Statistics and big data

Prof. Giuseppe Arbia; Prof. Sophie Dabo-Niang

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

The course aims to introduce the student to the world of Big Data. The course will be divided into two parts. In the first part (taught by Prof. Giuseppe Arbia) the student will use the R language and its packages, to apply the basic statistical methods learned in a first three-years degree course. The second part (taught by Prof. Sophie Dabo-Niang) will instead be devoted to introducing statistical models and methods and machine learning approach for the treatment of large and complex amounts of data. The emphasis will be on the practical aspects of implementing the various methods and models and on the interpretation of the results. Computer laboratories using the language R will be held weekly by Dr. Niccolò Salvini.

At the end of the course the student:

– will know how to analyze structured statistical datasets through the use of the R language and the RStudio development environment;

– will know how to distinguish between structured and unstructured datasets;

– will know how to distinguish between supervised models and non-supervised models.

***COURSE CONTENT***

Part 1:

Introduction to Big Data and Data Science. The importance of correct data collection in the world of Big Data. Hypotheses testing on an average and on a percentage. Hypothesis testing on 2 averages and 2 percentages. Hypothesis testing on more than 2 averages (ANOVA) and on more than 2 percentages (CHI square). Multiple linear regression model. Nonlinear regression. Regression with dummy variables.

Part 2:

Factor analysis, Clustering, Discrimination and Classification (Bionomial and multinomial logistic, kernel, general additive regression models, other supervised classification models).

***READING LIST***

For the logic of the statistical analysis of Big Data it is recommended the following:

* Arbia, G. (2021) Statistics, new empiricism and society in the world of Big Data, Springerbrief in statistics, Springer Verlag

The textbook of the course is the following:

* Zelterman, D. (2022) Applied Multivariate Statistics with R (2nd edition), Spinger-Verlag

***TEACHING METHOD***

Theoretical lectures and lab sessions on the software R and RStudio ©

***ASSESSMENT METHOD AND CRITERIA***

Option 1:

Optional intermediate exam on PC after week 6. In the computer lab, students will perform practical exercises using RStudio © software. If successful, the intermediate exam will account for 50% of the final grade. Final examination will be carried out with the same criteria as the intermediate test with a possible additional oral examination. Those who will successfully pass the intermediate exam, at the final exam will be tested only the second part of the course. The intermediate exam can only be used during the winter session at the end of the course, in the January and February appeals.

Option 2:

Full exam in any of the sessions planned during the year. In the computer lab, students will perform practical exercises using R and RStudio © software. Those passing the practical test in the computer lab could be called for an oral examination.

***NOTES AND PREREQUISITES***

Prerequisites:

* A basic course in statistics at a bachelor level including descriptive statistics, probability, inductive statistical inference (point and interval estimation).
* Basic notions of the language R. Students who do not possess the necessary prerequisites with the software R will be helped with *ad-hoc* sessions organized outside the official timetable in the first 2 weeks of the course.

***Recommended texts for prerequisites:***

For the statistical background:

Levine, J-Szabat-K. and Stephan, D. (2018) Business Statistics, Pearson.

For the language R:

Wickham, H., Grolemund G. (2018) R for Data Science, O’Reilly. Freely available on-line at <https://r4ds.had.co.nz/index.html>