Technical University of Denmark

Written examination: 19. Dec 2020

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

Duration: 4 hours

Aids and facilities allowed: All

The questions were answered by

____________________ (student number)

____________________ (signature)

__________ (table number)

This exam consists of 30 questions of the “multiple choice” type, which are divided between 10 exercises. To answer the questions, you need to fill in the “multiple choice” form (6 separate pages) on CampusNet with the numbers of the answers that you believe to be correct.

5 points are given for a correct “multiple choice” answer, and -1 point is given for a wrong answer. ONLY the following 5 answer options are valid: 1, 2, 3, 4, or 5. If a question is left blank or an invalid answer is entered, 0 points are given for the question. Furthermore, if more than one answer option is selected for a single question, which is in fact technically possible in the online system, 0 points are given for the question. The number of points needed to obtain a specific mark or to pass the exam is ultimately determined during censoring.

The final answers should be given by filling in and submitting the form online via CampusNet. The table provided here is ONLY an emergency alternative. Remember to provide your student number if you do hand in on paper.

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

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

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

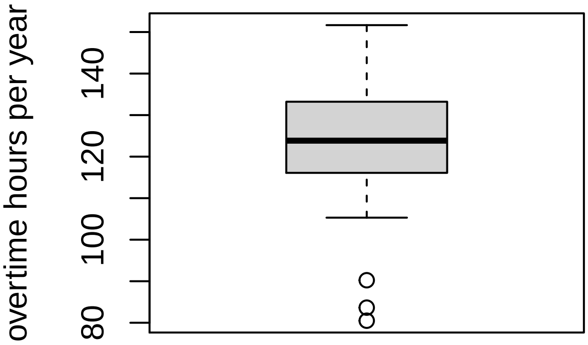
The exam paper contains 38 pages.

Continue on page 2

Multiple choice questions: *Note that in each question, one and only one of the answer options is correct. Furthermore, not all the suggested answers are necessarily meaningful. Always remember to round your own result to the number of decimals given in the answer options before you choose your answer. Also remember that there may be slight discrepancies between the result of the book's formulas and corresponding built-in functions in R.*

Exercise I

A city department has introduced a quality improvement program which allows employees to get credit for overtime hours when attending meetings. The total number of overtime hours per year for 36 employees is visualized in the boxplot below.



Question I.1 (1)

Which of the following statements is correct?

1 □ $IQR = Q1 - Q3 \approx 17$ hours.

2* □ $IQR = Q3 - Q1 \approx 17$ hours.

3 □ $IQR = Q4 - Q1 \approx 48$ hours.

4 □ The IQR cannot be determined because the boxplot contains three outliers.

5 □ $IQR = Q3 - Q1 \approx 48$ hours.

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

Using the boxplot above we can find $Q3 \approx 133$ and $Q1 \approx 116$, hence $IQR = Q3 - Q1 \approx 17$

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

Continue on page 4

Exercise II

The table below shows the number of persons tested positive for coronavirus that were admitted to hospitals in Denmark on 3 different dates during the spring of 2020. Furthermore, the table shows the numbers of those persons that were also in an intensive care unit (ICU).

Date	ICU	Admitted
April 30	62	255
April 10	113	433
March 20	37	153

Question II.1 (2)

Based on the numbers above, what is the usual 95% confidence interval for the probability that, given you are admitted, you are also in an intensive care unit? Assume that the model assumptions are fulfilled.

1 □ $[0.72, 0.78]$

2* □ $[0.22, 0.28]$

3 □ $[0.18, 0.22]$

4 □ $[0.16, 0.35]$

5 □ $[0.12, 0.28]$

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

The best estimate is to pool the observations from the three dates (this is what we must do if told nothing else about changing conditions etc.):

```r
icu <- c(62,113,37)
n <- c(255,433,153)
ph <- sum(icu)/(sum(n))
ph + c(-1,1) * qnorm(0.975) * sqrt(ph*(1-ph)/sum(n))

## [1] 0.2227350 0.2814268

# The result from the built-in function is slightly different (it's using the t-distr
prop.test(sum(icu), sum(n), correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  sum(icu) out of sum(n), null probability 0.5
## X-squared = 206.76, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2239073 0.2825089
## sample estimates:
##         p
## 0.2520809
```

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

Question II.2 (3)

In order to investigate the development over time, the numbers from April 30th and March 20th are now compared. With the null hypothesis that the proportions of patients in ICU are equal on the two dates, what is the p-value and the conclusion given a significance level $\alpha = 0.05$?

1 □ p-value=0.476 and the difference is significant.

2 □ p-value=0.029 and the difference is not significant.

3* □ p-value=0.976 and the difference is not significant.

4 □ p-value=0.024 and the difference is significant.

5 □ p-value=0.060 and the difference is not significant.

