UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici II

2021-2-F8203B011-F8203B013M

Learning objectives

The aim of the course is to provide analytic and inferential advanced statistical procedures also conducted by simulations. The content is presented to develop a critical understanding of the underlying assumptions. The main arguments are bootstrap and mixture models.

Knowledge and understanding

- to develop a critical knowledge of the assumptions;
- to develop simulation methods;
- to apply the models using the R semantic with the RMarkdown interface;
- to interpret and explain the results with rigor and to provide a clear description of the same.

Ability to apply knowledge and understanding

- to apply statistical inferential methods such as bootstrap;
- Estimate, select and interpret mixture models for heterogenous populations;

- Apply latent variable models;

- Apply models to data arising in the fields of epidemiology, medicine, biology, environmental, genetics and public health.

The student is encouraged to explain the theory and the results of the applications by providing written text and at the oral part.

The course provides the main concepts for parametric and non-parametric statistical models which are essential for the analysis of the data arising in the working contexts of biostatistics, statistics, demography and public health. It is compulsory for the next course of student' studies.

Contents

In the first part of the simulation methods are introduced to generate pseudo-realizations from random variables. The student is introduced to some resampling methods: bootstrap and jackknife along with their use for inferential purposes.

The Expectation-Maximization (EM) algorithm is illustrated for incomplete-data problems through the estimated parameters of a generalized linear model. Then it is illustrated as an optimization method for the estimation model parameters of finite mixture and latent variable models. The course provides skills in use of the semantic of the software R.

Detailed program

The first part of the course deals with simulation methods and linear congruential methods to generate pseudo-random numbers. Graphical tools for testing the series are illustrated along with some statistical tests such as Kolmogorov-Smirnov and Chi-Squared tests. Transformations of uniform deviates and simulation of random numbers from specific distributions are considered. Some theoretical features of the exponential, binomial and Gaussian distributions and convolution of random variables are exposed.

The main resampling methods such as the jackknife and the bootstrap are introduced. The bootstrap is applied for bias adjustment, and for the estimation of dispersion. Bootstrap confidence intervals based on the percentile method and the bias corrected accelerated bootstrap method are explained. Applications of the bootstrap are provided involving the skewness estimator, the relative risk estimator and some estimators derived from the linear regressions coefficients.

Among the optimization methods the Expectation-Maximization Algorithm is considered and explained first as a tool to impute missing values through a generalized linear model and then as a tool to maximize the log-likelihood function for incomplete data problems. Finite mixture models are introduced both for continuous and categorical data and a special focus is given on the mixture of Gaussian distributions and on latent variable models for categorical data.

Some amount of time is devoted to explain the theory by imparting flavor of the empirical applications on real data collected from different fields arising in epidemiology, pharmacoepidemiology medicine and biology as well as ecology and environmental sciences. They are developed within the statistical environment R, RStudio with the

RMarkdown interface so as to provide live code and make reproducible documents. The main R packages are boot, bootstrap MultiLCIRT and mclust.

Prerequisites

Knowledge on Probability and Statistical Inference is required as well as the basic knowledge of the R programming language.

Teaching methods

The lectures are held in the lab since the theoretical part is placed side by side with the applications carried out with the computer. During the lectures, many practical examples based on real and simulated data referred to different contexts are proposed to the students to be solved with R through the RMarkdown interface. The student is also encouraged to develop the cooperative learning in order to interact each other and finalize the required steps of the analysis. Exercises are carried out to report in a written form the results by adding critical comments and create reproducible documents.

During the Covid-19 emergency period the lessons will take place in the online asynchronous mode (videotaped lessons) with scheduled videoconferences meetings and some live meetings according to the scheduled days indicated at the elearning page of the course

Assessment methods

The following assessment methods are valid also for students not attending lectures. The written examination is performed in the lab where the student has to solve the exercises by showing that she/he is able to apply simulations and statistical models to real data in the field of biostatistics. The exercises are planned to evaluate the analytical skills of the students and his/her ability to solve the problem with R and the RMarkdown interface as well as to provide a reproducible document.

The following assessment methods are valid also for students not attending lectures. The written examination is performed in the lab where the student has to solve the exercises by showing that she/he is able to apply simulations and statistical models to real data in the field of biostatistics. The exercises are planned to evaluate the analytical skills of the students and his/her ability to solve the problem with R and the RMarkdown interface as well as to provide a reproducible document.

The results of the written examination are published in the e-learning page. With a positive score (from 18/30 and above) the student has to sustain an oral exam where she/he is explaining the theoretical features raised in the written part and the theory in the program of the course. In this way it is possible to evaluate the comprehension of the theoretical models. The written and the oral part are compulsory and should be carried out during the same examination term. Intermediate assessments are not planned.

During the Covid-19 emergency the exam will be same but it will carried out in the lab or in videoconference through Webex according to the guidelines of the University.

Textbooks and Reading Materials

The teaching material is made by the lecture notes concerning the theory and the applicative examples. Slides, R scripts, datasets and exercises with solutions are available after each lecture. The material is downloadable from the web page of the e-learning platform of the university.

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. and Hwang J. (2014). *Introduction to probability,* Chapman & Hall/CRC.

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

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

Pennoni, F. (2020). *Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R*. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Rizzo M. L. (2008). *Statistical Computing with R*, Chapman & Hall/CRC, New York.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Semester

Semester I, cycle I, October-November 2019.

Teaching language

The teaching language is Italian. Erasmus students can meet the professor to define proper English textbooks and require to carry out the exam in English.