

Institute of Mathematics and Statistics
Faculty of Science
Masaryk University

Statistics for Computer Science

Assignment
spring semester 2020

Stanislav Katina and Markéta Janošová

`katina@math.muni.cz`

May 11, 2020

Instructions

Template. You can find sample template for the assignment in the file `mj-template-eng-2020.pdf`, it was generated using file `mj-template-eng-2020.Rnw`. Keep the style used in the template when creating the pdf file you will be submitting. Comments marked `%%%` denote lines where you need to fill your information.

File `mj-template-eng-2020.Rnw` is a Sweave file, Sweave combines RScript with L^AT_EX.

More information about L^AT_EX can be found for example here: The Not So Short Introduction to L^AT_EX.

Further information about Sweave can be found for example here: Chunk options and package options.

When writing your R-code, follow the instructions in the presentation **Standards of programming in R: R style guide**.

For each part of the assignment you will submit two files:

1. `UCO-surname-firstname-STAT-CS-2020-part1.pdf` (or `-part2.pdf`) which contains your solutions of exercises, tables, figures, comments and preview of R-code,
2. `UCO-surname-firstname-STAT-CS-2020-part1.Rnw` (or `-part2.pdf`), which is the Sweave file that was used to generate the final pdf file and contains all necessary code to each exercise. This file needs to be **compilable** in the form you submit.

Do not use diacritics when naming your files and abide uppercase and lowercase letters as set by the instructions.

Each student submits his/her own solution. This is NOT a team project.

Assignment is marked on the following:

1. *presence of all above mentioned files and their names (when uploading the files, DO NOT choose 'Insert UCO, Surname and First Name' and upload **individual files**, NOT *.zip, *.rar or other archives!),*
2. *completeness and correctness of solutions,*
3. *sufficient description of your thought process, your approach and your interpretations of results, either graphics or tables,*
4. *clarity of R-code and compliance with the instructions in presentation **Standards of programming in R: R style guide**.*

Assignment needs to be submitted into the Homework vault ('Odevzdávárna') by the date specified with each part. If you fail to do so, you get 0 points for that part.

*In order to be allowed to sit the exam, you need to gain at least 60 % of overall points from the assignments. If you score less than 60 %, you will get mark **X** and cannot sit the exam.*

Assignment - Part 1

Deadline for these exercises is **13 April 2020 (incl.)**.

Exercise 1 File *Howell.csv* contains craniometric measurements from several populations. We are interested in maximum cranial breadth (variable *XCB*, in mm) of males (denoted *M* in variable *Sex*) from populations *AUSTRALI* and *PERU*. Missing values in this database are coded as 0.

1. Create your own **R**-functions for calculating estimates of the following characteristics: sample mean, sample five number summary, sample skewness, sample kurtosis, sample variance, sample standard deviation, sample range, sample decile range, sample 0.1-trimmed average and sample 0.1-trimmed variance. Use Cramér's estimates for skewness and kurtosis. Do not use in-built functions *min()*, *max()*, *mean()*, *quantile()*, *var()*, *sd()*, *range()* or functions from external libraries.
2. Separately for each population calculate sample size and these characteristics of maximum cranial breadth. Print them in a table with values rounded to 4 decimal places.
3. Create boxplots of maximum cranial breadth of each population side by side in one figure. Set the width of boxes to be proportional to sample sizes, add notches and arithmetic averages for both groups.
4. Separately for each population create histogram of maximum cranial breadth. Make sure that the histograms can be easily compared (without using back-to-back histogram).
5. Create normal qq-plot of maximum cranial breadth for each population.
6. Interpret your results and graphics.

Don't forget to properly label your plots.

Source of data: The William W. Howells Craniometric Data Set <https://web.utk.edu/~auerbach/HOWL.htm>

Exercise 2 File *area_spanish_provinces.csv* contains the area of each spanish province (in km²). File *population-spain-1998-2018.csv* contains population numbers for each province in years 1998, 2003, 2008, 2013 and 2018.

1. Calculate the number of men and women in Spain for each year and print them in a table together with the total population.
2. Display barplot plot of total population of Spain in each of the years, with each bar divided between men and women.
3. Display barplot of relative proportions of men and women within each province in 2018.
4. Calculate population density (= number of people per km²) in each province in 1998 and in 2018.
 - (a) For both years calculate sample size and estimates of these characteristics of population density: sample mean, sample five number summary, sample skewness, sample kurtosis and sample standard deviation (using functions you created for the previous exercise). Print them in a table with values rounded to 4 decimal places.

(b) Create boxplot of population density in 1998 and in 2018.

(c) Create histograms of population density in 1998 and in 2018. Make sure that the histograms can be easily compared.

5. Interpret your results and graphics.

Don't forget to properly label your plots.

Source of data: Instituto Nacional de Estadística (INE.es).


Assignment - Part 2

Deadline for these exercises is **24 May 2020 (incl.)**.

Exercise 3 (6 points) Over 20 days we noted the number of customers that made purchase at a local fruit and vegetable shop. We want to model this variable by Poisson distribution.

day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
customers	117	109	109	89	120	88	99	103	109	91	107	101	109	117	96	95	129	96	105	98

Table 1: Number of customers of the shop during 20 days

1. Write down the formula for likelihood function of Poisson distribution.
2. Write down the formula for log-likelihood function of Poisson distribution.
3. Write down the likelihood equation and work out the exact formula for $\hat{\lambda}$.
4. Create your own -function for calculating the value of log-likelihood function of Poisson distribution.
5. Using function `optimize()` find $\hat{\lambda}$. Compare it to the estimate you get from the exact formula.
6. Plot the log-likelihood function, highlight the maximum and denote the maximum likelihood estimate in plot margin.

Exercise 4 (8 points) File `body-measurements.txt` contains body measurements of young adults: body weight (`body.W`, in kg), body height (`body.H`, in mm), waist circumference (`waist.C`, in mm), hip circumference (`hip.C`, in mm), wrist circumference (`antb.C`, in mm) and neck circumference (`neck.C`, in mm), we also know their sex. We are interested in body height and neck circumference of females. Assuming the variables follow bivariate normal distribution, test hypothesis that they are uncorrelated using test based on Fisher Z-variable on $\alpha = 0.05$.

1. Write down the null and the alternative hypotheses in mathematical form.
2. Calculate the value of test statistic.
3. Calculate the critical region and make your decision.
4. Calculate the p-value and make your decision.

5. *Plot the density of the distribution, that the test statistic follows, and visualise the p-value.*
6. *Calculate the confidence interval for ρ and make your decision.*
7. *Interpret your conclusion.*

Results must be present in your pdf and rounded to 5 decimal points.

Exercise 5 (6 points) *Two factories are manufacturing the same product. In a sample of 200 products from the first factory, 32 were faulty. In a sample of 230 products from the second factory, 21 were faulty. Using Wald test statistic on $\alpha = 0.05$, test hypothesis that the probabilities of getting a faulty product are the same in both factories.*

1. *Write down the null and the alternative hypotheses in mathematical form.*
2. *Calculate the value of test statistic.*
3. *Calculate the critical region and make your decision.*
4. *Calculate the p-value and make your decision.*
5. *Calculate the confidence interval and make your decision.*
6. *Interpret your conclusion.*

Results must be present in your pdf and rounded to 5 decimal points.