Course Syllabus- **Modern Statistical Computing in R**

Language of Instruction: English
Professor: Albert Satorra & Ferran Carrascosa
Professor's Contact and Office Hours: Albert Satorra (albert.satorra@upf.edu),
Course Contact Hours: 45 hours
Recommended Credit: 6 ECTS credits
Weeks: 4
Course Prerequisites: The equivalent of a regular first course in statistics
Language Requirements: None

Course Description:

Over the recent years, R (https://www.r-project.org) has become the leading tool for statistical computing and graphics. The basic language of R is greatly enhanced by numerous contributed packages submitted by users. The majority of computing in the leading applied statistical journals is done in R, and it is used almost exclusively in some of the leading-edge applications, such as in genetics and data sciences. This software permits data analysts to interact with their data and to design personalized protocols for statistical analysis. R is free software that can run in most of the computer platform systems (Windows, OS of Mac, Unix, etc.). The purpose of this course is to set a foundation for full exploitation and creative use of this statistical modern language for computing and graphics.

Much of the statistical methodology implemented in software packages are used in the form of a black box. The computer language R facilitates opening this black box, permitting users better control of the method and adapt it to their particular research.

The course will introduce students to the syntax and inner workings of R, to become proficient in everyday computational tasks with datasets of all kinds. It will be emphasized the practice of applications of elementary and advanced statistical methods, with an emphasis on (initial) data exploration and simple graphics.

Learning Objectives: At the end of the course, students will have learned

- The use of programming tool for statistical analysis in small or large databases.

- Data wrangling, transformations, sub-setting, exploratory data analysis, probability distributions and simulations, linear and non-linear regression methods, effective graphics.

- Computational concepts of modern statistics (bootstrap, simulations)

- Using a reporting system for reproducible statistical writing (the use of R Markdown)

Course Workload:

The course consists on online sessions of 3 hours, four days a week, for four weeks.

Timetable of the online classes:

Monday to Thursday from 12:00 to 15:00

First Day of classes: 5 July. Last Day of classes: 29 July.

The Final Exam will take place during the class of 29 July.

Methods of Instruction:

Each online session (3h) include both the lectures and the practical classes. The third hour will be devoted to practical exercises and preparation/discussion of the final project. The second instructor of the course will teach this third hour.

From first day of class, students require to have R installed in their laptops.

Method of Assessment

Assessment is composed of the following inputs:

1. Continual Evaluation: contribution to class +homework's (25%)
2. Main Project (45%)
3. Final Exam (30%)

(A minimum of 10 points, out of 30, is required in the final exam to pass the course)

The main project will involve some computing in R and submission of a report of up to 6 typed pages (not counting appendices). Students will select their projects from topics of their own interest (upon the acceptance of the instructors) and will make a brief oral **presentation** at the end of the course.

Non-UPF students (6 ECTS) will need to do an extended project as a part of the course.

Absence Policy

Attending class is mandatory and will be monitored daily by professors. The impact of absences on the final grade is as follows:

Absences	**Penalization**
Up to two (2) absences	No penalization.
Three (3) absences	1 point subtracted from final grade (on a 10-point scale)
Four (4) absences	2 points subtracted from final grade (on a 10-point scale)
Five (5) absences or more	The student receives an INCOMPLETE for the course

The BISS attendance policy does not distinguish between justified or unjustified absences. The student is deemed responsible to manage his/her absences.

Emergency situations (hospitalization, family emergency, etc.) will be analyzed on a case-by-case basis by the Academic Director of the UPF Summer School.

Course Contents:

IMPORTANT: There will be no class on Thursday July 22th

1. General introduction to computing

Using R as a calculator
Numbers, words and logicals; missing values (NA)
Vectors and their attributes (names, length, type)
System- and user-defined objects
Accessing data (data ()). Data in the system and date outside the system (read.table, scan)

2. First steps in graphics

The basics of R syntax
The R workspace
Matrices and lists
Subsetting
System-defined functions; the help system Errors and warnings; coherence of the workspace

3. Data input and output; interface with other software packages

Writing your own code; R script
Good programming practice
R syntax -- further steps
The parentheses and brackets; =, == and <–

4. Exploratory data analysis

Range, summary, mean, variance, median, sd, histogram, box plot, scatterplot

5. Probability distributions. Simulations

Random number generation Distributions, the practice of simulation,

6. Apply-type functions Compiling and applying functions Documentation

Conditional statements Loops and iterations

7. Statistical functions in R

Statistical inference, contingency tables, chi-square goodness of fit, regression, generalized linear models, M-estimation, non-linear regression modeling, the bootstrap method for assessing confidence

8. Graphics; beyond the basics

Graphics and tables using ggplot and tidyverse
Working with larger datasets, tidyverse
Principles of exploratory data analysis

9. Dataframes in R

Defining your own classes and operations Models and methods in R Customising the user's environment

Required Readings: Handout material will be posted on the web as the course evolves.

Recommended bibliography:

Students are encouraged to consult the following sources on their own.

Dalgaard, P. (2002), Introductory Statistics with R, Springer

Dennis, B. (2013). The R Student Companion, Taylor & Francis Group

Matloff, N. (2011). The Art of R Programming: A Tour of Statistical Software Design, William

Philip H. Pollock (2014). An R Companion to Political Analysis, CQ Press

Chihara, L. and Hesterberg, T. (2011), Mathematical statistics with resampling and R, Wiley

Lander, J. P. (2014) R for Everyone: Advanced Analytics and Graphics, Addison-Wesley Data & Analytics Series

Last revised, February 2021