Sample exam

1. (3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 = 30 points)
(a) Give one example of a random variable with Bernoulli(0.1) distribution. What values does it take?

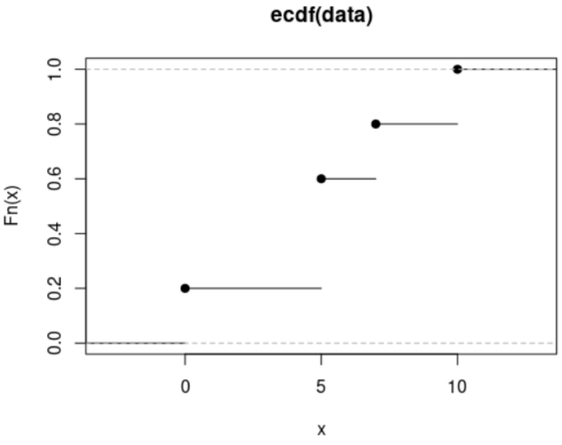
For example a biased coin flip, where P(heads) = P(1) = 0.1 and P(0) = P(tails) 0.9, so it takes values 0 and 1 representing tail and head.

(b) Give an example of a dataset (realisation of a random sample) with 0 skewness.

Any data that is symmetric over the mean, for example: 0, 9, 1, 10, 5

(c) For the following data draw a plot of the empirical distribution function. Data: 5, 0, 7, 10, 5.

```
data<- c(5,0,7,10,5)
plot(ecdf(data))
```



(d) Explain what is the imputation of a missing value. Give two particular examples how you can impute the value of a numerical variable age.

Imputation of missing value means replacing of the missing value, for example by predicting based on other variables in the observation or by a mode, etc. For example if our dataset has redundancy like birth date, we can easily impute the age based on this property. If there is no dependency, we might replace it with a mode or by a random Guess.

Imputation – replace the missing value with an arbitrary value (typical value from the data, predicted value from the data).

1.Mean/Median Imputation:- In a mean or median substitution, the mean or a median value of a variable is
used in place of the missing data value for that same variable.
2.Mode substitution:- In mode substitution,the highest occurring value for categorical value is used in
place of the missing data value of the same variable. Better for categorical variables
3.Cold-Deck Imputation:-A systematically chosen value from an individual who has similar values on other
variables.This is similar to Hot Deck in most ways, but removes the random variation.

(e) Let u = 0.5478 be a realisation of a random variable with uniform U(0, 1) distribution. Based on u, generate a realisation of a random variable with uniform U(−5, 12) distribution. Explain your approach in detail.

Using a generator U1(0,1) we can create a generator U2(-5, 12) by shifting and scaling the original one. We shift U1 by -5 and scale it by 12 + 5 = 17 -> -5 + U1 * 17 to get a generator U2. To translate the realisation of the R.V. from one interval U1 to interval of U2 we do the same with the realisation directly getting -5 + 0.5478 * 17 which is equal to 4.3126

(f) Assume that the probability that a newborn baby is a boy is 0.53. With the aid of the central limit theorem, compute the approximate probability that the number of newborn boys among 5 000 babies is greater than 2 600. Explain your approach in detail.

As the underlying model is sequence of bernoulli trials, so a binomial distribution of Bi(5000, 0.53). Based on the assumptions, the mean of newborn boys should be 0.53 * 5 000 = 2650 (n*p) and variance is n*p*q = 5000*0.53*0.47 = 1245.5 Then approximating using CLT and normal distribution of X ~ N(2650, 1245.5) we look at the

P(X >= 2600) which is equal to 1 - F(2600)
pnorm(2600,mean=2650,sd=1245.5) =0.483989
1 - pnorm(2600, 2650, 1245.5) = 0.516011

(g) Let 15, 18, 15, 19, 16, 17, 13, 14 be a realisation of a random sample. Write down one parametric bootstrap sample based on the above realisation (when assuming normality of the data). Explain your approach.

When assuming that the data comes from normal distribution, we need to estimate the unknown parameters: mean and standard deviation.

We can estimate the parameters using MLE, simply by computing sample mean for mean estimation and computing MLE of variance.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}, \ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Mean = 15.875
Variance = 4.125

Then we generate random sample of size n (n = sample size of the original data), the bootstrap from the estimated distribution and compute

```
data<- c(15, 18, 15, 19, 16, 17, 13, 14 )
n<- length(data)
rnorm(n=n, mean=mean(data), sd=var(data))
```

```
14.22817 15.41793 12.64028 18.54200 20.35352 16.54897 22.60417 13.99523
```

(h) Consider the standardized residuals $r_1^*, \ldots, r_n^*$ in linear regression model. Are they independent? Are they identically distributed? Explain.

Linear regression slide 31 (580 merged), Unlike model errors ei, ri are neither independent, nor identically distributed. Unlike model errors, the residuals are neither independent or i.i.d., due to the estimation of linear regression parameters, the residuals sum up to 0.

But standardized residuals are still not independent nor iid., but they have normal distribution, so they have all of these things only approximately, so they are almost i.i.d and almost independent.

2. (10 points)
Let 6, 5, 7, 3, 4, 4, 6 be a random sample from Pois(λ) distribution, where λ > 0 is the unknown parameter.
(a) Write down likelihood function for your data.

$$L(\lambda; x_1, \ldots, x_n) = \prod_{j-1}^{n} \exp(-\lambda) \frac{1}{x_j!} \lambda^{x_j}$$

(b) Find the maximum likelihood estimate of λ (either analytically or numerically).

$$\widehat{\lambda}_n = \frac{1}{n} \sum_{j-1}^{n} x_j$$

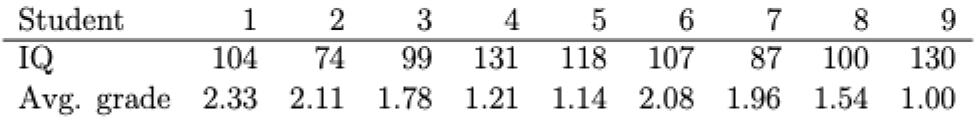
Therefore, the estimator is just the sample mean of the observations in the sample.

This makes intuitive sense because the expected value of a Poisson random variable is equal to its parameter, and the sample mean is an unbiased estimator of the expected value.

6+5+7+3+4+4+6/n = 5

(c) Find the maximum likelihood estimate of the parametric function λ 2 .
(6+5+7+3+4+4+6/n)^2 = 25

3. (10 points)
We randomly selected 9 students to investigate the dependency between their IQ and average grade. The corresponding values are listed below:



Student	1	2	3	4	5	6	7	8	9
IQ	104	74	99	131	118	107	87	100	130
Avg. grade	2.33	2.11	1.78	1.21	1.14	2.08	1.96	1.54	1.00

Is there any dependency between IQ and average grade?
To measure depedency, we can measure correlation between these 2 values, if there is any linear relationship, the corr. coefficient will be large

```
IQ<- c(104,74,99,131,118,107,87,100,130)
avg_grade <- c(2.33,2.11,1.78,1.21,1.14,2.08,1.96,1.54,1.00)
```

(a) Define the mathematical model – random variables and their distributions, null and alternative hypotheses, model assumptions. Explain all the symbols you use in detail.
Model, our data is generated by a random vector X(IQ, Grade), representing students attributes, IQ is a R.V. that is normally distributed by definition of IQ and similarly I will assume that avg. grades are normally distributed from CLT. In other words, it is a random sample from bivariate normal distribution.

