

Project in Data analysis

Instructions

Below there are five tasks described; answer to all of these. The first four tasks consist of shorter problems, each of which is answered with an R program code that solves the problem and with possible written explanations.

The fifth task is a broader research task that corresponds to half of the whole scope of the assignment. The fifth problem is answered with a PDF report in which in addition to graphs and analytical considerations, the research question, the material used, how the research was made, what methods (which may include, for example, appropriate statistical tests) used and what are the results in a plain English to the reader of the report. In addition, the R program code used in the report must also be returned. The materials used for the tasks can be downloaded from the course's Moodle page.

Follow these steps in order to organize your submission. The submitted solutions to tasks 1-4, **one R program script in which all these tasks are clearly separated from each other as described in the example on the last page of this assignment document**, and which includes in addition to the codes, also the necessary written answers in comments.

If in any task you are asked to extend the code from the previous task, then copy-paste the codes and continue from there. Submission of the broader research assignment 5, a PDF report is provided as well as a clearly organized R code script which you used to while preparing it, which is a separate file from the R code script you provided as a submitted solution to tasks 1-4. The submission consist, therefore, of exactly three files (one PDF file and two R code script files)

Note: There are several different versions of the assignment document, and not all of the students completing the course at the same time may not receive the same assignments.

So don't answer to the tasks using your friend's assignment. Instead always fetch the tasks from the Moodle logged in with your own credentials. The purpose of the exercise is test your skills independently, so absolutely **no collaboration with friends is allowed in any form!** Plagiarism and fraud, for example copying responses from a friend and returning them as your own, work is denied. The submission can be used sent to Urkund plagiarism detection system and any the cheating that is discovered will always leads to the rejection of the course.

Data sets for tasks 1-2

The data file `pkseutu2015.csv` contains the results of the 2015 parliamentary elections in the Helsinki Metropolitan Area voting areas (it contains information on turnout and party support).

The data file `hki2019.csv` contains the results of the 2019 parliamentary elections in Helsinki from different voting areas (`Alue`). In addition, the voting areas of Helsinki are joined with the Helsinki major area (`Suuralue`) to which the voting area belongs to.

The data file `ek2023.csv` contains the results of the ~~2019~~ ²⁰²³ parliamentary elections in Finland divided by municipality (`Alue`). In addition, a constituency to which the voting area is belongs to is also included.

Task 1 (3 p.)

- a) Load into R the 2015 metropolitan area parliamentary election results in different voting areas into a data frame `pkseutu`. Use scatterplots to examine the turnout and the dependencies between party support while the voting area acts as the observation unit. Find such a party whose support seems to be growing as the turnout increases and draw a scatterplot corresponding to this dependence.

Which of the following interpretations (one or many) do you think can be made from the picture you have drawn? Justify your answer.

- i) Support for this party is apparently high in those areas where the turnout is high.
 - ii) If the turnout in the election is high, then this party will obviously get a lot support.
 - iii) If the support of this party in the metropolitan area were to rise, obviously the turnout would also increase.
- b) Find such a explanatory party, that its support in the voting areas of the metropolitan area explains well, in your opinion, the support of Finns Party in the voting areas of the Helsinki metropolitan area. Try to model the dependence with a linear regression model. What your model predicts to be the support for Finns Party in such a (fictional) area where the explanatory party's support is 15.5 percent?
- c) It is noted that the support of small parties (category "Muut") is quite explanatory for National Coalition Party support in the metropolitan area with one obvious exception. What is this area? Draw a picture showing the names of the voting areas instead of dots to illustrate this the deviation.

Fit a linear regression model in which you explain the support of National Coalition Party by the support of the small parties (i.e. the party "Muut"). However, exclude the previous exceptional area and make the estimates without this observation. Draw the corresponding scatterplot and add the estimated regression line to the plot.

Also add another party to the model as an explanator so that the coefficient of determination is as high as possible with these two explanators. Predict the support of the National Coalition Party in an area where the support of small parties ("Muut") is 3.5%, and the support of the explanator you choose is 10.2 percent. Give a brief evaluation of the suitability of this two explanators model.

- d) Fit a linear regression model in which you explain the support of the National Coalition Party with the support of all other parties. What do you notice? Explain what causes this result.