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

```r
## Q2: Compare two prop
phs <- icu[c(1,3)] / n[c(1,3)]
ph <- sum(icu[c(1,3)])/sum(n[c(1,3)])

(z <- diff(phs)/sqrt(ph*(1-ph)*(1/n[1]+1/n[3])))

## [1] -0.02981863

2*(1-pnorm(abs(z)))

## [1] 0.9762117
```

```
# or with the build in function
prop.test(icu[c(1,3)], n[c(1,3)], correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  icu[c(1, 3)] out of n[c(1, 3)]
## X-squared = 0.00088915, df = 1, p-value = 0.9762
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08457431  0.08718868
## sample estimates:
##    prop 1    prop 2
## 0.2431373 0.2418301
```

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

Question II.3 (4)

The distribution of patients across different regions is now investigated. The table below shows the number of persons admitted to hospital on different dates in the 5 regions of Denmark, we assume here that the same person is not admitted on more than 1 date.

Date	Nordjylland	Midtjylland	Syddanmark	Hovedstaden	Sjælland	All DK
April 30	13	33	12	144	53	255
April 16	21	54	35	183	60	353
April 2	32	77	85	251	86	531
March 18	10	16	12	64	27	129
Total	76	180	144	642	226	1268

We will now investigate if the proportion of admitted patients in the different regions is the same over time (the null hypothesis) or if it changes. Formally, this can be written as

$$\mathrm{H}_0: \quad p_{ij} = p_i$$

for all i.

Under the null hypothesis, what is the contribution to the test-statistics for “Nordjylland” on March 18?

1 □ 7.73

2 □ 0.59

3 □ 5.14

4* □ 0.67

5 □ 10

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

Under the null hypothesis we assume that the proportion is equal for each row across the groups. Then we can calculate the expected count in the cell and then the contribution to the χ^2 statistic:

```
(e <- 129/1268*76)

## [1] 7.731861

(10-e)^2 / e

## [1] 0.6653577
```

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

Question II.4 (5)

The test statistics is calculated to $\chi^2_{obs} = 29$. Given a significance level $\alpha = 0.05$, what is the p-value and conclusion for the corresponding hypothesis test? (Both argument and conclusion must be correct)

1 □ p-value=0.0012 and there is a significant difference

2 □ p-value=0.0099 and there is not a significant difference

3 □ p-value=0.024 and there is a significant difference

4 □ p-value=0.088 and there is not a significant difference

5* □ p-value=0.0039 and there is a significant difference

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

```
1 - pchisq(29, df = 12)

## [1] 0.00393999
```

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

Question II.5 (6)

If we on a given day assume that 4% of the population is infected with a virus, how many people should then be tested at random in order to get a margin of error on maximum 1% using significance level $\alpha = 0.05$?

1 □ 1039

2* □ 1476

3 □ 369

4 □ 9603

5 □ 6764

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

```
0.04 * 0.96 * (qnorm(0.975)/0.01)^2

## [1] 1475.12
```

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

Continue on page 9

Exercise III

The 2008-09 nine-month academic salary for Professors in a given U.S. college is to be assessed. The data includes salaries of 125 male Professors working in applied departments (in US dollars). It is of interest to find out if the salary depends on the years of work since obtaining a Ph.D. degree and years of service.

Question III.1 (7)

An initial multiple linear regression model was established. The model summary is given below. Assume that the model assumptions are fulfilled!

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72479 -20472   -288  16051  92778
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   130213.8     6956.5  18.718   <2e-16 ***
## yrs.since.phd   -304.2      430.1  -0.707    0.481
## yrs.service      529.3      378.5   1.398    0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26450 on 122 degrees of freedom
## Multiple R-squared:  0.02085,Adjusted R-squared:  0.004803
## F-statistic: 1.299 on 2 and 122 DF,  p-value: 0.2765
```

Which of the following statements is correct given a significance level $\alpha = 0.05$? (Both conclusion and argument must be correct)

1 □ The Professor `salary` depends on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.05.

2 □ The Professor `salary` does NOT depend on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.025.

3 □ We are not given sufficient information to make a conclusion about the relation between Professor `salary` and `yrs.since.phd` and `yrs.service`.

4 □ The Professor `salary` depends on `yrs.since.phd` and `yrs.service` because the respective p-values are less than 0.5.

5* □ The Professor `salary` does NOT depend on `yrs.since.phd` and `yrs.service` because both p-values are greater than 0.05.

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

The p-values for `yrs.since.phd` and `yrs.service` are both greater than the significance level $\alpha = 0.05$ (Note: *alpha* is not 0.025). We can therefore state that there is no significant correlation between Professor `salary` and `yrs.since.phd` and `yrs.service`. This is equivalent to stating that the Professor `salary` does not depend on `yrs.since.phd` and `yrs.service`.

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

Question III.2 (8)

Backwards model selection was performed for the multiple linear regression model above, resulting in the following R output:

