Introduction to R

David Smith, REvolution Computing
LUGOD, Nov 16 2009
Open-source statistical analysis and visualization
What is R?

• Open-source software for advanced statistical analysis and data visualization
• Interpreted language *designed* for statistical programming
  – Virtually all statistical and predictive analytic methods available
  – Create analysis in a fraction of the time compared to C++, Python, SAS, SPSS, ...

• [www.r-project.org](http://www.r-project.org)
R History

• Late 70’s / Early ’80s: S language invented at AT&T Bell Labs
• 1983-1992: S version 3
  – 1998: Chambers wins ACM Software Systems award for S
• 1993: R Gentleman and R Ihaka create R
  – 1995: Released under GPL v2
• 1997: R Core Group formed
• 2001: R 1.0.0 released
• Oct 2009: R 2.10.0 released
• 2M+ users (estimated)
  – Google, Facebook, Pfizer, Bank of America, New York Times
• 2000+ contributed packages
  – cran.r-project.org
• Annual international user group meeting
• Commercially supported by Revolution Computing
Momentum behind \textit{R}\textsuperscript{LUGOD}, Nov 16 2009

Data Analysts Captivated by R's Power

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

\textbf{Related}
- Bits: R You Ready for R?
- The R Project for Statistical Computing

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly...
What is Statistics?

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.

THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.

SOUNDS LIKE THE
CLASS HELPED.

\WELL, MAYBE.

http://xkcd.com/552/
REvolution Computing: The “R” Company

- REvolution R
  - Free, high-performance binary distribution of R
- REvolution R Enterprise
  - Subscription-based version, bundled with proprietary extensions
  - Fully supported, validated
- R Consulting and Training
- R Community Development
  - R Evangelism (blog.revolution-computing.com)
  - R community portal
  - Technical and financial contributions to R Project
What can you do with R?

• Mash-up messy data sources to analyze the foreclosure crisis

From O'Reilly’s *Data Mash-ups in R.*
What can you do with R?

• Find a clean place to surf in the Bay Area

John Oram, a scientist at the San Francisco Estuary Institute (SFEI) uses R to collect and monitor environmental data from the waters and wetlands of the Bay Area.
What can you do with R?

• Compare baseball player performance

Clay Buchholz (BOS)
PitchFX data exists for 572 pitches thrown in 2008.

Velocity distribution by pitch type

Location and velocity by pitch type

PitchFX Viewer, by Mike Driscoll: labs.dataspora.com/gameday/
R saves time for the New York Times

- Published 3 hours after Jackson’s death:
Getting Started with R on Ubuntu (Karmic)

$ sudo apt-get revolution-r
$ R

• R is interactive: type statements at the prompt, and the result is printed:

```r
> x <- rnorm(10, mean=1, sd=2)
> mean(x)
[1] 1.695564
```
• *Introduction to R* (free)
  – 20-minute tutorial: Appendix A

• *Introductory Statistics with R*
  – Peter Dalgaard

• *R in a Nutshell* (O’Reilly)

• Other resources for beginners:
  – blog.revolution-computing.com/beginner-tips/
ESS: An emacs-based GUI for R

• `sudo apt-get install ess`
• Start R in Emacs: `M-x R`

• `http://ess.r-project.org`
• ESS Reference Card:
  – `http://ess.r-project.org/refcard.pdf`
Let’s do a simulation!

- Is this your birthday?

January 6
Simulating birthdays

- A simple simulation:

```r
birthday <- function(n) {
  ntests <- 10000
  pop <- 1:365
  anydup <- function(i)
    any(duplicated(sample(pop, n, replace=TRUE)))
  sum(sapply(seq(ntests), anydup)) / ntests
}

x <- foreach (j=1:100) %dopar% birthday (j)
```
Birthday Simulation

```r
> x <- foreach (j=1:100) %dopar% birthday (j)
> plot(1:100, unlist(x), type="l")
```
More Examples

• Plotting unemployment data on a map
• Displaying financial data as a calendar heatmap
• Analyzing Twitter hashtags
Unemployment data map
Calendar heat map

Calendar Heat Map of MSFT Adjusted Close

2006

2007

2008

2009

Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec

LUGOD, Nov 16 2009
Social network analysis with Twitter
Thank You!

• David Smith
  – david@revolution-computing.com, @revodavid
• Revolutions, the R blog
  – blog.revolution-computing.com
• R Project
  – www.r-project.org

• Links for this talk:
  – tinyurl.com/lugod-nov09