



UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Analisi Esplorativa

2627-2-E4104B010-E4104B010-1

Learning objectives

The Exploratory Analysis module introduces the main descriptive statistical methods for studying two or more phenomena jointly observed on a set of statistical units. These methods aim to explore multidimensional data to detect their underlying structures and reduce their dimensionality while preserving the main observed features. From an application perspective, data analysis is carried out using R software within the RStudio environment.

Knowledge and understanding. This course will provide expertise and understanding concerning:

- Main basic exploratory methodologies of multivariate statistical analysis aimed at the classification of statistical units and the synthesis of observed variables into a reduced number of indicators
- Practical application of the exploratory techniques through numerical exercises solved with a pocket calculator (i.e., without the aid of statistical software)
- The underlying logic and functioning of the R language and its use in applying the main statistical analyses for multidimensional data, along with the related graphical representations
- Reading and interpreting the outputs of analyses generated with R.

Ability to apply knowledge and understanding. By the end of the course, students will be able to:

- Choose the most appropriate basic multivariate exploratory analysis methods according to the purposes of the analysis and the nature of the available data
- Reduce the dimensionality of a dataset by aggregating statistical units into groups and/or setting up summary indicators of the observed variables
- Interpret and compare the results of analyses obtained through different methods to determine which approach is most appropriate based on specific *a priori* criteria
- Import external data files of different sources and formats into R and autonomously use the basic syntax of the R language.

The course allows students to acquire solid theoretical and practical foundations regarding the main exploratory methods of multidimensional data analysis. These skills are necessary in any professional context involving the use

of data files and represent an essential foundation for the continuation of university studies.

Contents

Introduction to multivariate statistical analysis. Quantitative, qualitative, and mixed data matrices. Graphical representations of multidimensional data. Cluster Analysis: Hierarchical and non-hierarchical clustering methods. Principal component analysis. Linear discriminant analysis. Integrated use of exploratory multivariate methods. Applications to real data with software R (RStudio environment).

Detailed program

- Introduction to the multivariate statistical analysis: French and Anglo-Saxon schools, classification of multivariate analysis methods
- Quantitative, qualitative, and mixed-type data matrices. Main syntheses and transformations. Data representation, individual space, and variable space. Dissimilarities and distances between units, distances between variables
- Cluster analysis: Hierarchical and non-hierarchical clustering methods, goodness of classification, applications to quantitative and qualitative variables
- Principal component analysis: Extraction of the principal components, stopping criteria, evaluation of the reproduced variability, interpretation of the principal components, applications
- Linear discriminant analysis: set-up of linear discriminant functions in the presence of two or more populations, alternative method for constructing the discriminant functions, decision rules and evaluation of results, applications
- Integrated use of exploratory multivariate techniques
- Analyses of empirical cases with RStudio

Prerequisites

Successful completion of the following first-year prerequisite courses of Statistics I, Mathematical Analysis I, Linear Algebra, and Probability Theory

Teaching methods

Theoretical lectures in the classroom and practical exercises in the statistical-computing laboratory using R software within the RStudio environment.

All lectures will be delivered in person through traditional face-to-face instruction. However, asynchronous video recordings of both theoretical lectures and laboratory exercises will be made available.

Assessment methods

The exam consists of a written test (total duration: 2 hours) with three questions (divided into multiple points) regarding both theoretical and practical aspects of the methodologies covered during the course. A fourth question is optional and concerns programming with R software. The theoretical questions concern the methodological aspects of the topics covered in the course and allow for the verification of the theoretical knowledge acquired regarding the basics of multivariate statistical analysis (in particular, main matrices and their properties, main data types) and the methodologies of cluster analysis, principal component analysis, and linear discriminant analysis. The practical questions concern both numerical exercises (to be solved with a calculator) and the reading and commenting of parts of R output, and allow for the verification of the capacity for understanding and applying the theory, calculation, interpretation and commenting of results, and choosing among analyses obtained with different options of the methods. Furthermore, the written exam allows for the verification of the ability to express oneself through the appropriate use of technical statistical language.

The oral exam is optional (at the request of either the instructor or the student) and covers both theoretical and practical topics. Access to the oral exam is conditional upon passing the written test with a mark of at least 18/30 (both overall and in the theoretical part). Please note that the oral exam can result in an increase, maintenance, or decrease of the mark obtained in the written test.

Considering the abundance of teaching materials (in Italian) made available by the instructor on the course e-learning platform, no distinction is made between exams for attending students and exams for non-attending students. Finally, no mid-term exams are scheduled.

Students may request the English translation of the written exam text up to a maximum of 5 days before the date of the written exam.

Textbooks and Reading Materials

- Teaching material uploaded on the course e-learning website (restricted access)
- Frosini, B.V. (2014). Complementi di analisi statistica multivariata, EDUCatt, Milano
- Zani, S., Cerioli, A. (2007). Analisi dei dati e data mining per le decisioni aziendali, Giuffrè Editore, Milano
- Gherghi, M., Lauro, C. (2004). Appunti di analisi dei dati multidimensionali. Metodologia ed esempi, RCE Edizioni, Napoli
- Bolasco, S. (1999). Analisi multidimensionale dei dati: strategie e criteri di interpretazione, Carocci, Roma
- Dillon, W.R., Goldstein, M. (1984). Multivariate Analysis, J. Wiley, New York
- Everitt, B.S., Hothorn, T. (2011). An Introduction to Applied Multivariate Analysis with R, Springer, Berlin

Semester

First semester, second period

Teaching language

Italian

Sustainable Development Goals

QUALITY EDUCATION