```
##
## Call:
## lm(formula = salary ~ yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73189 -20581     29  15226  92951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126901.7     5133.7  24.719   <2e-16 ***
## yrs.service    307.7      212.0   1.451    0.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26400 on 123 degrees of freedom
## Multiple R-squared:  0.01684,Adjusted R-squared:  0.008846
## F-statistic: 2.107 on 1 and 123 DF,  p-value: 0.1492
```

```
##
## Call:
## lm(formula = salary ~ 1, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65959 -19018   -693  16858  98027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   133518       2372    56.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26510 on 124 degrees of freedom
```

Which R code results in the correct 95% confidence interval for the mean of the Professor salary?

1 □

```
133518 + c(-1, 1) * qt(0.95, 124) * 2372
```

2 □

```
133518 + c(-1, 1) * qt(0.975, 123) * 2372
```

3* □

```
133518 + c(-1, 1) * qt(0.975, 124) * 2372
```

4 □

```
126902 + c(-1, 1) * qt(0.975, 124) * 5134
```

5 □

```
130214 + c(-1, 1) * qt(0.95, 124) * 6957
```

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

The correct 95% confidence interval is found using the fully reduced model. `yrs.since.phd` and `yrs.service` were not significant and were removed step-wise. Hence, the mean Professor salary can be found using the following R-command:

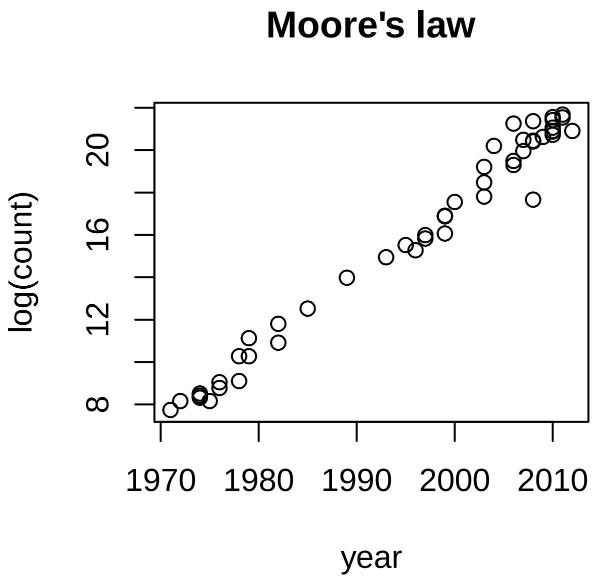
```
133518 + c(-1, 1) * qt(0.975, 124) * 2372

## [1] 128823.1 138212.9
```

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

Continue on page 13

Exercise IV

Moore's law is about the observation that the number of transistors in a dense integrated circuit doubles about every two years. The observation is named after Gordon Moore, the co-founder of Fairchild Semiconductor. In the figure below the transistor count has been transformed using the natural logarithm and plotted against year.



```
##
## Call:
## lm(formula = log(count) ~ year, data = moore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60701 -0.26843 -0.01245  0.35038  1.67737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.786e+02  1.414e+01  -48.01        ?
## year         3.481e-01  7.083e-03       ?   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6762 on 46 degrees of freedom
## Multiple R-squared:  0.9813,Adjusted R-squared:  0.9809
## F-statistic:  2415 on 1 and 46 DF,  p-value: < 2.2e-16
```

Question IV.1 (9)

Calculate the test statistic which is missing in the model summary above (missing values have been replaced by question marks in the table above). Which of the following answers is correct?

1 □ $t_{\text{obs}} = 0.02$

2 □ $t_{\text{obs}} = 12.25$

3 □ $t_{\text{obs}} = 0.49$

4* □ $t_{\text{obs}} = 49.15$

5 □ $t_{\text{obs}} = 12.49$

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

The test statistic, t_{obs}, describes how many standard errors the estimated slope $\hat{\beta}_{year}$ is away from the hypothesized slope $\beta_{year,0} = 0$ and can be calculated using the formula:

$t_{obs} = \frac{\hat{\beta}_{year} - \beta_{year,0}}{\hat{\sigma}_{year}}$, where $\hat{\sigma}_{year}$ is the standard error for the slope of `year`.

Using the model summary we obtain:

$t_{obs} = \frac{3.481 \cdot 10^{-1}}{7.083 \cdot 10^{-3}} = 49.15$

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

Question IV.2 (10)

We want to test the hypothesis $H_0 : \beta_0 = 0$, where β_0 represents the model intercept. Which of the following statements is correct (given $\alpha = 0.05$)? (Both argument and conclusion must be correct!)

1 □ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.96$. We reject H_0 because $|t_{obs}| > t_{crit}$.

2* □ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 2.01$. We reject H_0 because $|t_{obs}| > t_{crit}$.

3 □ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.68$. We reject H_0 because $|t_{obs}| > t_{crit}$.

4 □ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 2.01$. We accept H_0 because $|t_{obs}| > t_{crit}$.

5 □ We compare the absolute value of the corresponding test statistic $|t_{obs}| = 48.01$ with the critical t-value, $t_{crit} = 1.96$. We accept H_0 because $|t_{obs}| > t_{crit}$.

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

We reject H_0 because $|t_{obs}| > t_{crit}$