The H0: corr. coefficient = 0 (they is no linear relation) H1: corr. coefficient != 0

If correlation coefficient is equal to 0 we know that these are independent

(b) Is there any significant dependency between IQ and average grade? Explain in detail. You do not need to check the assumptions.
c) How strong the dependency is? Give one measure for it and interpret its value.
Based on basic graphical technique, looking at the data and scatterplot, we can see there is some kind of relationship between higher iq and lower average grade (lower the better)
The best grades are among the people with high intelligence, but there is a lot of mixup. Using a single measure - Pearson's correlation coefficient which is equal to -0.753, we can see that there is strong negative relationship, this means that higher IQ correlates with better (closer to 1.0) grades.

```
plot(IQ,avg_grade)
```

```
cor.test(IQ,avg_grade)
cor(IQ,avg_grade)
```

(d) Compute the corresponding confidence interval. How would you interpret it? Use it for testing the null hypothesis.

95 percent confidence interval: -0.9447040 -0.1778664, meaning that 95 times out of 100 sampling of another new n samples, the true coefficient will be withing the bounds 95% of the times

```
cor.test(IQ,avg_grade)
```

The confidence interval is wide, but the negative correlation is significant. The p value is 0.01917 which is smaller than 5%, this means we reject the null hypothesis that the IQ and Grade are not correlated.

4. (10 points)

For a dataset containing data about protein consumption (red meat, white meat, eggs, milk, fish, cereals, starch, nuts and fruits and vegetables) in 25 European countries in 1985, we performed PCA with sample correlation matrix. Its eigenvalues are 4.01, 1.63, 1.13, 0.95, 0.46, 0.33, 0.27, 0.12, 0.10. Corresponding first two eigenvectors are displayed in Figure 1.

(a) What is the equation for the first principal component. How can you interpret it?

$$Y_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p$$

C = weight of that specific variable a.k.a. Eigenvectors

Xi = variable

Y1 = -0.3026094 * RedMeat + (-0.3105562) * WhiteMeat + (-0.4266785) * Eggs + …

How can we interpret it: Countries in Europe can be divided into small groups with similar features of protein consumption based on these values. If we take the highest values from column PC1(either positive or negative), then we can say that based on the first principal component we can split countries into those that have high intake of nuts and cereals, but also low intake of protein from eggs and other countries that don't have this.

(b) How can you interpret the second principal component?

Countries may be split into subsequent categories based on their fish and fruits&vegetable protein intake.The second PCA has significant negative associations with Fish and fr. veg. and Fish

(c) How much of the variability is explained by the first, respectively second, principal component?

We need to look into eigenvalues. We can either sum it up or just use their number(it is the same thing). In order to explain variability of the first component we would 4.01 / 9 = 0.4455556. Almost 45% of the variability in our data can be explained by the first principal component. 1.63 / 9 = variability of PC2 and so on. We can do cumulative sum to explain it by the first two and so on.

4.01/ (4.01 + 1.63 + 1.13 + 0.95 + 0.46 + 0.33 + 0.27 + 0.12 + 0.10) = 0.44555555555

1.63/ (4.01 + 1.63 + 1.13 + 0.95 + 0.46 + 0.33 + 0.27 + 0.12 + 0.10) = 0.18111111111

(d) For Yugoslavia with standardized values (−1.62, −0.78, −1.55, −1.07, −1.08, 2.16, −0.78, 1.32, −0.52)T ,compute the projection into the space defined by the first two principal components.

We will substitute these values into equation from A) as Xi values. These are standardized values so just based on them we can understand a great deal about Yugoslavia compared to other countries. 0 value means that it is just average, if we go into negative numbers then we are below average and the other way around.

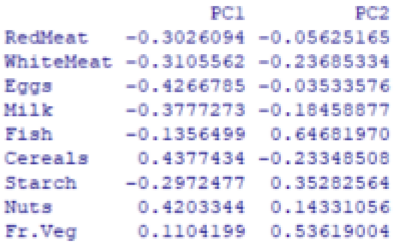
PC1 = -0.3026094 * -1.62 + (-0.3105562) * -0.78 + (-0.4266785) * -1.55 + … = 7(guessed)
PC2 = -0.05625165 * -1.62 + (-0.23685334) * -0.78 + (-0.03533576) * -1.55 + … = 2

Projection: Projection is transforming data from 3D into 2D space. We will create a cartesian coordinate system and PC1 is X axis denoted as Y1, PC2 is Y axis denoted as Y2 and we will put it into the graph and mark it as Yugoslavia. Yugoslavia [7,2]

(e) Based on the previous projection, what can you say about the protein consumption in Yugoslavia? Explain.

Projection states that Yugoslavia is much more affected by the PC1 so we will just reuse the explanation from A). Protein consumption in Yugoslavia is heavily based on intake of nuts and cereals, but with the addition of PC2 what is focused on fish. These values are arbitrary and would depend on specific values. But if we calculated this for the 7 or 8 countries we would see clusters of states based on their protein consumption.

Projection is split on coordinate system according to the principal components. PC1 splits countries into those that have their protein intake more focused around nuts and cereals yet lack eggs in their cusine. PC2 split coordinate system that top parts are countries with more fish intake and fruits and vegetables.



```
                  PC1         PC2
RedMeat   -0.3026094 -0.05625165
WhiteMeat -0.3105562 -0.23685334
Eggs      -0.4266785 -0.03533576
Milk      -0.3777273 -0.18458877
Fish      -0.1356499  0.64681970
Cereals    0.4377434 -0.23348508
Starch    -0.2972477  0.35282564
Nuts       0.4203344  0.14331056
Fr.Veg     0.1104199  0.53619004
```

Figure 1: Results of PCA from Task 4.

First exam

1.

(a) Give an example of a dataset (realisation of a random sample) with mean 5 and 0 skewness.

we can for example generate values from 0 and 10, keeping the data symmetric over the mean, so the skewness is 0

data: 0,10,5,2,8,1,2,3,9,8,7,6,4

(b) Let 15, 18, 15, 19, 16, 17, 13, 14 be a realisation of a random sample. Write down one nonparametric bootstrap sample based on the above realisation. Explain your approach.

We can use nonparametric boostrap to estimate the distribution, we generate bootstrap samples from the realized sample with replacement so an example nonparametric bootstrap sample: 15, 16, 17, 17, 18, 14, 15, 18

Each sample as uniform probability of being sampled (1/n) and each sampling is independent. So we just take values form default sample and blindly choose only values from it and repeat x times.

(c) For the following data compute 0.1, 0.4 and 0.8 - quantiles. Use the basic definition. Data: 5, 0, 7, 10, 5.

```
data <- c(5,0,7,10,5)
quantile(data, probs = c(0.1,0.4,0.8))
```

(d) For nominal variable hair_color with values blonde, brunette, red, and black construct corresponding dummy variables. How would you do that? How many variables do you need?

A dummy variable is a type of variable that we create in regression analysis so that we can represent a categorical variable as a numerical variable that takes on one of two values: zero or one to indicate the absence or presence of some categorical effect.

We can use blonde as the reference variable, the total number of distinc values is 4, so we need 4-1=3 dummy variables to express the color.
For example I will name them, is_brunette, is_red and is_black, if all of them are equal to 0, this would mean that the color is blonde.

(e) Let 0.548, 0.124, 0.874, 0.214, 0.495 be 5 independent realisations of a random variable with uniform U(0, 1) distribution. Based on it, estimate the value of the following integral from 0 to 1 sin(x)x dx. Explain your approach in detail.

We can use Monte Carlo integration, we take the 5 samples and compute the integrated function values, f(0.548), …

This way we get 5 values of the functions at the random points, their average is then the approximation of the integral. The more samples we have, the closer we get to the real value of the integral we were assigned.

```
realisations<-c(0.548, 0.124, 0.874, 0.214, 0.495)
n<- length(realisations)
fnc_results<- c()
for (i in 1:n){
  x <- realisations[i]
  fnc_results[i] <- sin(x)*x
}
fnc_results
sum(fnc_results)
sum(fnc_results)/n
```

Integral is equal to 0.2503389.

