**Applied statistics for business analytics**

Prof. Giovanni Viganò; Prof. Emilio Gregori

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

This course will give you the fundamentals of probability, and methods for statistical inference, with focus on estimation and tests of hypotheses. Students will learn how to perform regression analysis for datasets with quantitative and qualitative variables, especially in the context of applied works in business analytics. More advanced statistical techniques based on the partition of the covariate space, and unsupervised learning methods for clustering will be presented.

The course deals with concepts, methods and techniques of multivariate analysis of data for estimating and predicting purposes. The focus is on analyzing real data, using the software R.

The following learning abilities are provided and expected to be achieved by participants at the end of the course:

1. Knowledge of concepts, terms and methods of statistical inference, regression models, tree-based methods and clustering techniques (DD1- Knowledge and understanding);
2. Ability to correctly apply methods of estimation and prediction through the use of the statistical software R. (DD2- Applying knowledge and understanding);
3. Quantitative thinking addressed to make independent evaluations, driven by estimations and predictions (DD3- Making judgements);
4. Ability to read and interpret data and communicate results, through the extraction of qualitative information from quantitative data (DD4-Communication);
5. Mastery of tools useful for quantitative analyses required in future careers involving management of data, rigorous reasoning and data-driven decision-making (DD5- Lifelong learning skills).

***COURSE CONTENT***

*Module I* (Prof. Giovanni Viganò)

– Introduction: Business Analytics and data mining (the CRISP-DM methodology).

– Basics of probability theory: random variables, distribution theory.

– Point estimation for population mean, proportion and variance. Interval estimation for the population mean (in the normal case) and for the population proportion (binomial model). Extension to two sample problems.

– Hypothesis testing for population mean and proportion. Extension to two sample problems.

– Linear regression: inference and prediction, related t and F tests.

– Logistic regression: inference and prediction.

*Module II* (Prof. Emilio Gregori)

– – Tree-based methods: regression and classification trees.

– Clustering techniques: K-means clustering, hierarchical clustering.

***READING LIST***

The textbook of the course is:

A. Agresti, *Statistical Methods for the Social Sciences*, Pearson, 2018, 5th edition.

A selection of notes and other materials, available in the course reserve on Blackboard.

Suggested optional material for an introduction to R: W.J. Owen, *The R Guide*, CRAN, 2010.

***TEACHING METHOD***

The course involves lectures, exercise sessions using PC and discussion of use cases.

***ASSESSMENT METHOD AND CRITERIA***

Two alternative type of exams are foreseen:

*Option 1* - Two written partial exams, one mid-course exam and one final exam at the end of the course. Moreover, students can present a group-assignment about carrying out a research project through the use of R-Studio on a real dataset. The first partial exam will cover approximately 60% of the topics of Module I, the second partial exam will cover the rest of Module 1 and the topics of Module II. Both the first partial and the second partial exam will foresee close-ended and open-ended questions about theoretical concepts and about analyzing and interpreting results. The maximum grade for each of the two partial exams is 28. The assignment will foresee an autonomous activity of analysis on a real dataset through the software R, according to some basic instructions; the work is carried out in groups and must be submitted by mid-March. The maximum grade for the assignment is 3 points.

The aim of the exams is to assess reasoning analytic abilities on the course subjects; language properties and presentation communication abilities are also assessed. The final grade is given by the weighted average between the grade of the first partial exam (50%) and the grade of the second partial exam (50%) plus the grade of the assignment.

*Option 2* - One general written exam, covering both the topics of Module I and the topics of Module II. Moreover, students can present a group-assignment about carrying out a research project through the use of R-Studio on a real dataset. The exam will foresee close-ended and open-ended questions about theoretical concepts and about analyzing and interpreting results. The aim of the exam is to assess reasoning analytic abilities on the course subjects; language properties and presentation communication abilities are also assessed.

The maximum grade for the general exam is 28. The assignment will foresee an autonomous activity of analysis on a real dataset through the software R, according to some basic instructions; the work is carried out in groups and must be submitted by mid March. The maximum grade for the assignment is 3 points. As alternative to the assignment, students can attend an oral exam with maximum grade of 3 points, immediately after the general exam.

The final grade is given by the grade of the general exam plus the grade of the assignment, or of the oral exam.

 For the exam sessions other than the first, only Option 2 is applied; the grade of the assignment will be valid for the entire academic year.

Other details will be provided on Blackboard.

***NOTES AND PREREQUISITES***

Students enrolling in this course should ideally have a basic understanding of Statistics with regard to data analysis, probability and inference, at the level of the undergraduate courses taught at this University, namely Statistics or Statistica (analisi dei dati e probabilità).

Correspondence between students and teachers will be managed through Blackboard.

For office hours please contact:

* giovanni.vigano@unicatt.it,
* emiliofabio.gregori@unicatt.it.