Mathematical Statistics

Class 2 - Practical session - 6/9/2022

Bernardo Nipoti

Follow the instructions below, you can either work individually or in small groups. Points that are not covered during today's class should be completed as homework.

Part 0: material available on Blackboard

0. If you are not familiar with R and the environment RStudio, you can check the material that is available on Blackboard:

   - a guide for installing R
   - R tutorials and basic exercises (with solutions)

Part 1: fertility rate data

1. Download the file fertilityrate.csv from blackboard and create a data object in R. The dataset provides the fertility rate (defined as the average number of children per woman) for 229 countries from 1960 to 2014.

2. Within one R plot generate four boxplots with the fertility rates for 1984, 1994, and 2004, 2014. Make the plot 'nice', i.e. add some title and name the boxplots according to the year they are based on.

3. Choose one country and produce a plot showing the evolution of the fertility rate for that country during the period 1960 to 2014.

4. Compute summaries (mean, median, quartiles) for the years 1980 and 2000 (use the R command `summary`).

5. (Missing values) Make yourself familiar with the `is.na` command in R. In particular notice that R thinks about TRUE as 1 and FALSE as 0; for instance `is.na(NA)+is.na(1)` gives 1+0=1. Use this to compute the number of missing values (=NA's) for the years 2001 and 2007.

6. Learn what a function is, for example from the webpage

   http://www.statmethods.net/management/userfunctions.html

Write your own function which, given a vector as input, returns the number of missing values.

7. Compute the mean and the median for every year (not by hand) and plot them. For the mean you can use the function `colMeans`; for the median a similar function is provided by the R package miscTools.

Part 2: student survey data

8. Read Survey2017.csv into R, file available in blackboard. The dataset consists of the answers given, in 2017, to 16 questions by 111 students of a large university.

9. Compute the percentage of male and female participants. Can you say if the gender mix in the sample is representative for the student body of the university?

10. The survey question corresponding to the variable ComparisonOfFinancialSituation was "In comparison to other students how do you identify your financial situation?". Make yourself familiar with the function `table` in R. Use this to summarise the data in ComparisonOfFinancialSituation. Interpret the outcome.

11. Make yourself familiar with the `gsub` function in R. Use this R function to remove the Euro-symbols and the expressions "euros" and "euro" in the vector WeeklySpending.

    Hint: The Euro-symbol is: euro="\u20AC"

12. Produce and compare the histograms of the variable Happiness for students who have financial stability and students who do not. For this consider only rows whose entry for the variable Happiness is numeric.