(f) Let (2, −1), (0, 0), (−1, −2), (−2, 3), (3, 2), (1, 1) be a realisation of a bivariate random sample. Compute the number of concordant and discordant pairs for the sample.

Any pair of observations (xi, yi) and (xj, yj), where i < j, are said to be concordant, if either xi > xj and yi > yj, or xi < xj and yi < yj.

Otherwise, they are said to be discordant.

(2, −1),(0, 0) -> 2 goes into 0, so it is decreasing, but -1 goes to 0, so it is increasing, thus it is disconcordant

```
tab <- as.table(rbind(c(2,0,-1,-2,3,1),c(-1,0,-2,3,2,1)))
library(asbio)
crabm<-ConDis.matrix(tab[1,],tab[2,])
concordant <- sum(crabm == 1, na.rm = TRUE)
disconcordant <- sum(crabm == -1, na.rm = TRUE)
```

(g) Consider the standardized residuals r ∗ 1 , . . . , r∗ n in linear regression model. Are they independent? Are they identically distributed? Explain.

Linear regression slide 31 (580 merged), Unlike model errors ei, ri are neither independent, nor identically distributed. Unlike model errors, the residuals are neither independent or i.i.d., due to the estimation of linear regression parameters, the residuals sum up to 0.

But standardized residuals are still not independent nor iid., but they have normal distribution, so they have all of these things only approximately, so they are almost i.i.d and almost independent.

(h) Let X1, . . . , Xn be a random sample from normal distribution N (μ, σ2). Derive in detail a formula for 99% confidence interval for parameter σ 2 .
Kukni v papieroch

2. (10 points)
Let 10, 15, 7, 2, 14, 11, 8 be a random sample from log-normal distribution with the unknown parameter μ ∈ R. It has the following pdf:

$$f(x) = \frac{1}{x\sqrt{2\pi}} \mathrm{e}^{-\frac{(\log x-\mu)^2}{2}}, \text{ for } x > 0.$$

(a) Write down likelihood function for your data.
Kukni papiere
(b) Write down log-likelihood function for your data.
(c) Find the maximum likelihood estimate of μ (either analytically or numerically).

3. (10 points)
Are IQ scores of the students from Faculty of Informatics above the population average? To find out, we randomly selected 10 students and measured their IQ. The obtained IQ scores are listed in the following table:

93 129 116 116 106 118 147 105 132 120

Can we claim that the average IQ of the students from Faculty of Informatics is greater than 100?
(a) Define the mathematical model – random variables and their distributions, null and alternative hypotheses, model assumptions. Explain all the symbols you use in detail.
Random Variable is X ~N(100,var) by the definition of IQ, we have normal data
H0: mu = 100; H1: mu >100
We use a one-sided test to see if information students have an average IQ > 100. We assume normality of our data and i.i.d. We can use one sample T-test.

(b) Are IQ scores of the students from Faculty of Informatics above the population average (100)?

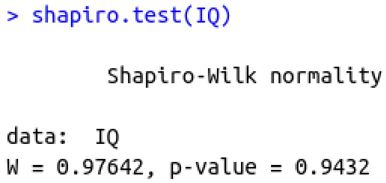
```
IQ<- c(93, 129, 116, 116, 106, 118, 147, 105, 132, 120)
t.test(IQ, mu=100, alternative = 'greater')
```

The mean of IT students is 118.2 so it is large, question is if it is statistically significant

(c) Check all the model assumptions you considered.

From the definition of IQ test and assuming every student is there only once, the sample points are iid and they should also be normal, we can test that with Shapiro-Wilk

shapiro.test(IQ)



```
> shapiro.test(IQ)

	Shapiro-Wilk normality test

data:  IQ
W = 0.97642, p-value = 0.9432
```

Since this value is not less than 0.05, we can assume the sample data comes from a population that is normally distributed.

(d) Construct the corresponding confidence interval. Use it for testing the null hypothesis.

CI [109.3292 ; Inf)

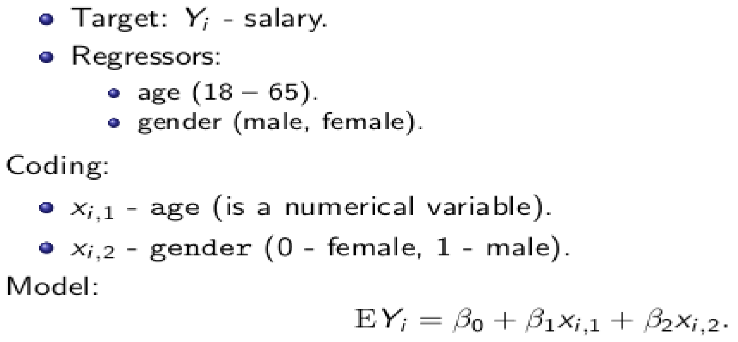
the right side is Inf as we are doing one sided testing. Our mean estimate is within the confidence interval, the p-value is also smaller than 5% (0.002239)

Thus we reject the H0 that the mean IQ of IT students is 100.

4. (10 points)

We randomly selected 100 people to investigate the relationship between their salary and age and gender. We will model salary with the aid of linear regression model. As possible regressors we will consider age and gender. Target: salary (in CZK), regressors: gender (1 - male, 0 - female), age (in years), and possibly its quadratic term. We fitted 2 different models: Model 1 with quadratic dependency (see Figure 1) and Model 2 with linear dependency (see Figure 2).

(a) For both models, write down the equation for the regression model.



- Target: Y_i - salary.
- Regressors:
  - age $(18 - 65)$.
  - gender (male, female).

Coding:

- $x_{i,1}$ - age (is a numerical variable).
- $x_{i,2}$ - gender (0 - female, 1 - male).

Model:

$$\mathrm{E}Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}.$$

Model1: salary = β_0 + β_1(gender) + β_2(age) + β_3(age)^2

Model2: salary =β_0 + β_1(gender) + β_2(age)

(b) For both models, compute the expected salary for 25-year old female.

Model1: salary = 3123.6563+ 3417.2701(gender) + 583.3339(age) + -0.3394(age)^2
Model2: salary =3573.13 + 3436.55(gender) + 556.93(age)

These models tell us what the baseline is for a person that is 0 years old and what is their salary on average and for each additional age year gained it would increase by coefficient close to age.

25years old
Model1: salary = 3123.6563+ 3417.2701*(gender) + 583.3339*(25) + -0.3394*(25)^2
Model2: salary =3573.13 + 3436.55*(gender) + 556.93*(25)
Female
Model1: salary = 3123.6563+ 3417.2701*(0) + 583.3339*(25) + -0.3394*(25)^2
Salary = 17494.8788
Model2: salary =3573.13 + 3436.55*(0) + 556.93*(25)
Salary = 17496.38

