# Statistics and big data

## Prof. Giuseppe Arbia; Prof. Daniel Zelterman

***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 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 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.

At the end of the course the student:

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

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

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

***COURSE CONTENT***

Introduction to the R package. Basic statistics (Descriptive statistics. Point and interval estimation. Test of statistical hypotheses 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. Binomial and multinomial logistic regression. Factor analysis. Discrimination and classification. Other supervised classification models: outline of regression trees approach. Time series. Canonical Correlations.

***READING LIST***

The textbook of the course id the following:

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

**For the logic of the statistical anlaysis of Big Data it is recommenmded the following:**

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

**Futher recommended reading:**

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

***TEACHING METHOD***

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

***ASSESSMENT METHOD AND CRITERIA***

Optional intermediate exam on PC after week 6. In the computer lab, students will perform practical exercises using R and RStudio © software. Under some exceptional circumstances, they may also carry out the test using their own PC on which the necessary programs will be installed. 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. Those who will successfully pass the intermediate exam, will have to perform only the second part of the final exam. The intermediate exam can only be used in the winter session at the end of the course, in the January and February appeals.

***NOTES AND PREREQUISITES***

Warnings: In the first lesson of the course the professor will indicate to the students how to download the R and RStudio software and the main R packages used throughout the course and how to get the codes for their installation on their own PC.

Prerequisites: a basic course in statistics at a bachelor level including descriptive statistics, probability, inductive statistical inference (point and interval estimation) and hypothesis testing.

Recommended text for prerequisites:

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

*In case the current Covid-19 health emergency does not allow frontal teaching, remote teaching will be carried out following procedures that will be promptly notified to students.*