UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2627-2-F8205B011-F8205B011-1

Learning objectives

The course falls within the learning areas of statistical sciences, computer science and social sciences. It aims to provide students with knowledge of the following inferential approaches: nonparametric bootstrap, the multivariate Gaussian distribution, univariate and multivariate Gaussian mixture models, and predictive models.
During the learning activities, students develop a critical understanding of the assumptions underlying theoretical models through empirical applications using real and simulated data. Students also acquire skills related to the implementation of reproducible and replicable research. In addition, they develop written communication skills, since they are required to write texts accompanying the results of the analyses carried out.
Knowledge and understanding
The course enables students to:
• Analyze data using advanced statistical models developed for univariate and multivariate response variables, both categorical and continuous.
• Develop knowledge of simulation methods.
• Use the syntax and semantics of the R software, also through the RMarkdown environment, to develop a replicable and reproducible research approach. The documents produced include code, results, and comments on the code and on the analyses carried out.
• Interpret the results of data processing rigorously, developing expressive and text-summarizing skills also for dissemination purposes aimed at a non-academic audience. In this way, students develop independent judgment and refine their communication skills.
Applying knowledge and understanding
The course enables students to:
• Conduct statistical inference using resampling techniques (bootstrap).
• Estimate, select and interpret mixture models of distributions for heterogeneous populations.
• Conceptualize latent variable models, estimate their parameters using the maximum likelihood principle, and interpret the results.
• Apply theoretical knowledge to analyze different types of data arising from the application fields of the degree program, such as epidemiology, medicine, biology, genetics and public health.
• Implement code using the open-source R software language for descriptive and inferential analyses, adopting an open-source approach that ensures the reproducibility and replicability of the analyses.

The course enables students to acquire solid theoretical foundations and to develop practical applications through a problem-solving approach. The course is part of the field of data science, a knowledge area that is now essential for the professional contexts available to graduates in Biostatistics. At the end of the course, thanks to the material provided (the instructor's handouts accompanied by an extensive bibliography, R software code and the RMarkdown interface), students are able to continue studying this discipline independently.

Contents

In the first part of the course, the main probability distributions used to simulate realizations from random variables are reviewed. The resampling procedure known as bootstrap is introduced in order to obtain measures of precision in a nonparametric framework for selected estimators of interest. In the second part of the course, after presenting the multivariate Gaussian distribution, Gaussian mixture models are illustrated. The steps of the EM algorithm for maximum likelihood estimation of the parameters of these models and of latent variable models with a discrete distribution are described. The theoretical lectures are complemented by practical exercises. The course provides skills in using the syntax and semantics of the R software, also using the RMarkdown library through the knitr package to integrate code, analysis results and comments.

Detailed program

The first part of the learning activities concerns linear congruential methods for generating pseudo-random numbers and graphical tests for assessing pseudo-randomness. The theoretical discussion is accompanied by examples of data simulation from selected probability distributions. In the second part of the learning activities, after a brief introduction to the conceptual framework of statistical inference, the resampling procedure known as bootstrap is presented in order to obtain measures of precision in a nonparametric framework for selected estimators of interest. Confidence intervals obtained using both the percentile method and the BCa method, which corrects for bias, are illustrated. Univariate and multivariate finite mixture models for quantitative response variables are illustrated, assuming Gaussian distributions for the mixture components. The maximum likelihood estimation method based on the Expectation-Maximization algorithm is presented. In particular, density estimation and classification of statistical units using the maximum a posteriori probability method are considered.
Theory is complemented by practical exercises in which numerous applications are developed in the R environment with the support of the RMarkdown markup interface. These applications are aimed at the analysis and fitting of statistical models for real and simulated data concerning the fields of biostatistics. The main R packages used are skimr, MASS, boot, bootstrap, mclust. Students are encouraged to prepare reproducible documents in which they provide written comments on the code and critically discuss the results of the analyses, also through cooperative learning. Exercises are assigned weekly, and students are encouraged to write reports in which they comment on the code, provide an explanation of the analytical procedure carried out, and critically describe the results obtained. During the learning activities, the solutions to the assigned exercises are discussed.

Prerequisites

For an easier understanding of the course content, students must know the basic notions of Probability and Statistical Inference, as well as the basic syntax and semantics of the programming language in the R environment.

Teaching methods

Lectures are held in the computer laboratory or in computer-equipped classrooms. Theoretical lectures are complemented by practical exercises that allow students to learn through problem solving by analysing real and simulated data. Weekly review exercises are assigned on the topics covered in the course. During the course, with the support of R in the RStudio environment and the R Markdown interface, students learn to produce reproducible documents containing code, descriptions, and comments on the results of the analyses. Students are encouraged to collaborate with one another in solving applied problems, in order to promote cooperative learning. The course consists of 30 hours of lectures, devoted to the presentation of theoretical and methodological aspects, and 12 hours of interactive teaching, devoted to guided exercises, applications using R/RStudio/R Markdown, discussion of results, and problem-solving activities. The asynchronous video recordings made available on the e-learning platform constitute supplementary study support material and do not replace in-person teaching activities.

Assessment methods

The following methods for assessing learning apply both to students attending and to students not attending the lectures. The examination consists of a written test with an optional oral examination. No mid-term tests are scheduled. The written examination lasts a maximum of two hours and takes place in the computer lab. The open-ended theoretical questions aim to assess the understanding of the essential concepts of statistical inference conducted with advanced methods. The applied exercises, carried out using the R environment, RStudio and RMarkdown, allow the assessment of students' ability to apply the proposed methodologies and to produce reproducible reports describing the data, the procedures and the results obtained.
The written examination is graded on a scale out of 30, according to a scoring rubric communicated to students before the exam. The assessment takes into account: the theoretical correctness of the answers; the appropriateness of the model choice; the correctness of the computational implementation; the critical interpretation of the results; and the clarity and completeness of the written communication. The optional oral examination, if requested by the student or by the instructor, consists of a discussion of the written examination and of the theoretical topics covered in the course, and may confirm, supplement, or modify the final grade. During the examination, students are allowed to use the teaching materials and the materials provided by the instructor, including the R and SAS code developed during the course activities. The exam is considered passed with a minimum grade of 18/30.

Textbooks and Reading Materials

The main teaching material consists of handouts prepared by the instructor, covering theoretical topics, applications developed with the R software, exercises and solutions. These handouts will be made available on the university e-learning platform page dedicated to the course. In addition, at the end of each lesson the instructor publishes the slides, calculation programs and datasets used.
Exercises are assigned weekly, together with the corresponding solutions. Examples of examination texts are available on the same webpage.
The main reference texts are listed in the bibliography of the handouts. Some of them are the following, also available as eBooks in the University Library:

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). Latent Markov Models for longitudinal data, Chapman and Hall/CRC, Boca Raton.

Bishop, Y. M., Fienberg, S. E., Holland, P. W. (2007). Discrete multivariate analysis: theory and practice. Springer Science & Business Media, New York.

Blitzstein, J. K., Hwang, J. (2014). Introduction to probability, Chapman & Hall/CRC.

Gentle, J. E., Härdle, W., Mori Y. (2004). Handbook of computational statistics. Springer-Berlin.

Lange, K. (2010). Numerical analysis for statisticians, 2nd Edition, Springer, New York.

Pennoni, F. (2026). Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2026). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Semester

Semestre I, ciclo I, Settembre-Novembre 2025

Teaching language

The course is delivered in Italian. Erasmus students may use the teaching material prepared in English and provided by the instructor upon request. They may also request to take the examination in English.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | REDUCED INEQUALITIES | CLIMATE ACTION