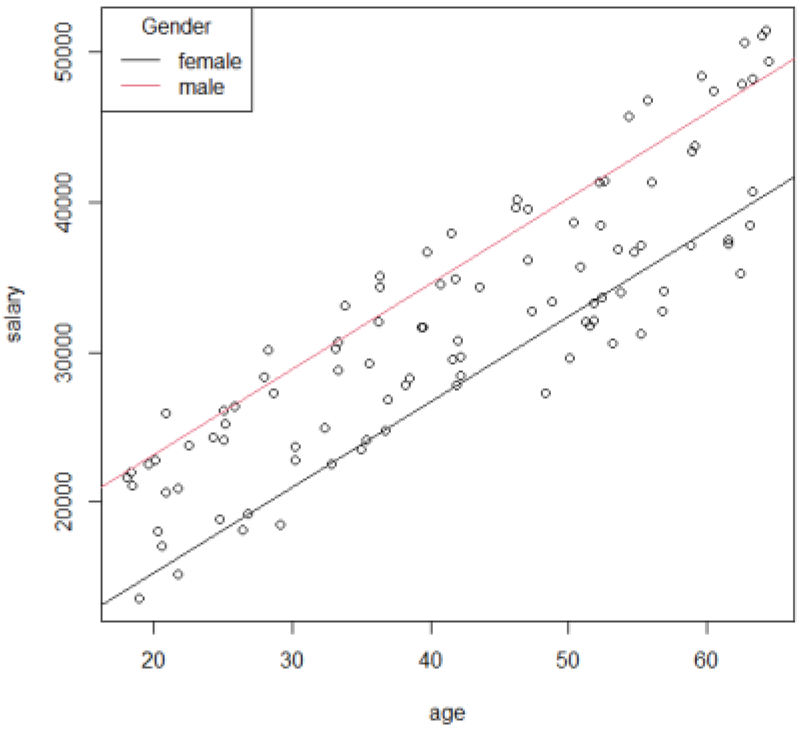
(c) When performing descriptive modeling, which model would you choose? Why?
Descriptive modeling is used to understand the relationship between response and regressors. Which components are significant, so we want a model that is as simple as possible, so interpretability is better.

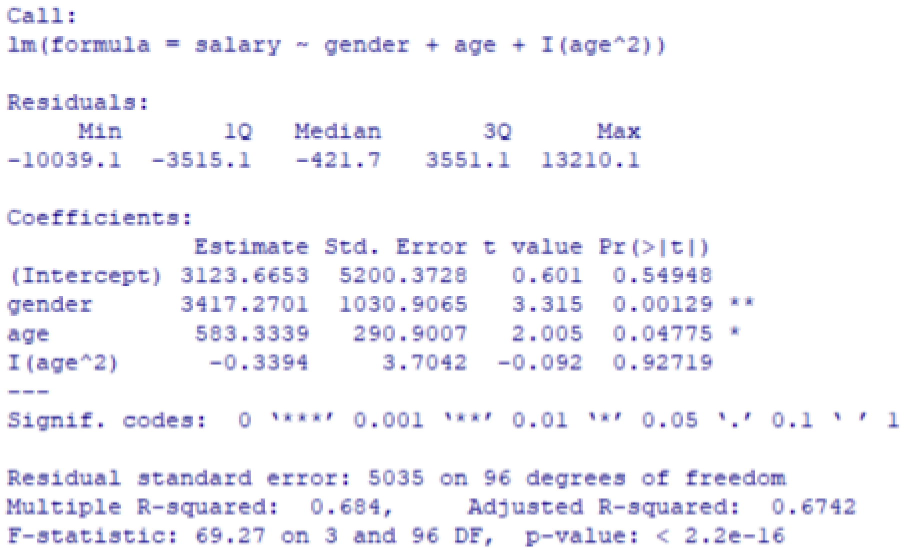
From the Adjusted R-squared value from the figures we can deduce that it is almost the same if we use model1 or model2. We look into adjusted one due to the fact that we have different numbers of regressors.
But we also need to consider Residual standard error and choose a model with the lower value, so in this case it would be Model2.

(d) For Model 2, does gender have a significant influence on salary? If yes, quantify it.
Pr(>|t|) and Signif. Codes. The p-value, in association with the t-statistic, help us to understand how significant our coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as significant. It means we are confident that the coefficient is **not** zero, meaning the coefficient does in fact add value to the model by helping to explain the variance within our dependent variable.

Yes because the p.value is smaller than 0.05. So if a person is a man his salary is higher by 3436.55 Czk on average.

(e) Visualise (plot) Model 1.





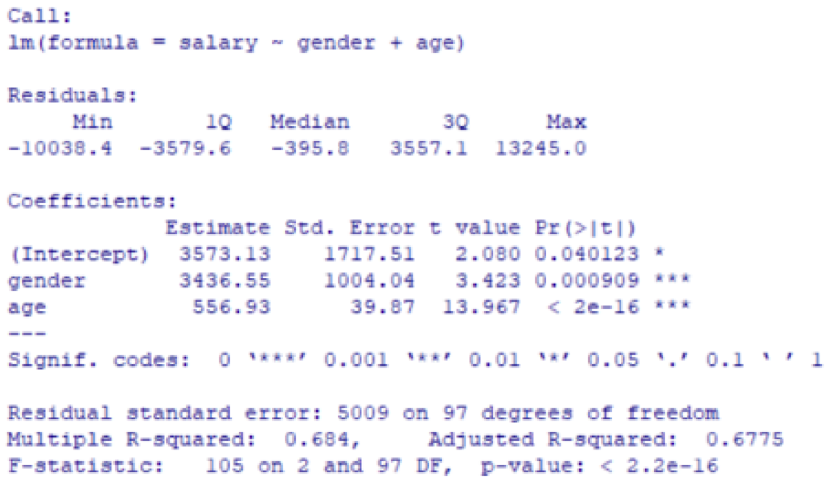
```
Call:
lm(formula = salary ~ gender + age + I(age^2))

Residuals:
     Min       1Q   Median       3Q      Max
-10039.1  -3515.1   -421.7   3551.1  13210.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3123.6653  5200.3728   0.601  0.54948
gender      3417.2701  1030.9065   3.315  0.00129 **
age          583.3339   290.9007   2.005  0.04775 *
I(age^2)      -0.3394     3.7042  -0.092  0.92719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5035 on 96 degrees of freedom
Multiple R-squared:  0.684,     Adjusted R-squared:  0.6742
F-statistic: 69.27 on 3 and 96 DF,  p-value: < 2.2e-16
```

Figure 1: Model 1 from Task 4.



```
Call:
lm(formula = salary ~ gender + age)

Residuals:
     Min       1Q   Median       3Q      Max 
-10038.4  -3579.6   -395.8   3557.1  13245.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3573.13    1717.51   2.080 0.040123 *  
gender       3436.55    1004.04   3.423 0.000909 ***
age           556.93      39.87  13.967  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5009 on 97 degrees of freedom
Multiple R-squared:  0.684,     Adjusted R-squared:  0.6775 
F-statistic:   105 on 2 and 97 DF,  p-value: < 2.2e-16
```

Figure 2: Model 2 from Task 4.

Second exam

1

d) Give one example of a random variable with exponential Exp(0.02) distribution. What is its expectation? Interpret it for your particular example.

It is used for modeling waiting times in a queue. The Poisson rate for X Exp(0.02) will therefore be 0.02. The event occurs 0.02 times on average during a unit of time, whether you can take it as a second, minute, hour, week, month or year.

If I am considering minutes as my unit of time, it will take 50 minutes which is a reciprocal of 0.02 for the event to occur, for example getting to the front of the queue.

e) Let u = 0.961 be a realization of a random variable with uniform U(0; 1) distribution. Based on u, generate a realization of a random variable with binomial Binomial(2; 0.2) distribution. Explain your approach in detail.

We can use inverse transform method, we just need the inverse of Binomial CDF
Formally, we want to get a realization of a random variable X with known cdf F.