```
(t_crit <- qt(0.975, df = 46))

## [1] 2.012896
```

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

Question IV.3 (11)

According to the linear model above, what is the expected transistor count increase from 2010 to 2015?

1 □ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015) - \ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

2 □ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

3* □ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

4 □ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

5 □ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015}$

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

We use the estimated linear model parameters as shown in the summary above to find expected transistor counts for years 2015 and 2010. These expected counts are on the natural logarithmic scale and require back-transformation. After we have performed the back-transformation we can subtract the expected count of 2010 from the expected count of 2015.

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

Continue on page 16

Exercise V

Question V.1 (12)

One is interested in determining the density of a liquid. To do so, the mass, m, and the volume, V, of the liquid are measured. The density of the liquid is given by

$$\rho = \frac{m}{V}$$

What is the precision (standard deviation, σ_ρ) of the determined density if the mass and the volume can be measured with a precision $\sigma_m = 0.2$ and $\sigma_V = 0.4$, respectively? Assume that mass and volume measurements are independent and normally distributed.

1 □ $\sigma_\rho \approx \frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})$

2* □ $\sigma_\rho \approx \sqrt{\frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})}$

3 □ $\sigma_\rho \approx \frac{1}{V^2}(0.4^2 + \frac{0.2^2 m^2}{V^2})$

4 □ $\sigma_\rho \approx \frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}$

5 □ $\sigma_\rho \approx \sqrt{\frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}}$

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

We can find the precision σ_ρ using the error approximation rule for non-linear functions (see slides week 7).

$$\sigma_\rho^2 \approx (\frac{\partial \rho}{\partial m})^2 \sigma_m^2 + (\frac{\partial \rho}{\partial V})^2 \sigma_V^2$$

$$\sigma_\rho^2 \approx \frac{1}{V^2}\sigma_m^2 + \frac{m^2}{V^4}\sigma_V^2$$

$$\sigma_\rho^2 \approx \frac{1}{V^2}(\sigma_m^2 + \frac{\sigma_V^2 m^2}{V^2})$$

$$\sigma_\rho \approx \sqrt{\frac{1}{V^2}(\sigma_m^2 + \frac{\sigma_V^2 m^2}{V^2})}$$

$$\sigma_\rho \approx \sqrt{\frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})}$$

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

Question V.2 (13)

Let X_i be a random variable. The following code is run in R to draw 100 random numbers X_i from a given distribution.

```r
x <- rnorm(100)^2 + rnorm(100)^2 + rnorm(100)^2
```

Which of the following statements is correct?

1 □ X_i follows a χ^2-distribution with 1 degree of freedom.

2 □ X_i follows a standard normal distribution with mean 0 and variance 1.

3 □ X_i follows a χ^2-distribution with 2 degrees of freedom.

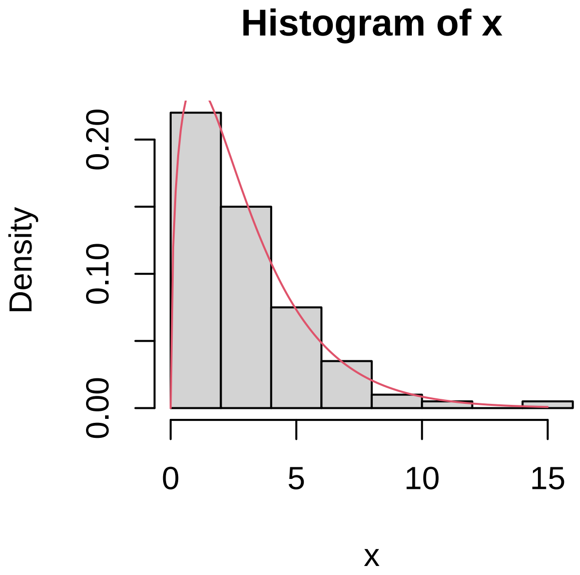
4* □ X_i follows a χ^2-distribution with 3 degrees of freedom.

5 □ X_i follows a normal distribution with mean 0 and variance 3.

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

We draw 100 standard normal numbers three times, and for each element we sum them, so we according to 2.79 we have 3 df. Check it by:

```r
x <- rnorm(100)^2 + rnorm(100)^2 + rnorm(100)^2
length(x)