Task 2 (4 p.)

ek2023 **and 2023**

Let's move on to the results of the 2019 elections. Load the national election results for the whole country (~~ek2019~~.csv) into an R data frame `finland`. Also download the results by voting areas in Helsinki (`hki2019.csv`) into a data frame `hki`.

- a) Write a function `largestSupport`, which takes as an argument the election statistics as a data frame and the abbreviations of the abbreviation of the party name as a string similarly as in the material (eg "KESK" or "VIHR"). The function should look from the data frame provided as an argument the area in which the support of the party, given as an argument, is the largest

of in all the voting areas in the data. Finally, the function should return this result and tell how much support there is.

If the party given as the second argument cannot be found from data frame given as the first argument, the function must return text:

Please check the party name

For example, the command `largestSupport(hki, "PIR")` should return

"PIR support is highest in the Sörnäinen region and its support is 3.7 percent"

or the command `largestSupport(finland, "SDP")` should return

"SDP support is highest in the region of ~~Imatra~~ **Rautjärvi** and its support is ~~40.8~~ **38.1** percent"

while the command `largestSupport(hki, "XYZ")` should return

Please check the party name

Tip: The `subset` function can be useful.

- b) Make the function the `largestParty` that takes as its arguments election statistics as a data frame, and the region name as a string. The name must be exactly as in the data set. The function must choose from the given data frame the area corresponding to the other argument and look for the largest party in the area and its support in that area. Finally, the function should return as a result what is the largest party in the region and how much its support is.

If the area given as the second argument cannot be found from the data frame given as the first argument, the function must return a text:

The area not found

For example, the `largestParty(hki, "Katajanokka")` should return

"The largest party in the Katajanokka region is KOK and its support is 32.9 percent"

or the `largestParty(finland, "Närpiö")` should return

"The largest party in the Närpiö region is RKP and its support is ~~88.9~~ **82.6** percent"

while the command `largestParty(finland, "Exactum")` should return

The area not found

- c) Make a function `visualize` that takes the Helsinki election statistics as its argument as a data frame, and two strings giving abbreviations for the names of the two parties. The function should draw a scatterplot of the dependency of the support of these parties in Helsinki's voting areas. The image drawn by the function must alongside visualizing the actual voting area also in some way visualize the major area (South = Eteläinen, East = Itäinen, South-East = Kaakkoinen, Central = Keskinen, Northeast = Koillinen, West = Läntinen and North = Pohjoinen), into which voting area belongs to.

Interpret this picture and the major area differences when the chosen parties are the National Coalition Party and the Left alliance, i.e. call your function with a command

`visualize(hki, "VAS", "KOK")`

A function that produces a particularly informative and visually clear image can provide a half point bonus.

This function does not have to take into account exceptional situations, i.e. incorrect arguments.

- d) Write a function `compare`, which takes its arguments the election statistics as a data frame, the region name as a string and the reference region name as a string and a number `k`. The function should draw an image that visualizes the differences between the support of parties between the chosen two regions. The number `k` given as an argument determines how many largest parties in the first region are taken in The image produced by the function may be, for example, similar bar plot as in the example at the end of the assignment paper. A function that produces a particularly informative and visually clear image, can provide a half point bonus.

For example, the command `compare(finland, "Helsinki", "Oulu", 10)` could draw a similar picture is given at the end of the assignments.

This function does not have to take into account exceptional situations, i.e. incorrect arguments.

Task 3 (3 p.)

Mr. Fisher receives as a gift a thousand-sided die (singular form of *dice*, tuhatsivuinen noppa) from Mr. Teller. Mr. Fisher has a suspicion that the die is weighted so that when thrown it stops the side number one thousand upwards more often than any other numbers. He decides to investigate the matter using Bayesian inference.

- a) Mr. Fisher sets the prior distribution of θ to a continuous distribution whose density function is

$$f(\theta) = \begin{cases} 200\theta, & \text{when } \theta \in (0, \frac{1}{100}] \\ \frac{20000}{9900}(1 - \theta), & \text{when } \theta \in (\frac{1}{100}, 1] \\ 0, & \text{otherwise} \end{cases}$$

Draw a picture of Mr. Fisher's density function of the prior distribution. What does the priordistribution tell about Mr. Fisher's prior expectations about the θ ?