We can use inverse transform method, we just need the inverse of Binomial CDF

pbinom Binomial distribution - (Cumulative distribution function)
qbinom Binomial quantile function

```
qbinom(p = 0.961, size = 2, prob = 0.2)
= 2
```

3. (10 points) Fuel economy is the relationship between the distance traveled and ` consumed, generally expressed in liters per 100 kilometers (l=100 km). A producer claims that the fuel economy of its car is 5,6 l=100 km. To check it, we performed 8 independent trials (measurements) and got the following results:

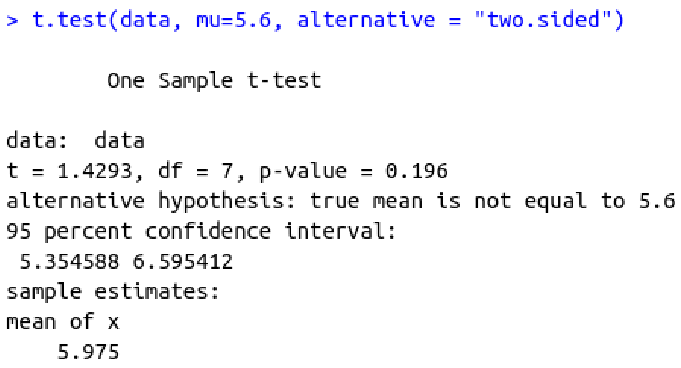
6.5, 5.2, 5.6, 6.8, 5.0, 5.6, 6.1, 7.0

(a) Does the producer keep the declared fuel economy? Explain in detail. Do not forget to define the mathematical model { random variables and their distributions, null and alternative hypotheses, model assumptions and all the symbols you use. You do not need to check the assumptions.

H0: mu=5.6 ; H1: mu! = 5.6
X 1 , . . . , X n is a random sample from normal distribution N (μ, σ^2), where μ is parameter of our interest. (σ^2 is the nuisance parameter)
X ~ N (5.6, σ^2)

Test statistic $\sqrt{n}\frac{\overline{X}-\mu}{\sqrt{S^2}} = \sqrt{n}\frac{\overline{X}-\mu}{S} \sim \mathrm{t}(n-1).$

```
t.test(data, mu=5.6, alternative = "two.sided")
```



```
> t.test(data, mu=5.6, alternative = "two.sided")

        One Sample t-test

data:  data
t = 1.4293, df = 7, p-value = 0.196
alternative hypothesis: true mean is not equal to 5.6
95 percent confidence interval:
 5.354588 6.595412
sample estimates:
mean of x 
    5.975 
```

We do not reject null hypothesis on the 95% significance level because the p value is higher than 0.05.

4) We would like to compare some numerical quantities in 5 different groups A,B,C,D,E. First, we performed 10 two-sample t-tests for each pair of groups. The corresponding p-values are:

A-B	**A-C**	**A-D**	**A-E**	**B-C**	**B-D**	**B-E**	**C-D**	**C-E**	**D-E**
0.012	**0.009**	**0.95**	**0.120**	**0.920**	**0.004**	**0.219**	**0.003**	**0.177**	**0.075**

(a) Based on the previous t-tests, which groups do significantly differ on average? Consider the overall significance level = 0:05. Explain your approach.

If we reject the null hypothesis at least for one pair, then there is a difference between the groups. In addition, we can see which groups differ.

We reject the null hypothesis if p.value is < significance level. So each pair of groups with p.value lower than 0.05 differ.

A-B A-C B-D C-D. These pairs differ

Third exam

1.(3 + 3 + 3 + 4 + 4 + 4 + 4 + 5 = 30 points)

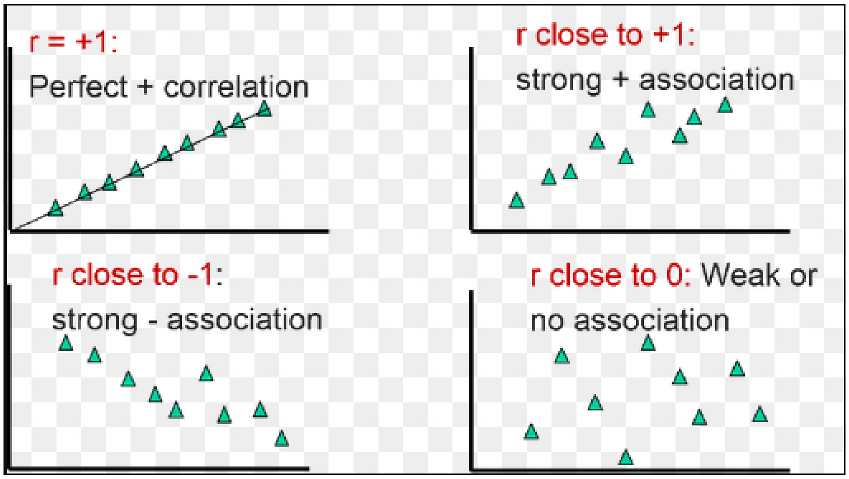
(a) Give an example of a contingency table for which the contingency coefficient is 0.

```
data<- c("a","a","b","b","c","c","d","d")
                  table(data)
 contigency_table <- table(data)
 Contingency <- function(x) {
           chi <- chisq.test(x)
           unname(sqrt(chi$statistic / (chi$statistic + sum(x))))
 }
    Contingency(contigency_table)
```

Visualization - table

a	b	c	d
2	2	2	2

(b) Draw 3 scatter plots for two-dimensional data for that correlation is close to −1, 0 and 1 respectively.



(c) Give two particular examples of a problem (with data) when integrating data about people's height from various data sources.

(d) For the following data (education level), compute the mean, the median, the first and the third quartile. Data:
`none`, `elementary`, `university`, `elementary`, `high`, `elementary`, `elementary`.

(e) Let u_1 = 0.146 and u_2 = 0.588 be two independent realisations of a random variable with uniform U(0, 1) distribution. Based on u_1 and u_2, generate a realisation of a random variable with $X^2(2)$ distribution. Explain your approach in detail.

First we need to generate from uniform distribution to the normal distribution with the use of Box-Muller method

$$X_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2),$$
$$X_2 = \sqrt{-2 \log U_2} \cos(2\pi U_1).$$

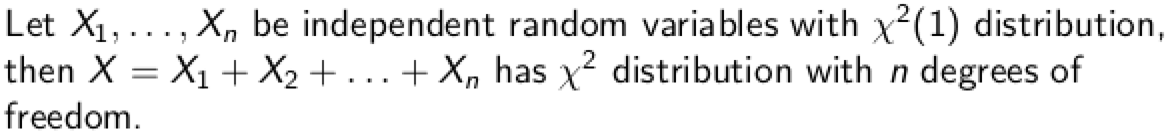
```
u1<-0.146
u2<-0.588
x1 <- sqrt(-2*log(u1))*cos(2*pi*u2)
x2 <- sqrt(-2*log(u2))*cos(2*pi*u1)
```

X1 = -1.669402
X2 = 0.6265095

Only after that we can take the realization and square it to get X^2 distributions with one degree of freedom.

```
x1^2
x2^2
```

X1^2 = 2.786905
X2^2= 0.3925141

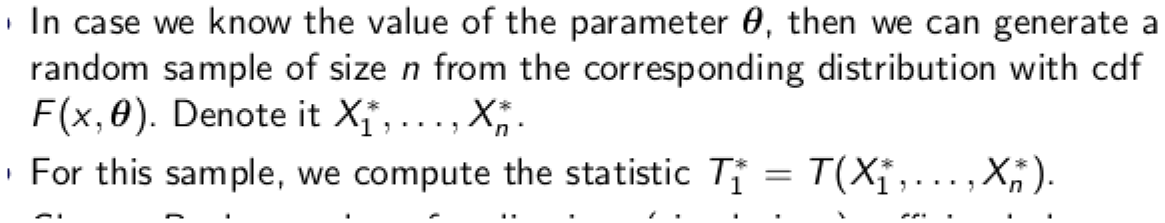


Let $X_1, \ldots, X_n$ be independent random variables with $\chi^2(1)$ distribution, then $X = X_1 + X_2 + \ldots + X_n$ has χ^2 distribution with n degrees of freedom.

```
x1^2 + x2^2
```

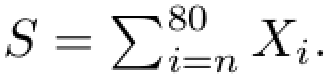
X = X1 + X2 = 3.179419
X is a random variable with 2 degrees of freedom.

