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  
Recommended Credit: 5 ECTS credits  
Course Prerequisites: Basic course in statistics  
Language Requirements: None

Course Description:

Over the recent years, R ([https://www.r-project.org](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 is used in the form of a black box. This is advantageous for a user who is not interested in the details of the methods, but the result is often a second-rate application, because the implementation, even if of high quality, is often meant for a different context, small details in the setting of options are ignored or misunderstood, and the orientation in the output, formatted for general interest, is difficult.

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, skilled in applications of elementary statistical methods, with an emphasis on (initial) data exploration and simple graphics. Focus will also be placed on opportunities to enhance the learning experience in other statistical courses by illustrating and applying basic statistical concepts using R.

Learning Objectives:

At the end of the course, students will have learned

- to use a fundamental tool for computing in the practice of quantitative analytical methods (the 'paper-and-pencil' tool of the 21st century), that can work for the small jobs (like a pocket calculator) as well as for the big jobs (complex statistical data analysis).

- programming, data handling, transformations, subsetting, exploratory data analysis, probability distributions and simulations, regression and linear models, summarising data, how to handle large data sets, effective graphics.
modern concepts of statistics based on simulations and writing a report of a quantitative analysis.

Course Workload

The course is divided into lectures, discussions, and field studies.

Methods of Instruction:

Classroom sessions are of 2h. 15 m and include both the lectures and the practical classes and seminars. From first day of class, students are required to come with their own laptops.

Method of Assessment

Class participation (15%) homework and mini project (20%) (the equivalent of the Midterm exam) the main function of which will be to prepare students for the main project (65%). This 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 by their own (upon approval of the instructor) and will make a brief oral presentation at the end of the course (the equivalent of the Final Exam).

Contribution to Class + Midterm Evaluation: 25%
Final Project: 50%
Final Exam: 25%

Absence Policy

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

<table>
<thead>
<tr>
<th>Absences</th>
<th>Penalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to two (2) absences</td>
<td>No penalization.</td>
</tr>
<tr>
<td>Three (3) absences</td>
<td>1 point subtracted from final grade (on a 10 point scale)</td>
</tr>
<tr>
<td>Four (4) absences</td>
<td>2 points subtracted from final grade (on a 10 point scale)</td>
</tr>
<tr>
<td>Five (5) absences or more</td>
<td>The student receives an INCOMPLETE for the course</td>
</tr>
</tbody>
</table>

The UPF Summer School 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.

Classroom Norms:

· No food or drink is permitted.
· There will be a ten-minute break during the class.
Course Contents:

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 <-

3. Exploratory data analysis
   Range, summary, mean, variance, median, sd, histogram, box plot, scatterplot

4. Probability distributions. Simulations
   Random number generation Distributions, the practice of simulation

5. Apply-type functions Compiling and applying functions Documentation
   Conditional statements
   Loops and iterations

6. Statistical functions in R
   Statistical inference, contingency tables, chi-square goodness of fit, regression, generalized linear models, advanced modeling methods

7. Graphics; beyond the basics
   Graphics and tables
   Working with larger datasets
   Principles of exploratory data analysis (big data analysis)

8. 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.


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


Data & Analytics Series

Last revised, March 2017