Statistics and Data Analysis Using R

Fall, 2012

Statistics and Data Analysis Using R was developed to be a brief but systematic introduction to the R statistical software suite for biomedical scientists. We assume that you have at least a passing familiarity with the plots and statistical analyses that are most commonly used in biomedical papers, but no formal background in statistics or programming is necessary. The primary objective is learning to use R, but the course also emphasizes the standards of practice that programmers and data analysts have implemented to ensure transparency, accuracy and accountability. To get the most out of the course, you should bring a laptop with R already installed on it (see installation instructions below) and try to complete the coding tasks yourself, as the class proceeds. We provide small exercises in each lesson to provide practice and a course project to help pull it all together at the end.

Why R?

R (http://cran.r-project.org/) is free open-source statistical software available for PC, OSX and Linux platforms. R is a *command-line* program, meaning that in order to perform analyses, the user must type commands into a terminal rather than make selections from a menu. If all you will ever need to do is a t-test there are simpler tools available, but no menu driven programs can offer the analytic flexibility that cutting-edge science will demand. There are a number of other excellent statistical programs available, including SAS, Stata, and S-Plus, but we like R best for several reasons:

- R is *free*
- it is available for any operating system
- R is supported by a huge library of user-contributed packages
- Its graphical capabilities are unparalleled.

These features have made R the first choice of professional statisicians working in public health and medicine, you will probably never need another program, and this is a great opportunity to learn the basics.

Installing R

You can download the latest versions of R for Windows, Mac OS X, and Linux as well as manuals and detailed installation instructions from the U.S homepage for R. `http://cran.us.r-project.org` R-2.15.0 is the current version but it is updated frequently and version numbers will change every few months.

R includes an editor and facilities for workspace management, but you may want to install Rstudio as well, `http://rstudio.org/`. Rstudio is a free and open source integrated development environment (IDE) for R that makes it much easier to manage your work sessions.

Instructors

Luigi Marchionni, PhD (marchion@jhu.edu)
Leslie Cope, PhD (cope@jhu.edu)
Elana Fertig, PhD (ejfertig@jhmi.edu)

Time and Place

The course meets on wednesdays and fridays between April 18 through May 23 from 10am-1pm. There will be no class on Friday May 4th. On wednesdays, class will meet in JHSPH W2300 and on fridays will meet in JHSPH W2033.

Course Website

Course materials can be found online at http://genomics.jhu.edu/R-Course/.

Course Textbook

There is no required course text book, and in fact we would rather have you become comfortable with the extensive built in documentation than depend on external sources for the basics. However, references to external sources will be provided for the interested reader as the course progresses.

Course Content

Day 1- Oct 24 Getting started in R (Elana)

Day 2 - Oct 26 Making calculations and manipulating data (Elana)

Day 3 - Oct 31 Data input and output (Rob)

Day 4 - Nov 2 Summarizing and plotting Data (Leslie)

Day 5 - Nov 7 Statistics in R (Rob)

Day 6 - Nov 9 Introduction to programming, loops and control structures (Luigi)

Day 7 - Nov 14 Crash class on regular expressions and string manipulations. (Leslie)

Day 8 - Nov 16 Lab day, hands on with the class project (Leslie, Luigi)

Day 9 - Nov 28 Writing your own functions (Rob)

Day 10 - Nov 30 Transparency, accuracy and accountability (Luigi)