(f) Let
8, 12, 6, 6, 5, 7
be a realization of a random sample. Write down one Monte Carlo sample based on the above realization (when assuming that the data come from Binomial(20, 0.3) distribution). Explain your approach.



- In case we know the value of the parameter θ, then we can generate a random sample of size n from the corresponding distribution with cdf $F(x, \theta)$. Denote it $X_1^*, \ldots, X_n^*$.
- For this sample, we compute the statistic $T_1^* = T(X_1^*, \ldots, X_n^*)$.

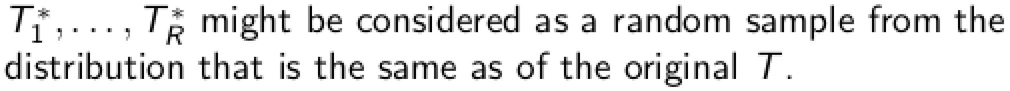
So we first need to know binomial test statistic - like this but not with 80 but only 20.



$$S = \sum_{i=n}^{80} X_i.$$

8 (X_1), 12 (X_2), 6 (X_3), 6 (X_4), 5 (X_5), 7 (X_6)

T_1* = 8 + 12 + 6 + 6 +5 + 7 = 44



$T_1^*, \ldots, T_R^*$ might be considered as a random sample from the distribution that is the same as of the original T.

And if we want to generate new T_r* then we can generate new data for this distribution and calculate its test statistics lke this .

```
random_samples<-matrix(rbinom(6,size=20, prob=0.3))
random_samples
```

```
sum(random_samples) / length(random_samples)
```

(g) Let
21, 20, 9, 17, 5, 4, 8
be a realisation of a random sample. Derive and compute the value of Wilcoxon signed rank test statistic for testing the null hypothesis that the median m is equal to 10.

$$W^+ = \sum_{i:X_i > m_0} R_i^+ .$$

Xi	21	20	9	17	5	4	8
Xi -10	11	10	-1	7	-5	-6	-2
|Xi-10|	11	10	1	7	5	6	2
Ri+	7	6	1	5	3	4	2

W+ = (7+6+5) = 18
We summed only those signed ranks of Xi higher than the median.

```
library(stats)
data<-c(21, 20, 9, 17, 5, 4, 8)
wilcox.test(data, mu =10, alternative = "two.sided", correct = T)
```

(h) Let $X_1, \dots, X_n$ be a random sample from normal distribution N (μ, σ_2). Derive in detail a formula for critical region W for testing H_0 : $\sigma_2 = 10$ against one-sided alternative H_1 : $\sigma_2 < 10$.

2. (10 points) Let
4, 5, 1, 1, 5, 5, 4
be a random sample from Binomial(20, p) distribution with the unknown parameter p ∈ (0, 1).
(a) Find the maximum likelihood estimate of p (analytically or numerically). Explain your approach in detail.
Kukni v papieroch
(b) Find the maximum likelihood estimate of the parametric function $20p(1 − p)_{19}$.

(c) How would you interpret the parametric function $20p(1 − p)_{19}$? What does it stand for?

3. (10 points) Is the probability of having a baby boy the same as a baby girl? To decide, we randomly selected 500 newborn babies; 279 of them were boys, and 221 were girls.

(a) Define the mathematical model – random variables and their distributions, null and alternative hypotheses, model assumptions. Explain all the symbols you use in detail.

H0: p = 0.5 ; H1: p != 0.5

$$X_i = \begin{cases} 1 & \text{newborn is male} \\ 0 & \text{newborn is female} \end{cases}$$

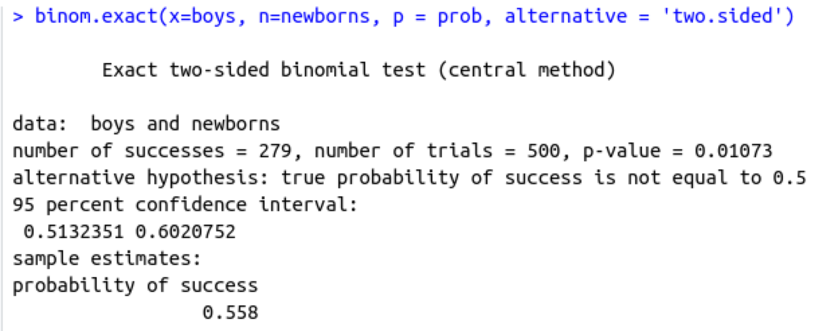
Xi is a random sample from Bernouilli distribution with parameter p.(probability that the newborn is a male).

Test statistics is $S = \sum_{i=n}^{80} X_i.$

Under H0 S ~ Binomial(500,0.5).

(b) Use the exact test. What is its actual (real) significance level? What is your Conclusion?

boys<-279
girls<-221
newborns<-500
prob<-0.5
binom.exact(x=boys, n=newborns, p=prob, alternative = 'two.sided', conf.level = 0.95)



```
> binom.exact(x=boys, n=newborns, p = prob, alternative = 'two.sided')

        Exact two-sided binomial test (central method)

data:  boys and newborns
number of successes = 279, number of trials = 500, p-value = 0.01073
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5132351 0.6020752
sample estimates:
probability of success
                 0.558
```

P.value is not higher than our significance level or 0.05, so we reject the null hypothesis.

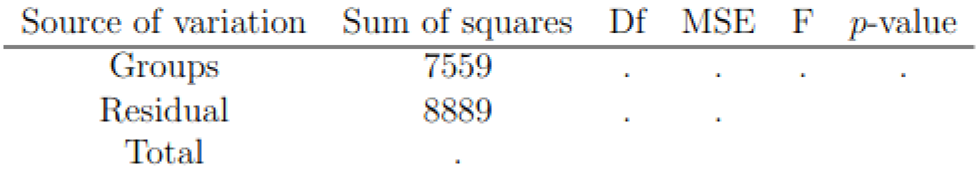
(c) Construct 95% confidence interval for the probability of having a baby boy. Indicate which one you used.

95 percent confidence interval: [0.5132351, 0.6020752]

4. (10 points) We would like to compare the IQ scores of the students from Faculty of

Informatics, Faculty of Law and Faculty of Science. We randomly selected 6 students from Faculty of Informatics, 8 from Faculty of Law and 9 from Faculty of Science and measured their IQ. Is there any significant difference between the IQ scores of the students from these three faculties?

For our problem, we decided to use ANOVA model. First, we checked all the assumptions and performed the analysis. Corresponding results are summarized in the following incomplete table:



Source of variation	Sum of squares	Df	MSE	F	p-value
Groups	7559	.	.	.	.
Residual	8889	.	.		
Total	.				

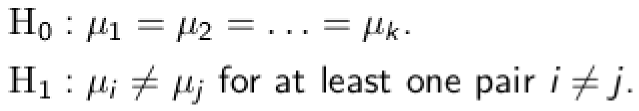
(a) Add the model assumptions and formulate null and alternative hypotheses.

Hypothesis is that the average of IQ is the same for each faculty.

H0: mu_1 = mu_2 = mu_3;   H1 : mu_i != mu_j for at least one pair i != j

Assumptions:

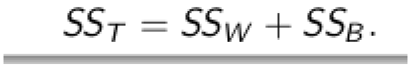
- All random samples are independent.
- All random samples are following normal distribution
- All random samples have the same variance - homoscedasticity.



