HUMANITAS UNIVERSITY

MEDTEC SHOOL

Course: Statistics

Year: 2nd

Period: 1st semester

Credits: 6

Objectives

The course aims to introduce students to statistics and its application to problems arising from medicine and clinical practice.

The main objectives of the course are:

- *To provide techniques and tools for the synthetic and graphical analysis of the information provided by clinical data sets;*
- *To introduce to the language and the models for the representation and the analysis of random phenomena, with particular focus on clinical problems;*
- *To introduce the methods and tools of descriptive and inferential statistics, as well as statistical modelling;*
- *To apply the methods and techniques of statistics to real data sets by means of the use of appropriate statistical software (e.g., R and RStudio).*

Moreover, the course aims at providing students with the ability of translating clinical research questions to a statistical setting and design, to define suitable methods to answer such questions, and, last but not least, to properly communicate statistical evidence to a clinical audience.

Prerequisites

Basic notions of calculus.
Some practice with basics of statistics and programming may help, despite they are not mandatory.

Contents

Block1: How to describe medical data. [8h]

- Samples and populations. Location and dispersion indexes (mean, median, quantiles, variance, standard deviation, range and IQR). Graphics for continuous, discrete and categorical variables (histograms, boxplot, pie charts, qqplot). How to describe dependency: variance-covariance matrix, correlation.

- The concept of Probability and Conditional Probability. Bayes formula and total probability theorem.

- Density Function (PDF) and Cumulative Density Function (CDF) for discrete variables (Bernoulli, Binomial and Poisson) and continuous variables (Uniform, Gaussian, Exponential). Focus on Gaussian computations.

- Point estimation: mean and variance estimators, bias and mean square error of an estimator.

Block2: How to test for evidence. [6+6 hours]

- Hypothesis testing procedure (Z and T test for the mean of Gaussian samples), significance level, power function, p-value. Confidence intervals (Z and T for one population).

- Nonparametric tests: Mann Whitney U Test, Wilcoxon matched pairs test, Spearman's and Kendall's rank correlation coefficients.

- The analysis of cross tabulation: Chi-squared test for association, tests for 2 by 2 tables (chi-squared test, Fisher's exact test), Yates' continuity correction. Odds and odds ratios.

Block3: How to make prediction. [12+8 hours]

- Linear models: least squares estimation, inference on parameters, inference on the mean, prediction of a new observation, model fit (index R2), analysis of residuals, multiple linear regression.

- Generalized Linear Models (Logistic Regression): the simplest model for dichotomous data, misclassification table, sensitivity, specificity, ROC curve.

- Models for time to event data: the Kaplan Meier product-limit estimate of survival, the log-rank test, survival function, hazard function, cumulative hazard function, censoring, Cox Proportional Hazards model.

Teaching Methods

Lessons + practical sessions (both traditional exercise session + lab sessions with R software). Practical sessions are finalized both to fix theoretical concepts exposed during the lessons and to allow students to learn how to conduct data analytics in a real world setting.
Group projects will be carried out during the course, in order to be discussed at the end as a part of the evaluation (see evaluation section below).
Students are encouraged to actively participate to the lectures with questions and comments.

Verification of learning

The course assessment will consist of two parts, namely an individual written exam and a team project. Both parts are mandatory.
The written exam will be taken in one of the dates scheduled by the School within the academic year; it will consist of some multiple choice questions (Part A) and exercises (Part B) possibly with the help of the software R, to be solved autonomously in maximum 2:00 hours. At the end of the exam the student will decide whether or not to have their exam evaluated.
The written exam will be evaluated with a score expressed in a scale from 0 to 30, the maximum evaluation being 32/30. The written exam will be passed upon obtaining a score greater than or equal to 18/30. The exam evaluation will account for the degree of clarity of the exposition and for the correctness of computations.
During the examination, being an open book mode, the students will be allowed to use the calculator, the statistical tables and a formulary.

The team project will consist of an analysis of a real dataset, to be conducted in teams of 3 to 5 students, using the models and methods introduced during the course.
The team projects will be presented at the end of the course in a seminar during an open workshop that will take place after the end of the semester.
Each team will receive an evaluation in a scale from 0 to 30.

The final evaluation of the course will be obtained as a weighted average of the scores obtained by the student in the written exam and project discussion, with weights 0.8 (written exam) and 0.2 (team project).
During the exam, the students will have to:

- *Demonstrate the degree of knowledge and comprehension of the key aspects of the course, presenting the used methodologies in a clear and exhaustive way;*
- *Demonstrate their ability to apply the learned notions to solve exercises and real problems, on any of the topics covered in the course.*

Texts

- *Bland, M.: An Introduction to Medical Statistics - 4th Edition, Oxford University Press, 2015.*
- *Hosmer, D., Lemeshow, S.: Applied Survival Analysis. John Wiley & Sons.*
- *Agresti, A.: Categorical Data Analysis, Wiley Interscience.*
- *Ieva, F., Masci, C., Paganoni, A.M.: Laboratorio di Statistica con R,*

Further readings and references will be suggested step by step by the teacher.