- b) Mr. Fisher rolls the die he has received 234 321 times. The die stops at a side number one thousand upwards 1234 times. Draw a picture of the logarithm of the density function of Mr. Fisher's posterior distribution (the density function does not have to be normalized). Find out what Mr. Fisher thinks is the best estimate for θ . Would the estimate have been different if the prior distribution had been a uniform distribution? Why / why not? Why should the posterior distribution be logarithmized?

Task 4 (2 p.)

- a) Make a function called a `lotterymachine`, which takes as an argument the number of lottery tickets you want and returns as a result how many tickets won. The lottery machine will draw a ticket independently and randomly, and the probability of winning with a single ticket should be set to 1/18643.

Using the function you wrote, simulate ten thousand lottery rounds with the number of tickets in each is 50. Save the results. Which is based on your results the ML estimate for a one-ticket winning probability? Calculate using the normal approximation found from the Statistical inference I -course material an approximate 95% confidence interval for the probability of winning with a single ticket Does it include the real probability of winning? What happens to

the confidence interval if do you increase the number of lottery rounds? What if you increase the number played tickets in each lottery round? Explain the reasons of your findings.

- b) Perform the simulations in part a) (i.e., 10,000 lottery rounds with 50 tickets) 100 times and save the ML estimate of a winning probability with single ticket for each simulation (thus resulting in one hundred ML estimates). What is the smallest interval containing 95% of the ML estimates? How do you interpret the above described matter? What happens between this if you increase the simulations the amount of? What if you increase the number of lines played in each round? How do you explain your findings?

Task 5 (12 p.)

The data file `income2017.csv` contains information about Finland the number of taxable income recipients, annual income and these of taxes paid by municipality in 2017. The variables of the data are:

- **Tulonsaajia:** The number of taxable income recipients in the area
- **Tulot:** The average taxable income of the region's income recipients in euros
- **Mediaanitulot:** The taxable median income of the income recipients in the region in euros
- **Ansiotulot:** The average earned income of the area's income recipients in euros
- **Pääomatulot:** The average investment (capital) income of the area's income recipients in euros
- **Valtionvero:** The average state tax paid by the income recipients of the region in euros
- **Kunnallisvero:** The average municipal tax paid by income recipients of the area in euros
- **Verot:** Total average taxes for income recipients in the area.
- **Tulot_miinus_verot:** The average income after the taxes of the income recipients in the area.

The number of municipalities differs slightly from the number of municipalities in the parliamentary election statistics ~~ek2019~~ ^{ek2023}.csv (e.g. Åland (Ahvenanmaa) municipalities are missing from the election statistics), but the data are can be connected by the name of the municipality. If the municipality doesn't have a pair with the another data set, you can leave the observations for that municipality as missing.

The task is to find out, based on these two data sets (`tulot2017.csv` and ~~ek2019~~ ^{ek2023}.csv), how the income level and the metropolitan area (e.g. the voting area) are related to voting behavior in Finland. You can limit yourself to considering the question from those perspectives you prefer after getting familiar with the data sets. A good starting point in writing your report to be submitted is that the reader is not familiar with the materials you are using nor with the assignment. Therefore, in addition to the making the analysis, be sure to describe from the data especially the things which you are studying in the research project.

Other instructions and tips:

- Remember that the observation units of the data are municipalities and e.g. the variables in the income statistics are condensed indicators (like averages and medians) of municipal residents.
- You can assume that the municipalities are independent of each other.

- If you wish, you can classify different variables into coarser scales than given.
- Not all possible comparisons between all possible groups worth reporting. Instead, you can focus on exploring the matter further from perspectives that seem appropriate, as long as you clearly justify your choice.
- Also remember to justify the statistical methods and the assumptions you use clearly.
- Essential issues are data processing (2p.), description of data and visualization (3p.), statistical testing and analyzes (4p.) and reporting the plain, written English (~~2p.~~).