$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k.$$

$$\mathrm{H}_1 : \mu_i \neq \mu_j \text{ for at least one pair } i \neq j.$$

(b) Complete the ANOVA table.



$$SS_T = SS_W + SS_B.$$

Total (SS_T) = 7559 + 8889 = 16448

SSw residuals

SSb groups

DFs are denominators

N- number of observations

K - is number of groups

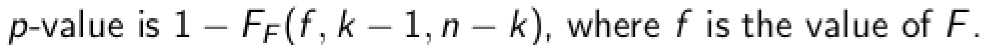
SS_w = n - k = (6+8+9) - 3 = 20

SS_b = k - 1 = 3-1 = 2

MSEs are the sum of squares divided by degrees of freedom. MSE is an estimate of sigma squared.
SS_w = 7559 / 20 = 377.95
SS_b = 8889 / 2 = 4444.5

For F we need to divide the MSE of Groups by the MSE of residuals.
F = 4444.5 / 377.95 = 11.759491996

For p.value we need to plug it into the formula.



p-value is $1 - F_F(f, k-1, n-k)$, where f is the value of F.

1-pf(11.759491996, 2, 20) = 0.0004202546

(c) What is the conclusion based on this ANOVA model? Formulate it both mathematically and with your own words.
We reject hypothesis that mean value of IQ across faculties is the same.

That means that there is a difference in IQ between students on different faculties and their mean is not the same. The result is really significant, thus there is probably high discrepancy between the student based on p.value measurement.

Fifth exam

1. (3 + 3 + 3 + 4 + 4 + 4 + 4 + 5 = 30 points)

(a) Let X be a random variable representing the number of heads when flipping a fair coin repeatedly until the first tail falls. What is the distribution of X? Name it and specify the value of the parameters.
X~ Geometric(n,p)

(b) Draw 3 boxplots: one for symmetric data with no outliers, the second for symmetric data with 2 outliers and the third for asymmetric data.

(c) Write down an example of nominal, ordinal and numerical variable from IT environment (one for each case).

(d) For the following data (`age at graduation`) compute one measure of location and one measure of variability (only one, the most suitable for the data). Explain your choice. Data:
22, 21, 22, 22, 23, 22, 25, 50, 22, 23.

(e) Let u = 0.754 be a realisation of a random variable with uniform U(0, 1) distribution. Based on u, generate a realisation of a random variable with Binomial(20, 0.1) distribution. Explain your approach in detail.

U(0,1) says that this is continuous uniform distribution

Formally, we want to get a realization of a random variable X with known cdf F.

We can use inverse transform method, we just need the inverse of Binomial CDF

pbinom Binomial distribution - (Cumulative distribution function)

qbinom Binomial quantile function

qbinom(p = 0.754, size = 20, prob = 0.1) = 3

(f) Let

5.9, 6.2, 5.9, 5.7, 5.6, 5.6, 6.7, 5.9

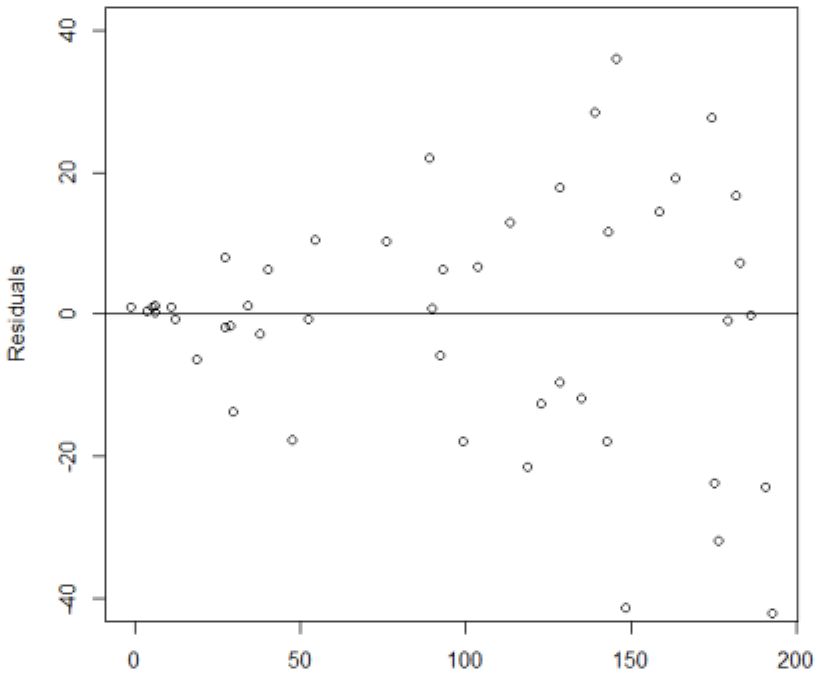
be a realisation of a random sample from normal distribution N (μ, σ_2). Compute (do not use R function) 99% confidence interval for σ_2.

Kukni v papieroch

(g) Draw two regression diagnostic plots for checking that all relevant regressors are present in the model in the right form: one indicating good model, while the second one indicating wrong model. Do not forget to describe both axes and explain what you see in the plots.

WRONG

On X axis should be on the regressor X1. With higher values, variability increases, so there is dependency on regressor X1 and its values increase variability.



Good

On X axis should be on the regressor X1. Dependency on regressor X1 and its values increase variability should not be significant so with increasing values spread of data around the 0 line should be the same..

(h) Let
44, 53, 81, 2, 46
be a realisation of a random sample from exponential distribution. Derive the value of AIC for the data.

Kukni v papieroch

2. (10 points) Let
17, 15, 53, 52, 93, 11, 95, 100, 2, 99
be a realization of a random sample. Our goal is to test the null hypothesis H_0 : m = 50, that the hypothetical median is equal to 50, against H_1 : m != 50 with the aid of Wilcoxon test.
(a) Give two arguments that the data do not look normal.

```
data<- c(17, 15, 53, 52, 93, 11, 95, 100, 2, 99)
hist(x=data, breaks= seq(1,105,by=20))
qqnorm(data)
qqline(data)
shapiroTest(data)
skewness(data)
kurtosis(data)
```

Kurtosis is not 0, thus there is tailness in data. Histogram shows data which are divided into three columns that don’t follow normal dist.
The Shapiro test has a p.value lower than 0.05, so we reject the null hypothesis that data follows normal distribution.

(b) Derive the value of the test statistic of Wilcoxon test.

$$W^{+} = \sum_{i:X_i > m_0} R_i^{+}.$$

Xi	17	15	53	52	93	11	95	100	2	99
Xi -50	-33	-35	3	2	43	-39	45	50	-48	49
|Xi-50|	33	35	3	2	43	39	45	50	48	49
Ri+	3	4	2	1	6	5	7	10	8	9

W+ = (2 + 1 +6 +7 +10 +9) = 35
We summed only those signed ranks of Xi higher than the median.

(c) Perform Wilcoxon test and comment on the results.

```
library(stats)
wilcox.test(data, mu =50, alternative = "two.sided", correct = T)
```

Correct= False would be used if we want to use an asymptotic test. And that is to be used when we have multiple values with the same rank. But we have unique values so it is ok to use correct=true.

p-value = 0.4922 so we dont reject the null hypothesis.

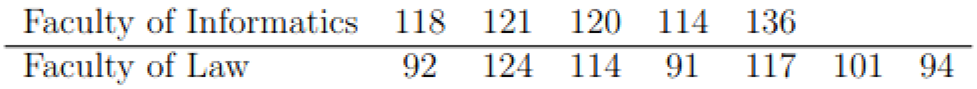
If it was a sign test then

```
library(BSDA)
SIGN.test(data, md = 50, alternative = "two.sided")
```

P.value is higher than 0.05 so we don’t reject the null hypothesis. If the value in the dataset is equal to median then we need to skip those values in the computation for the sign tests.

(d) Add the model assumptions you need for performing Wilcoxon test (you do not need to check it).
X 1 , . . . , X n is a random sample from a continuous distribution that is
symmetric around the median m.