## [1] 100

hist(x, prob=TRUE)
xseq <- seq(0,15,by=0.1)
lines(xseq, dchisq(xseq, df=3), col=2)
```



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

Question V.3 (14)

Which of the following R commands is drawing 10 random numbers from an exponential distribution?

1* □ `replicate(10, rexp(1, 2))`

2 □ `pexp(seq(0.1, 1, length.out=10), 2)`

3 □ `qexp(seq(0.1, 1, 0.1), 2)`

4 □ `rep(dexp(10, 2), 10)`

5 □ None of the above. The exponential distribution requires a second parameter, which is missing in all of the above

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

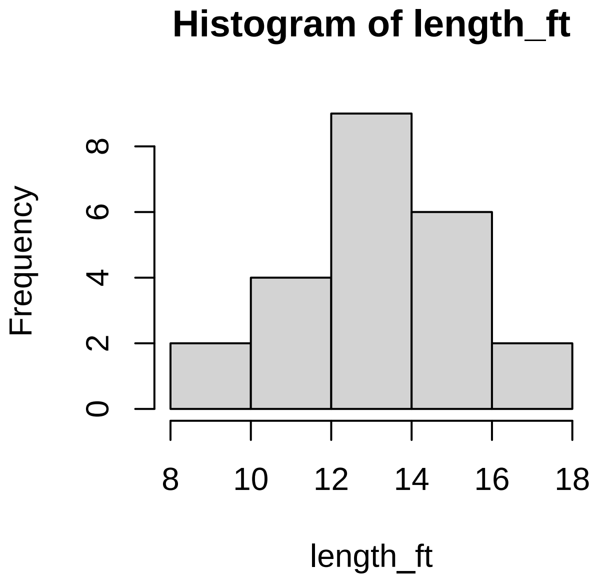
`rexp(1, 2)` draws 1 number from an exponential distribution with mean = 2. The `replicate` command ensures that this procedure is repeated 10 times. An easier way to draw ten random numbers from an exponential distribution would be to use the command `rexp(10, 2)`.

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

Continue on page 19

Exercise VI

Jesus Rivas, a herpetologist, is currently doing research on green anacondas. These snakes, some of the largest in the world, can grow up to 25 feet in length. They have been known to swallow live goats and even people. Jesus Rivas and fellow researchers walk barefoot in shallow water in the Llanos grasslands shared by Venezuela and Colombia during the dry season. When they feel a snake with their feet, they grab it and hold it with the help of another person. After muzzling the snake with a sock and tape, they measure the length of the snake. 23 green anacondas were captured and their length was measured in feet. The sample data is stored in `length_ft`. You can see the corresponding histogram of the sample below.



Question VI.1 (15)

Which of the following is the correct 99% confidence interval for the median anaconda length assuming that parametric bootstrapping was used for estimation of the interval?

```
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))

##     2.5%    97.5%
## 12.02935 14.59873
```

```
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.94304 14.68206

sim_samples <- replicate(k, rchisq(n, mean_ft))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 10.75972 16.24469

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.64546 15.04535

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft^2))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##      0.5%     99.5%
##  9.121213 17.500782
```

1 □ [12.03, 14.60]

2 □ [11.94, 14.68]

3 □ [10.76, 16.24]

4* □ [11.65, 15.05]

5 □ [9.12, 17.50]

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

Parametric bootstrapping requires knowledge regarding the population's distribution. As it can be seen from the histogram above the snake length follows approx. a normal distribution. As we are interested to simulate the 99% confidence interval for the median snake length only the fourth answer can be correct.

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

Question VI.2 (16)

Which of the following is the correct 99% confidence interval for the median anaconda length assuming that non-parametric bootstrapping was used for estimation of the interval?

```
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.93076 15.22501

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.61621 14.97613

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.01, 0.99))

##       1%      99%
## 12.08800 14.46791

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.01, 0.99))

##       1%      99%
## 12.48738 15.03513

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))

##     2.5%    97.5%
## 12.82957 14.46058
```

1* □ [11.93, 15.23]

2 □ [11.59, 15.05]

3 □ [12.13, 14.50]

4 □ [12.49, 15.04]

5 □ [12.83, 14.46]

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

In case of non-parametric bootstrapping we sample with replacement from our sample data. We are still interested in a 99% confidence interval for the median, hence only answer 1 can be correct.

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

Continue on page 24

Exercise VII

Question VII.1 (17)

You have been collecting amber with a friend and you found in total 20 pieces. You agreed to share it by randomly drawing 10 pieces each. Three of the pieces are very attractive. What is the probability that you will get all three attractive pieces?

1 □ 0.0877%

2 □ 0.877%

3* □ 10.5%

4 □ 13.0%

5 □ 24.0%

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

This is hyper geometric, since it is drawing without replacement, so

```
dhyper(3, 3, 17, 10)

