Data analysis with R

Helsinki University

Jouni Junnila

Course information

✓ Course details

  ✓ Lectures and demonstrations (3 per student)

  ✓ Lecture times:

    ✓ Mondays 16.15-17.45 & Thursdays 08.15-09.45

  ✓ Demonstrations:

    ✓ Tuesdays 10-12 & 12-14 / every other week per student

  ✓ 5 credits

  ✓ **Course book**: John Maindonald and John Braun: Data Analysis and Graphics Using R - An Example-Based Approach

  ✓ **Course lecturer**: MSocSc. Jouni Junnila

  ✓ **Demonstrations:** Ali Amiryousefi

Course goal

- This course provides an introduction to data-analysis based on the open source R environment and language which is a globally adopted tool for exploratory statistics and modelling.
- R is both a programming language developed for mathematical and statistical applications, as well as a extendable program for numerical computation.
- The large number of available extension libraries makes R an attractive choice for a wide range of application areas.
- During the course the participants will explore different kinds of datasets using both graphical and numerical approaches.
- After the course participants will handle the basic concepts to do diverse data analysis using R

Course contents

- Basic syntax
- Descriptive statistics
- Read in different kind of data to R
- Data manipulation
- Traditional tests & confidence intervals
- Graphical methods
- Simple statistical models
- Writing your own functions

Course lecturer

- Jouni Junnila
  - Graduated from University of Turku, Statistics department, majoring in biostatistics
  - Current job: Biostatistician / site manager in 4Pharma Ltd, Espoo branch office
  - Versatile experience in biostatistics, clinical studies, microarray analysis and data management using R and SAS programs, as well as teaching in University of Turku and Åbo Akademi.
  - Contact information: jouni.junnila@4pharma.com

What is R

- R is both an environment for mathematical computation as wells as a programming language with a rich syntax towards doing statistical modeling and data analysis.

- R is a open source software.

- It is designed for interactive use: the next step may depend on the previous result.

R system

- R is currently the environment of choice for
  - specialists who are implementing new methodology
  - highly trained professional data analysts as one choice (specially bioinformaticians)
  - increasingly, statistically skilled scientists.
- Specially popular in academia, a bit smaller part in private companies.
- Twice-yearly major releases bring improvements & new features.

R packages

- Packages are collections of R function and/or data.
- Because of the many developers, new things are being published all the time, that's why the base R can't include everything available.
  - There are many packages which you can download and install
    - By exploring them, you most certainly will find a function suitable for your needs.
    - OR you can of course write a function yourself.

Accessing R-packages

- First you have to install the package from CRAN, in R-homepage. (if not included in the recommended packages.

- To attach a package, type library(), eg library(DAAG).

- You can then get information about the package by writing library(help="DAAG")

R / S-PLUS

- Commercial program called S-PLUS uses in principal the same language as R.
  - There are naturally some minor differencies
  - S-PLUS is more used by companies, R by academia.
- With good knowledge of R, it is fairly easy to use S-PLUS as well.