3. (10 points) Are students from Faculty of Informatics smarter than students from Faculty of Law? To decide, we randomly selected 12 students from both faculties

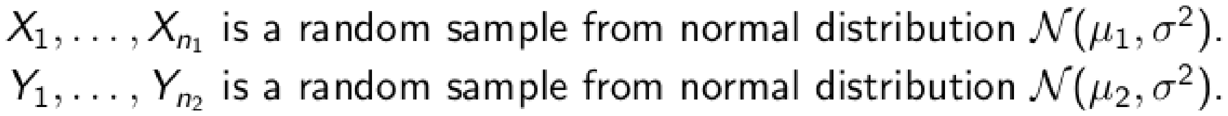


Faculty of Informatics	118	121	120	114	136		
Faculty of Law	92	124	114	91	117	101	94

Do students from Faculty of Informatics have on average higher IQ than students from the Faculty of Law?

(a) Define the mathematical model – random variables and their distributions, null and alternative hypotheses, model assumptions. Explain all the symbols you use in detail.

Mu1 = it
Mu2 = law

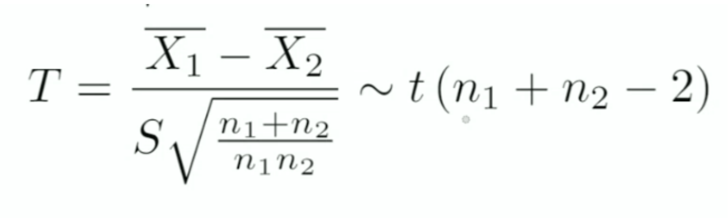


$X_1, \ldots, X_{n_1}$ is a random sample from normal distribution $\mathcal{N}(\mu_1, \sigma^2)$.
$Y_1, \ldots, Y_{n_2}$ is a random sample from normal distribution $\mathcal{N}(\mu_2, \sigma^2)$.

Assumptions: Both samples are mutually independent.

H0: mu1=mu2; H1: mu1 > mu2

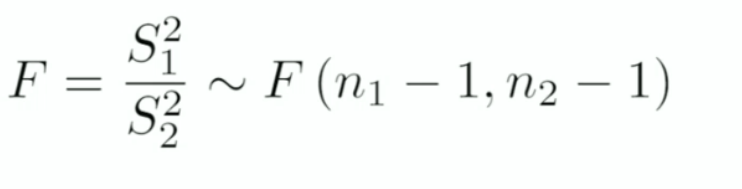
Test statistic:



$$T = \frac{\overline{X_1} - \overline{X_2}}{S\sqrt{\frac{n_1+n_2}{n_1 n_2}}} \sim t\left(n_1 + n_2 - 2\right)$$

b) Do students from Faculty of Informatics have on average higher IQ than students from Faculty of Law? Use a suitable statistical test and comment on your conclusions. You do not need to check the model assumptions.

For two sample test we expect to have equal variances. In order to check we first need to set this H0 to H0: σ1^2 / σ2^2 = 1. We need to check this first only then we can proceed to check our hypothesis.

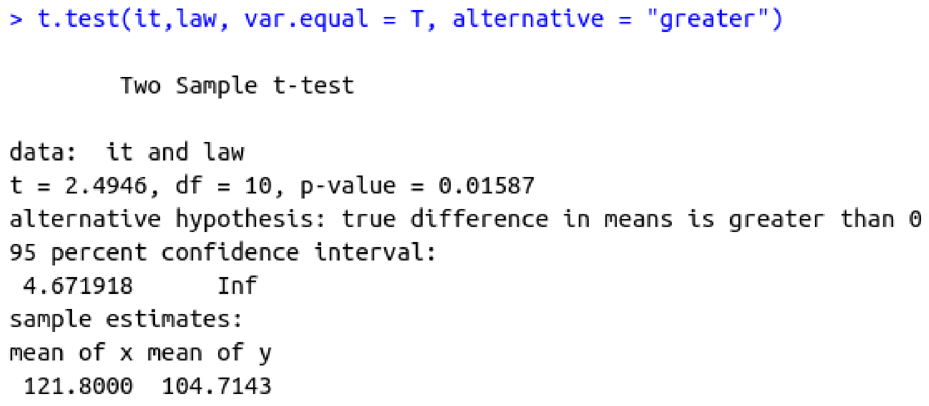


$$F = \frac{S_1^2}{S_2^2} \sim F\left(n_1 - 1, n_2 - 1\right)$$

it<-c(118,121,120,114,136)
law<-c(92,124,114,91,117,101,94)
F <- var.test(it, law, alternative = "two.sided")

CI does not contain 0, but H0 is formulated in a way that we look for 1. So if 1 is in CI then it is ok.

t.test(it,law, var.equal = T, alternative = "greater")



```
> t.test(it,law, var.equal = T, alternative = "greater")

	Two Sample t-test

data:  it and law
t = 2.4946, df = 10, p-value = 0.01587
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.671918      Inf
sample estimates:
mean of x mean of y 
 121.8000  104.7143 
```

We reject the null hypothesis based on p.value that is lower than 0.05.

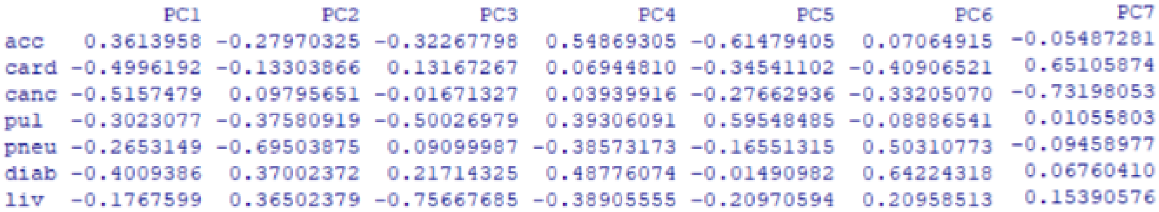
4. (10 points) For a dataset containing data about health (number of deaths caused by an accident, cardiovascular disease, cancer, pulmonar disease, pneumonia, diabetes and liver disease per 100 000 inhabitants) in 50 US states, we performed PCA with sample correlation matrix. Its eigenvalues are
3.401, 1.229, 1.061, 0.611, 0.434, 0.216, 0.047
and the corresponding eigenvectors are displayed in Figure 1.
(a) What is the equation for the first principal component. How can you interpret it?
2

(b) How can you interpret the second principal component?

(c) How much of the variability is explained by the second principal component?

(d) For Alaska with standardized values
$(3.78, −3.37, −3.10, −3.25, −2.09, −3.12, 0.16)^T$,
compute the projection into the plane defined by the first two principal compoNents.

(e) Based on the previous projection, what can you say about the health in Alaska? Explain.



```
            PC1         PC2         PC3         PC4         PC5         PC6         PC7
acc   0.3613958 -0.27970325 -0.32267798  0.54869305 -0.61479405  0.07064915 -0.05487281
card -0.4996192 -0.13303866  0.13167267  0.06944810 -0.34541102 -0.40906521  0.65105874
canc -0.5157479  0.09795651 -0.01671327  0.03939916 -0.27662936 -0.33205070 -0.73198053
pul  -0.3023077 -0.37580919 -0.50026979  0.39306091  0.59548485 -0.08886541  0.01055803
pneu -0.2653149 -0.69503875  0.09099987 -0.38573173 -0.16551315  0.50310773 -0.09458977
diab -0.4009386  0.37002372  0.21714325  0.48776074 -0.01490982  0.64224318  0.06760410
liv  -0.1767599  0.36502379 -0.75667685 -0.38905555 -0.20970594  0.20958513  0.15390576
```

Figure 1: Results of PCA (eigenvectors) from Task 4.