## [1] 0.1052632
```

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

Question VII.2 (18)

Let X represent the weight in grams of a new piece of amber that you find at your favourite location. From experience you know that when you find a piece of amber there, then its weight follows a log-normal distribution, such that $X \sim LN(1, 0.7^2)$.

What is the mean weight μ_X of amber pieces at your favourite location according to this model?

1 □ 2.01 g

2 □ 2.72 g

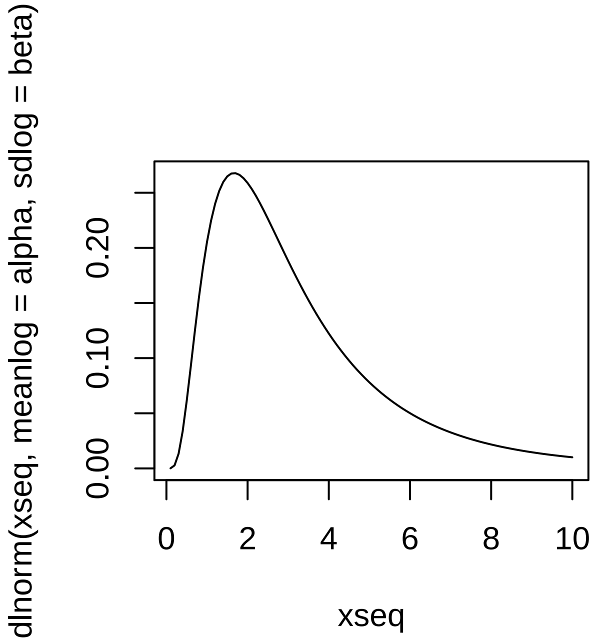
3* □ 3.47 g

4 □ 5.93 g

5 □ 9.21 g

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

```
alpha <- 1
beta <- 0.7
##
xseq <- seq(0.1, 10, by=0.1)
plot(xseq, dlnorm(xseq, meanlog=alpha, sdlog=beta), type = "l")
##
exp(alpha+beta^2/2)

## [1] 3.472935
```



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

Question VII.3 (19)

Based on the information given in the last question: If you find 20 pieces at your favourite location, what is the probability that at least 3 of them weigh more than 10 grams?

1 □ 0.31%

2* □ 2.36%

3 □ 3.14%

4 □ 4.24%

5 □ 12.31%

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

Now this is drawing with replacement, since every time we find a new piece it's from an “infinite“ sized population. So first we find the probability that a new piece is more than 10 grams i.e.

$$P(X > 10) = 1 - P(X < 10)$$

(X is weight), so in R:

```
(p <- 1 - plnorm(10, meanlog=alpha, sdlog=beta))

## [1] 0.03138368
```

and this is the success probability in the binomial drawing. The probability of finding 3 or more pieces is then

$$P(Y \geq 3) = 1 - P(Y \leq 2)$$

```
1 - pbinom(2, 20, p)

## [1] 0.02363236
```

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

Continue on page 27

Exercise VIII

Let the random variable X_i represent the i'th observation in a sample of n observations from a population which is uniformly distributed between α and β. The observations are sampled randomly and thus independently of each other. So $X_i \sim U(\alpha, \beta)$ and i.i.d.

Question VIII.1 (20)

The sample mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

What is the distribution of $\bar{X}$ as n goes to infinity?

1 □ $N(0, 1^2)$

2 □ $U(\alpha, \beta)$

3 □ t-distribution with $n-1$ degrees of freedom

4* □ $N(\frac{\alpha+\beta}{2}, \frac{(\beta-\alpha)^2}{12n})$

5 □ $U(\alpha^n, \beta^n)$

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

From the CLT Theorem 3.14 we know that the sample mean, i.e. a sum of i.i.d. random variables, is normal distributed, with same mean and variance divided by the number of variables n.

The mean and variance of the uniform distribution is found in Theorem 2.36, which inserted gives the answer.

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

Question VIII.2 (21)

Define $Y_i = 2 + \frac{1}{10}X_i$, which of the following statements is correct?

1 □ $\mathrm{E}(Y_i) = \frac{1}{10}\,\mathrm{E}(X_i)$

2 □ $\mathrm{E}(Y_i) = \frac{1}{100}\,\mathrm{E}(X_i)$

3 □ $\mathrm{V}(Y_i) = \frac{1}{10}\,\mathrm{V}(X_i)$

4* □ $\mathrm{V}(Y_i) = \frac{1}{100}\,\mathrm{V}(X_i)$

5 □ $Y_i \sim U(\alpha, \beta)$

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

We use the identities for linear variables in Theorem 2.54.

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

Continue on page 29

Exercise IX

In power systems the balancing power is the generation or load which can quickly be increased or decreased to stabilize the voltage on the grid. The balancing power is often traded on a market, as on the Dutch aFRR market, where bids are settled for 15 minute intervals. If you participate on such a marked, it is important to know how much energy is activated.

First the activated up-regulation volume is analyzed, that is how much energy in total was activated for increased generation per day. The average daily values in MWh for three winter months are read into the vector `xwinter` and the following analysis is carried out

```
t.test(xwinter)

