

Statistical Programming and Reproducible Research with **R** in **RStudio** A Brief Introduction

William "Hank" Murrah

November 22, 2013



Introduction to R by a familiar face

<http://www.youtube.com/watch?v=F1yPbalcs30>

What is R?

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues...

What is R?

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues...

R provides a wide variety of statistical (linear and non linear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. (R-project.org)

Pros

- ▶ FREE! *R is FREE and open source*

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).
- ▶ Easily extensible with (currently) over 5,000 packages listed on CRAN.

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).
- ▶ Easily extensible with (currently) over 5,000 packages listed on CRAN.
- ▶ Scriptable.

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).
- ▶ Easily extensible with (currently) over 5,000 packages listed on CRAN.
- ▶ Scriptable.
- ▶ Publication quality graphics.

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).
- ▶ Easily extensible with (currently) over 5,000 packages listed on CRAN.
- ▶ Scriptable.
- ▶ Publication quality graphics.
- ▶ Turing complete language.

Pros

- ▶ FREE! *R is FREE and open source*
- ▶ Available for multiple platforms (i.e. Windows, Mac, Linux).
- ▶ Easily extensible with (currently) over 5,000 packages listed on CRAN.
- ▶ Scriptable.
- ▶ Publication quality graphics.
- ▶ Turing complete language.
- ▶ Quickly becoming the *de facto* standard among statistician.

Cons

- ▶ Has a steeper learning curve.

Cons

- ▶ Has a steeper learning curve.
- ▶ Multiple ways of doing the same thing.

Cons

- ▶ Has a steeper learning curve.
- ▶ Multiple ways of doing the same thing.
- ▶ Can have difficulty with *very* large datasets.

How is R different from other software?

- ▶ R is FREE (as in free beer and free speech)

How is R different from other software?

- ▶ R is FREE (as in free beer and free speech)
- ▶ R is an Object-Oriented not Procedural language

R's Roots... S

- ▶ S is a language that was developed by John Chambers and others at Bell Labs.
- ▶ S was initiated in 1976 as an internal statistical analysis environment - originally implemented as Fortran libraries.
- ▶ Early versions of the language did not contain functions for statistical modeling.
- ▶ In 1988 the system was rewritten in C and began to resemble the system that we have today (this was Version 3 of the language). The book Statistical Models in S by Chambers and Hastie (the blue book) documents the statistical analysis functionality.
- ▶ Version 4 of the S language was released in 1998 and is the version we use today. The book Programming with Data by John Chambers (the green book) documents this version of the language.

History of S

- ▶ In 1993 Bell Labs gave StatSci (now Insightful Corp.) an exclusive license to develop and sell the S language.
- ▶ In 2004 Insightful purchased the S language from Lucent for \$2 million and is the current owner.
- ▶ In 2006, Alcatel purchased Lucent Technologies and is now called Alcatel-Lucent.
- ▶ Insightful sells its implementation of the S language under the product name S-PLUS and has built a number of fancy features (GUIs, mostly) on top of it-hence the "PLUS".
- ▶ In 2008 Insightful is acquired by TIBCO for \$25 million; future of S-PLUS is uncertain.
- ▶ The S language itself has not changed dramatically since 1998.
- ▶ In 1998, S won the Association for Computing Machinery's Software System Award.

In "Stages in the Evolution of S", John Chambers writes:

"[W]e wanted users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming, when the language and system aspects would become more important."

<http://www.stat.bell-labs.com/S/history.html>

History of R

- ▶ 1991: Created in New Zealand by Ross Ihaka and Robert Gentleman. Their experience developing R is documented in a 1996 JCGS paper.
- ▶ 1993: First announcement of R to the public.
- ▶ 1995: Martin Machler convinces Ross and Robert to use the GNU General Public License to make R free software.
- ▶ 1996: A public mailing list is created (R-help and R-devel)
- ▶ 1997: The R Core Group is formed (containing some people associated with S-PLUS). The core group controls the source code for R.
- ▶ 2000: R version 1.0.0 is released.
- ▶ 2012: R version 3.0.2 is released on September 9, 2013.
- ▶ There are now over 5,000 packages listed on CRAN.