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: 5
Course Prerequisites: Basic course in statistics
Language Requirements: None

Time modules: Monday to Thursday, 3pm to 5.15pm

Course Description:
Over the recent years, R has become the leading software tool for statistical computing and graphics. The software 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 R is used almost exclusively in some of the leading-edge applications, such as in genetics and data mining. The purpose of this course is to set a foundation for full exploitation and creative use of the statistical language for computing and graphics R. The course introduces 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 emphasis on (initial) data exploration and simple graphics. Focus is also placed on opportunities to enhance the learning experience in other statistical courses by illustrating and applying basic statistical concepts in R.

Keywords: R, statistics, methodology, statistical computing, statistical methods.

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
Assessment is composed of the following inputs:
1. Continual Evaluation: contribution to class +homeworks (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.

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 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.

Classroom Norms:
· No food or drink is permitted.
· There will be a ten-minute break during the class.
· Students must come to class fully prepared.

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, the bootstrap method to compute s.e.

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 February 2018