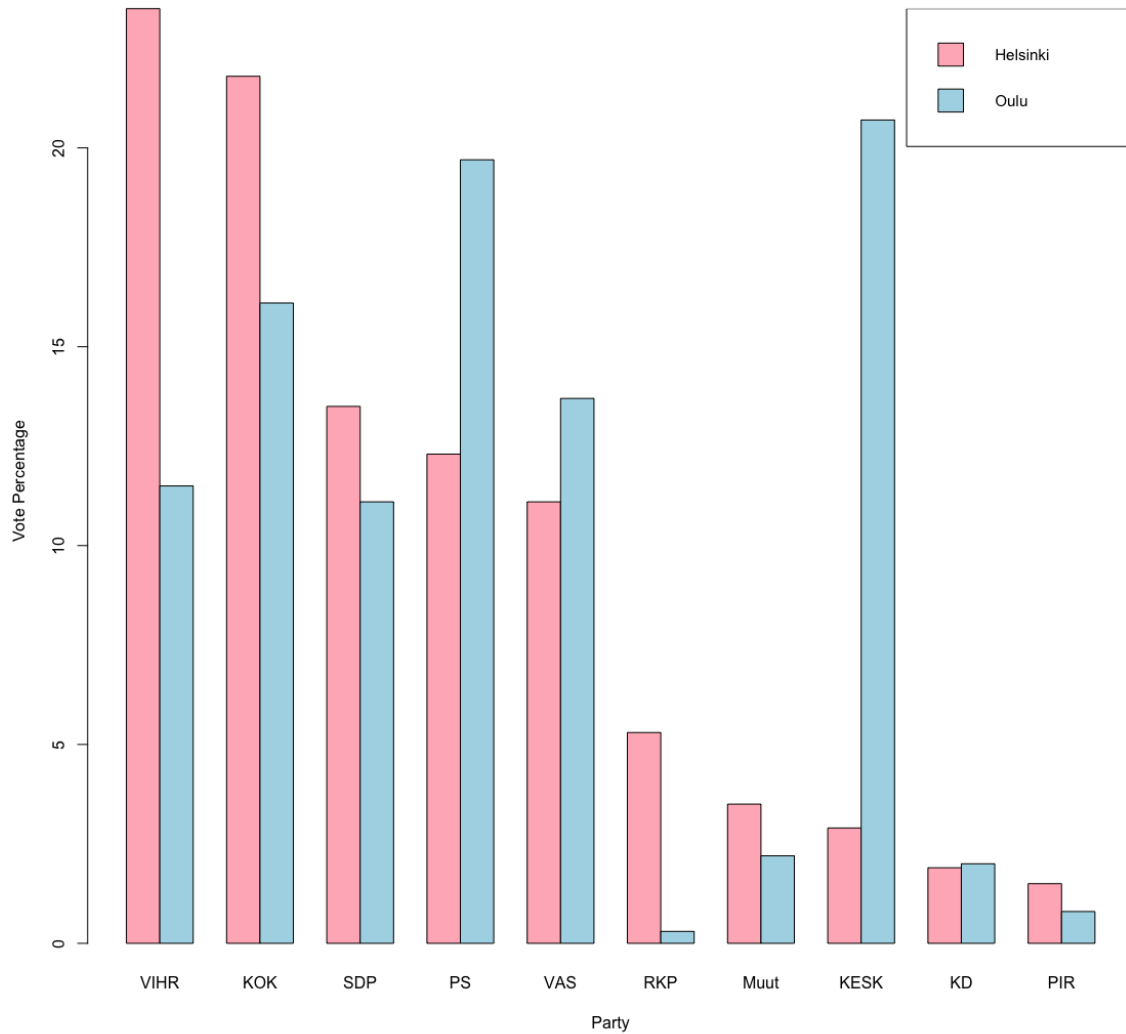
3p

However, even if you don't feel that you can meet the full requirements, prepare the report as carefully as you can!

Organization of the R program script (tasks 1-4)

```
1
2 ▾ #####
3  ## Task 1 ##
4 ▾ #####
5
6 ▾ #####
7  ## a) ##
8 ▾ #####
9
10 mysum <- function(a, b) a + b
11 mysum(1, 1)
12 # The answer is 2.
13
14 ▾ #####
15  ## b) ##
16 ▾ #####
17
18 mysum(2, 4)
19 # The answer is 6.
20
21 ▾ #####
22  ## Task 2 ##
23 ▾ #####
24
25 ▾ #####
26  ## a) ##
27 ▾ #####
28
29
30 ▾ #####
31  ## b) ##
32 ▾ #####
33
34
```

An example of the image of task 2d)



Sources of materials

- Statistics Finland (Tilastokeskus), Election Statistics, Data downloaded on 24 April 2019
- Statistics Finland (Tilastokeskus), Taxable income, The data was downloaded on 24 April 2019