##
##  One Sample t-test
##
## data:  xwinter
## t = 14, df = 89, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.346 12.341
## sample estimates:
## mean of x
##     10.84
```

Question IX.1 (22)

Let μ_{winter} be the mean up-regulation volume on winter days. Assuming a significance level $\alpha = 0.05$, what should be the conclusion on the following null hypothesis (both conclusion and argument must be correct)?

$$H_0 : \mu_{\text{winter}} = 10$$

1 □ The null hypothesis is rejected, since the p-value is below $2 \cdot 10^{-16}$ which is below 5%

2 □ The null hypothesis is accepted, since the p-value is below $2 \cdot 10^{-16}$ which is below 5%

3 □ The null hypothesis is rejected, since the p-value is below $2 \cdot 10^{-16}$ which is above 5%

4 □ The null hypothesis is accepted, since the p-value is below $2 \cdot 10^{-16}$ which is above 5%

5* □ The null hypothesis is accepted, since 10 is inside the 95% confidence interval

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

It's clear that the mean under the null hypothesis (μ_0) is inside the condidence interval, in which case we know that the null hypothesis will not be rejected, i.e. it must be accepted.

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

Question IX.2 (23)

What is the 99% confidence interval for μ_{winter}?

1 □ $[7.77, 13.91]$

2 □ $[8.01, 12.10]$

3 □ $[8.28, 13.41]$

4* □ $[8.86, 12.82]$

5 □ $[9.35, 12.34]$

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

Half the width of the confidence interval is

$$t_{0.975}\frac{s}{\sqrt{n}} = (12.341 - 9.346)/2 = 1.4975$$

so by looking up $t_{0.975}$ in R

```
qt(0.975, df=89)

## [1] 1.987
```

we find the standard error to

$$\frac{s}{\sqrt{n}} = \frac{1.4975}{t_{0.975}} = 0.7537$$

so we can find the 99% confidence interval by

$$\bar{x} \pm t_{0.995}\frac{s}{\sqrt{n}}$$

```
10.84 + c(-1,1) * qt(0.995, df=89) * 0.7537

## [1]  8.856 12.824
```

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

Question IX.3 (24)

What is the number of observations in `xwinter`?

1 □ 88

2 □ 89

3* □ 90

4 □ 91

5 □ 92

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

In a one-sample t-test we know that the degrees of freedom is $n - 1$, and since df is 89, then n is 90.

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

Question IX.4 (25)

In order to find out if there is a difference between winter and summer, the daily averages of up-regulation volume for the summer months in the same year are loaded into `xsummer`.

Based on the given data in the exercise, which of the following tests is best suited for concluding if there is a significant difference between the daily mean of up-regulation volume in winter and in summer?

1* □ A two-sample t-test

2 □ A paired two-sample t-test

3 □ A two-way ANOVA test

4 □ A test for the slope coefficient in a linear regression model

5 □ A χ^2-test

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

We have two samples, one from winter and one from summer, so it's a two sample test. They cannot be paired, since they are not on same dates and don't share other features that we are informed about.

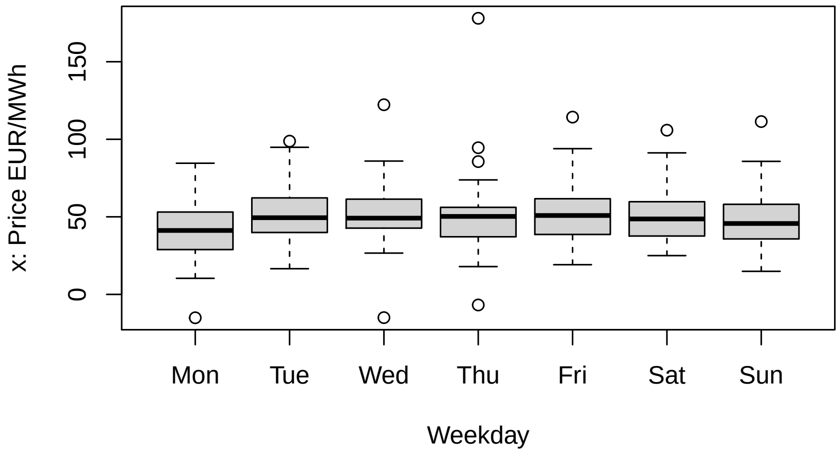
------------------------------------ FACIT-END ------------------------------------

Continue on page 33

Exercise X

This exercise is about the Dutch aFRR balancing power market as mentioned in the previous exercise. For providers of flexible power it is important to investigate the prices at which the balancing power is sold and bought on the market. A year of daily average price of down-regulation power is read into `x`. 364 observations (days) were included in the data.

To see if there are differences between the days of the week, box-plots are generated for each day (note that the prices are given per energy unit, this detail doesn't matter in this exercise):



A one-way ANOVA was carried out. The result are given below:

```
anova(lm(x ~ weekday))

## Analysis of Variance Table
##
## Response: x
##            Df Sum Sq Mean Sq F value  Pr(>F)
## weekday     6   4934  822.42  2.0969 0.05296 .
## Residuals 357 140016  392.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question X.1 (26)

Given a significance level of 5%, what is the critical value for the F-test of equal weekday means?

