2017/05/978-87-93614-01-7.pdf>

### Question IX.1 (28)

Which of the following statements is correct when using a significance level of  $\alpha = 5\%$ ?

- 1  From the box plot it can be seen that there is no significant difference in waste between the years, which is also the conclusion from the ANOVA test
- 2  Answers taken out of the exam.
- 3\*  From the box plot it is not possible to conclude if there is a significant difference in waste between the years, but from the ANOVA test no significant difference in waste between the years can be concluded
- 4  From the box plot it is not possible no conclude if there is a significant difference in waste between the years, but from the ANOVA test a significant difference in waste between the years can be concluded
- 5  None of the statements above are correct

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

Since we have two effects, we cannot make conclusions about the effects of years based on the box plot (since the effect of region is not being accounted for in the boxplot). Hence answer 1 and 2 are both wrong. The ANOVA test shows no significant difference between years, since the  $p$ -value= 0.1386 > 0.05.

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

### Question IX.2 (29)

Further, in the report it is listed how large a proportion of the waste is sorted in the five regions and the proportion of waste that is sorted is calculated for each year and each region. A 2-way ANOVA has been carried out on this data with the following result:

```
## Analysis of Variance Table
##
## Response: Proportion
##           Df      Sum Sq   Mean Sq F value    Pr(>F)
## Year       2 0.0109878 0.0054939  13.054 0.003026 **
## Region    4 0.0173773 0.0043443  10.323 0.003019 **
## Residuals 8 0.0033668 0.0004208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which one of the following conclusions is correct using a significance level of 5% (both argument and conclusion must be correct)?

- 1  Since the  $p$ -value  $> 0.05$  for the relevant test, a significant change in the sorted proportion over the years is not detected
- 2\*  Since  $P(F > 13.054) < 0.05$  where  $F$  follows the relevant  $F$ -distribution, a significant change in the sorted proportion over the years is detected
- 3  Since  $P(T > 0.003) > 0.05$  where  $T$  follows the relevant  $t$ -distribution, a significant change in the sorted proportion over the years is not detected
- 4  Since  $P(T < 10.323) < 0.05$  where  $T$  follows the relevant  $t$ -distribution, a significant change in the sorted proportion over the years is detected
- 5  Since  $1 - P(T > 10.323) > 0.05$  where  $T$  follows the relevant  $t$ -distribution, a significant change in the sorted proportion over the years is not detected

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

Both  $p$ -values are less than 0.05, hence Answer 1 is not correct. Since  $P(F > 13.054) = 0.00303 < 0.05$  there is a significant effect of years, hence Answer 2 is correct.

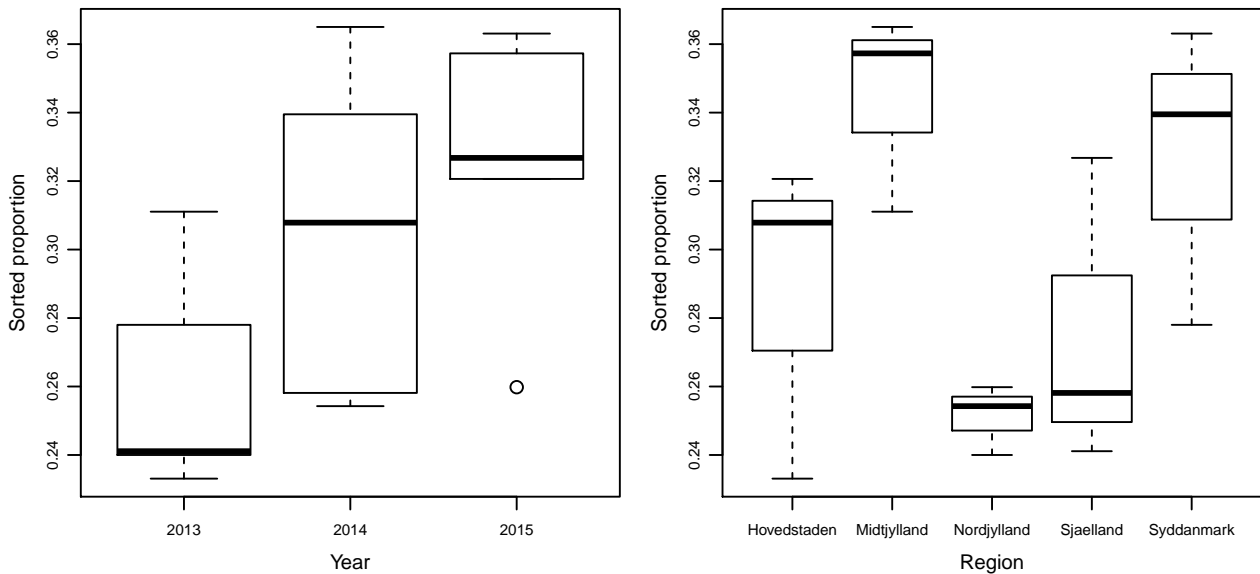
Answer 3, 4 and 5 all uses a  $t$ -distribution instead of an  $f$ -distribution and hence they are all wrong.

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

### Question IX.3 (30)

The box plots showing the proportion of sorted waste by year and by region are:





It is seen that in 2015 there is an observation which is low compared to the others, and it is identified as an outlier according to the modified box plot.

Which of the following statements is not correct (Tip: Remember that there is only one observation for each year in each region)?

- 1  The lowest observation in 2013 is from Hovedstaden
- 2  The lowest observation in 2015 (i.e. the outlier) is from Nordjylland
- 3\*  Each year Sjælland has had a higher observation than Hovedstaden
- 4  Nordjylland has the lowest median
- 5  The 75% quantile for 2014 is higher than the 25% quantile for 2015

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

Lets go through the answers:

- 1 In the right box plot it is seen that Hovedstaden has the lowest value of all the regions, and 2013 have the lowest value of all years, hence Hovedstaden must have the value in 2013. Thus TRUE
- 2 All regions have exactly one observation in 2015. Nordjylland has none higher than the outlier, hence it must belong to Nordjylland. Thus TRUE
- 3 The median for the regions mark exactly the middle observation, since there are only 3 values for each region. Since Hovedstaden has had 2 values above the 2'nd highest value

for Sjaelland, then Hovedstaden must have had a higher value one of the years. Thus NOT TRUE

4 The black bar in the box marks the median, which thus seen in the right-hand box plot to be lowest for Nordjylland. Thus TRUE

5 The 75% quantile is marked by the upper side of the box, thus on the left-hand box plot it is seen to be higher for 2014 than the 25% quantile for 2015 (marked by the lower side of the box). Thus TRUE

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

Continues on page 43

THE EXAM IS FINISHED. Enjoy the late summer!