UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

# COURSE SYLLABUS

## Esploratory Analysis

**2122-2-E4102B084-E4102B084M**

### Learning objectives

The Exploratory Analysis module introduces the main descriptive statistical methods addressed to study two or more variables jointly observed on a set of statistical units. These methods aim at exploring multidimensional data to detect underlying structures and reduce their dimensionality, however preserving the main observed features. From a practical point of view, data analysis is carried out through the R software (RStudio environment).

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

- the principal exploratory methodologies of multivariate statistical analysis aimed at classifying statistical units in groups and synthesizing observed variables in a reduced number of indicators
- the practical application of the exploratory techniques through numerical exercises to be solved with the pocket calculator (i.e., without using statistical software)
- the logic and working of the R language and its use in the application of the main statistical analyses for multidimensional data and the related graphical representations

- the reading and interpretation of the analysis outputs produced by the R software.

*Ability to apply knowledge and understanding*. At the end of the course, the 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 the statistical units into groups and setting up summary indicators of the observed variables

- interpret and compare the results of the analyses obtained with different methods to establish which

approach should be regarded as the most appropriate one according to specific, a priori fixed 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 the student to acquire solid theoretical and applicative bases relative to the main exploratory analysis methods for multidimensional data, which are necessary for any working context where data files are used and for the advancement of the 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

Passing of preliminary examinations of Calculus, Linear Algebra, Probability, Statistics I

## Teaching methods

Theoretical lectures in the classroom and practical exercises in the statistical-informatics laboratory with the R software (RStudio environment).

## Assessment methods

The exam consists of a written test (total duration: 2 hours) with three questions (divided into several points) that deal with both the theoretical and applicative aspects of the methodologies covered in 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 aim at verifying the theoretical knowledge acquired regarding the basic notions of multivariate statistical analysis (in particular, the fundamental matrices and their properties, the main data typologies) and the methodologies of cluster analysis, principal component analysis, and linear discriminant analysis. The practical questions involve both numerical exercises (to be performed with the pocket calculator) and reading and commenting on parts of R output, and aim at verifying the ability of comprehension and application of the theory, calculation, interpretation, comment on the results, and choice among analyses obtained with different method options. Furthermore, the exam in written form allows verifying the ability of expression through adequate use of the statistical technical language.

The oral exam is optional (on request by the professor or student) and covers both theoretical and practical topics. Access to the oral test is subject to passing the written test with a mark of at least 18/30. It should be noted that the oral test may involve either the increase, the maintenance, or the decrease in the evaluation achieved in the written test.

Given the abundance of teaching material uploaded on the e-learning platform of the course, no distinction is made between exams for attending students and exams for non-attending students. Finally, there is no ongoing test.

## Textbooks and Reading Materials

- Teaching material uploaded on the course e-learning website (restricted access with password)
- 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