Computational Statistics: An Introduction to R and RStudio For Scientists

Randall Pruim

Calvin College

November, 2011
Statistics + Computation = Computational Statistics

Modern statistics is done with statistical computing tools. There are many possibilities (Minitab, Excel, Stata, SPSS, StatCrunch, ...).

Different tools come with different sets of advantages and disadvantages.
Statistics + Computation = Computational Statistics

Modern statistics is done with statistical computing tools. There are many possibilities (Minitab, Excel, Stata, SPSS, StatCrunch, ...).

Different tools come with different sets of advantages and disadvantages.
The Lady Tasting Tea
Using R

RStudio online:

1. Open up Firefox
2. Navigate to http://dahl.calvin.edu
   
   This link and additional notes are available at http://www.calvin.edu/~rpruim/talks/SC11/
3. Select "RStudio"
4. Login, using the Username and Password you were provided.
5. In the packages tab, select fastR (which will automatically select a few other packages).
Why Use R?

Modern statistics is done with statistical computing tools.
Why Use R?

Modern statistics is done with statistical computing tools.

- R is one of many reasonable choices.
- Excel is NOT. [?]
Why Use R?

Modern statistics is done with statistical computing tools.

- R is one of many reasonable choices.
- Excel is NOT. [?] 

Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation. Additional errors in specific Excel procedures are discussed. Microsoft’s continuing inability to correctly fix errors is discussed. No statistical procedure in Excel should be used until Microsoft documents that the procedure is correct; it is not safe to assume that Microsoft Excel’s statistical procedures give the correct answer. Persons who wish to conduct statistical analyses should use some other package.
Why Use R?

Modern statistics is done with statistical computing tools.

- R is one of many reasonable choices.
- Excel is NOT. [?] Worse (from a teaching perspective), using Excel for statistics damages your brain.
Why Use R?

Modern statistics is done with statistical computing tools.

- R is one of many reasonable choices.
- Excel is NOT. 
  Worse (from a teaching perspective), using Excel for statistics damages your brain.
  - Excel users will constrain themselves to what Excel makes easy, whether or not it is the best way to think about things
  - Excel users are not forced to constrain themselves in ways that train the mind for statistical reasoning.
  - Aside from data entry and rudimentary data manipulation, Excel has no place in statistics education.
  - If you find that Excel "does everything you need" (statistically) this is an indictment on your statistical needs, not a strength of Excel.
R Advantages

• R is free and available on any platform (Mac, PC, Linux, etc.) and also, via RStudio, in a web browser. This means students have access to R whenever and wherever they need it.

• R is powerful – you won’t outgrow it. If one goal is to prepare students for future research, R will grow with them.
R Advantages

- R produces excellent quality graphics. R produces publication quality graphics. Via the `lattice` package, a wide range of useful plots are easy to produce. For those willing to learn a bit more, plots can be customized to your heart's contents.

- R helps you think about data in ways that are useful for statistical analysis. As with most true statistical packages, R prefers data arranged with variables in columns and observational units in rows. Being able to identify each of these is a crucial pre-requisite to understanding statistical procedures.
R Advantages

• R promotes reproducible research.
  R commands provide an exact record of how an analysis was done. Commands can be edited, rerun, commented, shared, etc.

• R is up-to-date.
  Many new analysis methods appear first in R.
R Advantages

- There are many packages available that automate particular tasks. The CRAN (Comprehensive R Archive Network) repository contains more than 3000 packages that provide a wide range of additional capabilities (including many with biological applications.) The Bioconductor repository (http://www.bioconductor.org/) contains nearly 500 additional packages that focus on biological and bioinformatics applications.

- R can be combined with other tools. R can be used within programming languages (like Python) or in scripting environments to automate data analysis pipelines.
**R Advantages**

- R is popular – including among biologists. R is becoming increasingly popular among biologists, especially among those doing work in genetics and genomics. R is very widely used in graduate programs and academic research, and is gaining market share in industry as well.

- R provides a gentle introduction to general computation. Although R can be used “one command at a time” (that’s the way we will be using it here, for the most part), R is a full featured programming language.
R Advantages

- The best is yet to come. A number of projects are underway that promise to further improve on R, including
  - RStudio and its `manipulate` features
  - Chris Wild’s new graphical interface (only videos available at this point)
  - `mosaic` and `mosaicManip` packages from Project Mosaic make learning statistics with Reven easier.
Faculty Testimonial

I would like to say again how much I appreciated the USCOTS workshop. As it happens, I elected to teach a very small summer offering of our elementary statistics class with just two students, both of whom had failed the course before and needed it to graduate.

My approach has been to use R as a supplement to a traditional course, but I find that I have been able to spend a lot more time on simulations of all sorts, including re-sampling approaches to hypothesis-testing – including contingency tables – early on.

The sample is small, but these students find the R-approach quite engaging and intuitive. In addition to the ready access to simulation and manipulation activities, they seem to get a conceptual leg-up when it comes to the handling the formulas that dominate the traditional approach.
Email from a student, 6 months after Intro Stats class:

Dear Professor Pruim,

This is Alec Boyd, I was in your spring semester for Math 143 and I just wanted to say thank you for teaching me the ways of Rstudio! It has really helped me with these various research opportunities and class projects. It is quite funny when a classmate is trying to do histograms and ratios in excel..... it just doesn’t work very well. Rstudio trumps it all!
References