1 □ 1.549

2 □ 1.791

3 □ 1.943

4* □ 2.124

5 □ 2.444

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

The test statistic follows an F-distribution under the null hypothesis, see Theorem 8.6 and the critical value is the $1-\alpha$ quantile in the F-distribution with $k-1$ and $n-k$ degrees of freedom, and $k = 7$ different weekdays, so

```
qf(0.95, 6, 357)

## [1] 2.123994
```

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

Question X.2 (27)

Assuming that all model assumptions are fulfilled, what is the estimate of the variance of the daily average down-regulation price on Fridays using this model (both value and explanation must be correct)?

1* □ $\hat{\sigma}^2 = 392.2$, since the variance estimate is pooled and thus it is the same for all weekdays

2 □ $\hat{\sigma}^2 = \frac{140016}{4934} = 28.38$, since the variance estimate is pooled and thus it is the same for all weekdays

3 □ $\hat{\sigma}^2 = \frac{140016}{7} = 20002$, since the variance estimate must be split on the different weekdays, thus adjusted by the degrees of freedom for `weekdays`

4 □ $\hat{\sigma}^2 = \frac{140016}{6} = 23336$, since the variance estimate must be split on the different weekdays, thus adjusted by the degrees of freedom for `weekdays`

5 □ This cannot be calculated with the given information

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

Since one of the model assumptions for the model in the ANOVA, is that the variance is homogeneous, meaning that it's the same for all groups, then the variance is pooled. We can read it off directly from the ANOVA table printed in the result.

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

Question X.3 (28)

What is the proportion of variance explained by the model?

1 □ 0.57%

2* □ 3.4%

3 □ 18.4%

4 □ 32.3%

5 □ 96.6%

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

It's the proportion of variance explained by the "treatment" (here `weekday`) of the total variance SST, hence

```
4934 / (140016 + 4934)

## [1] 0.03403932
```

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

Continue on page 36

Question X.4 (29)

Now the week number **week** is added as a second factor and a two-way ANOVA is carried out:

```
anova(lm(x ~ weekday + week))

## Analysis of Variance Table
##
## Response: x
##            Df Sum Sq Mean Sq F value Pr(>F)
## weekday     6   4934     822    2.97 0.0079 **
## week       51  55218    1083    3.91  6e-14 ***
## Residuals 306  84798     277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When comparing this result to the result of the one-way ANOVA it can be seen that the p-value for `weekday` decreased a lot. Which of the following statements is the correct explanation to this?

1 □ The variation explained by grouping into weekdays ($SS(weekday)$) increases, thus the p-value for the effect of weekdays decreases

2 □ The degrees of freedom for `Residuals` decreases, which leads to the decrease in p-value for the effect of weekdays

3 □ The variation explained by grouping into weekdays ($SS(weekday)$) decreases, thus the p-value for the effect of weekdays decreases

4* □ The residual sum of squares (SSE) decreases significantly, which leads to the decrease in p-value for the effect of weekdays

5 □ There must be a significant correlation between weekday group means and week group means and this leads to the decrease in p-value for the effect of weekdays

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

Compared to the one-way model, much more of the variance is explained by adding the weekdays to the model (compared to the additional number of parameters in the model), thus, when this variation is explained, the effect of weekday becomes significant. Softly put: “a lot of noise is removed (or explained) by the weeks, thus the effect of weekday can be seen more clearly”.

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

Question X.5 (30)

We are now interested in performing a post-hoc analysis concerning the ANOVA model shown in the question above. The following was run in R:

```
tapply(x, weekday, mean)

##  Mon  Tue  Wed  Thu  Fri  Sat  Sun
## 41.4 51.3 52.0 50.9 52.8 51.7 48.1

tapply(x, weekday, sd)

##  Mon  Tue  Wed  Thu  Fri  Sat  Sun
## 19.9 18.3 19.5 24.4 20.2 17.4 18.0
```

Which of the following R calls gives the correct single pre-planned 95% confidence interval for the difference in mean of the daily price between Saturdays and Sundays?

1 □ `t.test(x[weekday=="Sat"], x[weekday=="Sun"], conf.level=0.9976)`

2 □ `t.test(x[weekday=="Sat"], x[weekday=="Sun"])`

3* □ `51.7 - 48.1 + c(-1,1) * qt(0.975, 306) * sqrt(2 * 277 * 1/52)`

4 □ `51.7 - 48.1 + c(-1,1) * qt(0.9988, 52) * sqrt(2 * 277 * 1/52)`

5 □ `51.7 - 48.1 + c(-1,1) * qt(0.9988, 306) * sqrt(17.4^2/52 + 18.0^2/52)`

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

```
51.7 - 48.1 + c(-1,1) * qt(0.975, 306) * sqrt(2 * 277 * 1/52)

## [1] -2.82 10.02
```

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

Continue on page 38

The exam is over! Enjoy your Christmas holidays!