An Advanced Introduction to Data Analysis + R

Peter Li

9 November 2016

This workshop is an advanced introduction to manipulating, analyzing and visualizing data using R (www.r-project.org). It's advanced in the sense that it's designed for people who have *some* prior experience with either programming (e.g., C, Python). It's introductory in the sense that it's designed for people who may not have much (recent) experience analyzing data or using statistical software (e.g., SAS, Stata)

There are two goals. The first is to understand the fundamentals of R: its functional and object oriented nature, its grammar and syntax, its data types, and its flow control structures. The second is to understand the fundamental concepts and techniques of analysis and visualization: everything from cause and effect to histograms and scatterplots.

To best achieve these goals, there will also be lab sessions. In these sessions, we go beyond "textbook" examples, which illustrate how a command works and start with the punchline that reveal the insights that analysts work hard to uncover. Instead, mirroring what happens in the real world, we start from the beginning. We try to figure out what's happening in the data and then we try to assess whether what we find is "real". To do so, we'll use challenging and thought-provoking examples whose size and scale make analysis feasible in the workshop setting.

Why R?

1. R is a free, open-source programming language (GNU General Public License).
2. R is the *lingua franca* of statistics. Many new techniques and methods are written and developed in R.
3. Beyond statistics and data science, R is an important tool in many industries (e.g. bioinformatics, empirical finance).
4. Packages. User-written collections of software, data and documentation that add functionality to R. Be it a statistical technique, a machine learning algorithm, a file format (e.g., JSON, SAS), a visualization technique (e.g., igraph, ggplot2), an interface to a programming language (e.g., C++, Apache Spark, SQL, Stan), an interactive web application (e.g., shiny), or a dynamic document (e.g,. knitr), there's probably a package out there that can help you accomplish your task (cran.r-project.org/web/packages/).
5. A large community of users and experts. When it comes to getting help with a problem or question (large or small), there's probably someone who has already tackled it in task view (cran.r-project.org/web/views/), blog, or on stackoverflow (http://stackoverflow.com/questions/tagged/r).

Installing R

To install R, download your preferred binary distribution from cran.r-project.org. Linux (Debian, Red Hat, SUSE and Ubuntu), macOS or Windows binaries are available. If desired, you can also install from the source code. Please do so before coming to the first class.

Text Editors and IDEs

We'll primarily be using the R GUI application and its built-in text editor. But feel free to use whatever text editor and setup you like. For those who prefer an IDE, possibilities include RStudio (www.rstudio.com) and R Tools for Visual Studio (www.visualstudio.com/vs/rtvs/). For Emacs users, there is Emacs Speaks Statistics ESS (ess.r-project.org).

Schedule

Session I - Functions, Objects, Packages & Causality (2 hours)

getting started

R language

functional programming

object oriented programming

base R

packages

causality, correlation and conflation

Lab A: Detecting Discrimination (2 hours)

creating dynamic documents: knitr + rmarkdown + RStudio

Session II - Data & Flow (2 hours)

data types

data structures

flow control

R style guide

strategic and non-strategic behavior

Lab B: Making Decisions (2 hours)

Session III - Visualization & Significance (2 hours)

why visualize?

base R graphics

ggplot2

saving graphs

probability distributions, significance & confidence

Lab C: Predicting Elections (2 hours)

Why this workshop?

There are plenty of options, both online and off, to learn R. Three things make this class different. First, to ensure the quality of instruction, interaction and discussion, the number of participants will be small ($N \leq 5$). Second, the hands-on sessions, which are dedicated to analyzing data, are to my knowledge something unavailable from other classes. Third, while details have not yet been finalized, I hope to be able to use "pay what you want" pricing and to allow people to become sponsors (for this or future events) by contributing a caffè sospeso (http://www.nytimes.com/2014/12/25/world/europe/naples-suspended-coffee.html).