Data Analysis Workshop

Description

The *Data Analysis* workshop enables laboratory-based life scientists to use the *R statistical programming environment* to analyse their own data.

R is an open-source cross-platform software tool that combines data manipulation, statistical modelling and visualisation. This workshop focuses on data manipulation and biostatistics modelling using relevant examples from the life sciences.

Using plenty of hands-on exercises, participants will learn about different data structures and functions in *R*, how to manage and ask specific questions of their data, and use the results of statistical tests. Programming capabilities that make *R* well-suited to data manipulation (e.g. vectorization) and common beginner pit-falls, will be introduced.

Advanced topics will also be explored, including the `plyr` and `reshape2` packages, automating and reporting common work-flows using R markdown, and an introduction to control structures (e.g. loops and conditional statements). Extra material is provided in the reference book for specific problems, e.g. regular expressions, details on the `apply` functions, an introduction to the the `dplyr` package and dealing with missing data, as well as a case study involving ANOVAs and other inferential statistics. Students are also given time to work on their own data, with the goal of developing data analysis solutions as part of the workshop.

For further details, visit our web-site at `www.science-craft.com`.

The workshop instructor, Dr. Rick Scavetta, can be contacted directly via email at *rick.scavetta@science-craft.com*

Basic visualisations will be covered, but will be treated in more depth in the separate *Data Visualisation* workshop.

Requirements

The workshop does not set out to teach biostatistics, although how to *execute* various descriptive and inferential statistics in an efficient manner will be covered. Participants should be comfortable with computing and be familiar with basic biostatistics to take full advantage of the workshop.

> Participants are asked to bring in their own data sets and computers for practical work.

An introduction to biostatistics is given a proper treatment in our *Statistical Literacy* workshop.

In the simplest case, data should be in a flat text file, e.g. `.txt` and `.csv` files or something that students would view in Excel. Proprietary data formats may require advanced data structures which we will not encounter in this workshop, or they may not even be wholly accessible in R. Students are strongly encouraged to contact Rick (see above) if they have any questions concerning the type of data they should bring, in particular if they want to work on proprietary data formats.

Software

Students should have the following cross-platform software pre-installed. An additional script for R packages will be provided before the workshop.

R – v3.0 or later

`http://www.r-project.org/`

RStudio – v0.99 or later

`http://rstudio.org/download`