Statistics with Free and Open-Source Software

Wolfgang Viechtbauer
Maastricht University
http://www.wvbauer.com

Free and Open-Source Software

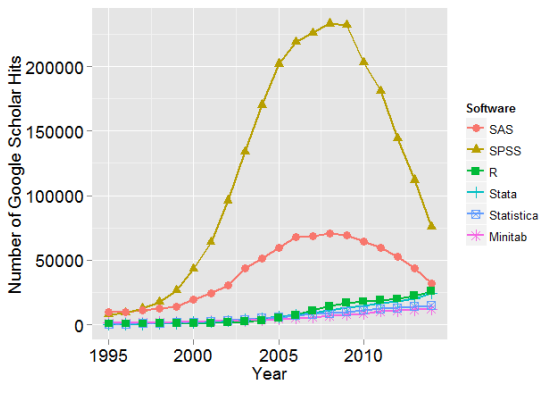
- the four essential freedoms according to the FSF:
  - to run the program as you wish, for any purpose
  - to study how the program works, and change it so it does your computing as you wish
  - to redistribute copies so you can help your neighbor
  - to distribute copies of your modified versions to others
- access to the source code is a precondition for this
- think of 'free' as in 'free speech', not as in 'free beer'
- maybe the better term is: 'libre'

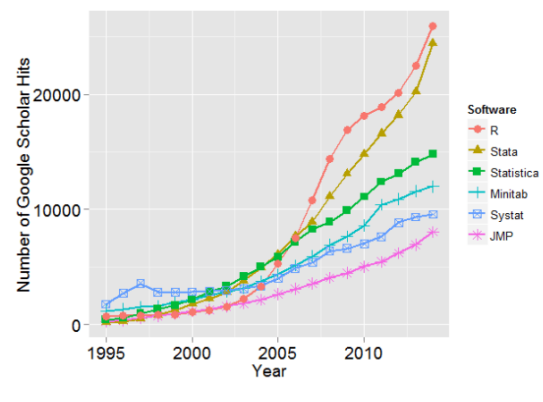
General Purpose Statistical Software

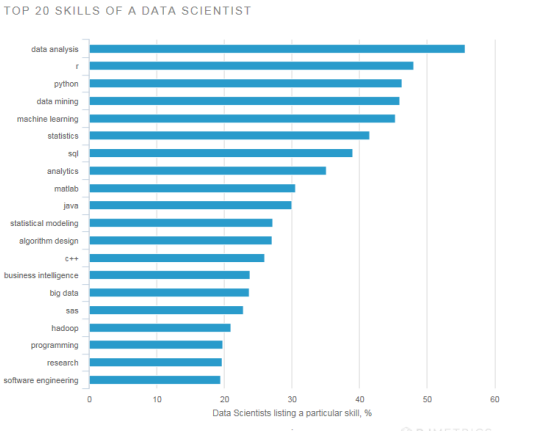
- proprietary (the big ones): SPSS, SAS/JMP, Stata, Statistica, Minitab, MATLAB, Excel, …
- FOSS (a selection): R, Python (NumPy/SciPy, statsmodels, pandas, …), PSPP, SOFA, Octave, LibreOffice Calc, Julia, …

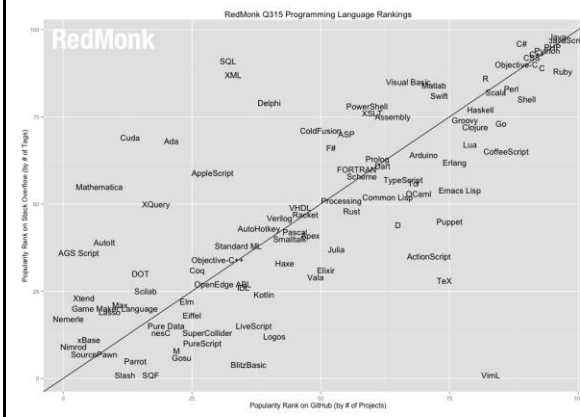
Popularity of Statistical Software

- difficult to define/measure (job ads, articles, books, blogs/posts, surveys, forum activity, …)
- maybe the most comprehensive comparison:
  http://r4stats.com/articles/popularity/
- for programming languages in general: TIOBE Index, PYPL, GitHut, Language Popularity Index, RedMonk Rankings, IEEE Spectrum, …
- note that users of certain software ~~may be~~ are heavily biased in their opinion









What is R?

- R is a system for data manipulation, statistical and numerical analysis, and graphical display
- simply put: a statistical programming language
- freely available under the GNU General Public License (GPL) → open-source
- cross-platform (can be used under Windows, Unix/Linux, Mac OS, …)

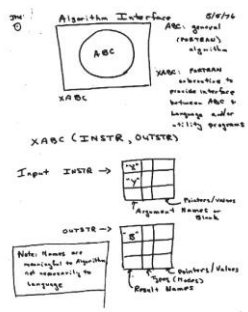
History of S and R

- … it began May 5, 1976 at:

Bell Laboratories, Murray Hill, New Jersey

History of S and R

- informal meetings to discuss development of a new system for statistical computing
- first implementation made by Rick Becker and John Chambers (and a few others)
- called "the system"



sketch of the system design made on the first meeting

History of S and R

- "the system" → "S" (the S language) (also play on name of programming languages such as C)
- first UNIX version of S in 1979 (version 2)
- distributed outside Bell Labs in 1980
- source code released in 1981, then licensed in 1984 for educational and commercial purposes

History of S and R

- history/development of S can be traced via a number of influential books:
  - Becker & Chambers (1984). S: An Interactive Environment for Data Analysis and Graphics.
  - Becker & Chambers (1985). Extending the S System.
  - Becker, Chambers, & Wilks (1988): The New S Language: A Programming Environment for Data Analysis and Graphics.
  - Chambers & Hastie (1991). Statistical Models in S.
  - Chambers (1998). Programming with Data: A Guide to the S Language.

13

History of S and R

- S-PLUS, a commercial implementation of S, released in 1988 by Statistical Sciences, Inc. (now TIBCO)
- Ross Ihaka and Robert Gentleman start developing a statistical programming language "not unlike S" at the University of Auckland (New Zealand) in the 1990s

14

History of S and R

- some R milestones:
  - first binary of R released in 1993
  - Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3), 299-314.
  - source code released in 1997 (CRAN is started)
  - R Core group is formed in 1997 with 9 members (now 20)
  - version 1.0.0 (2000), version 2.0.0 (2004)
  - first useR! conference in May 2004 in Vienna, Austria
  - version 3.0.0 released April 2013

15

History of S and R

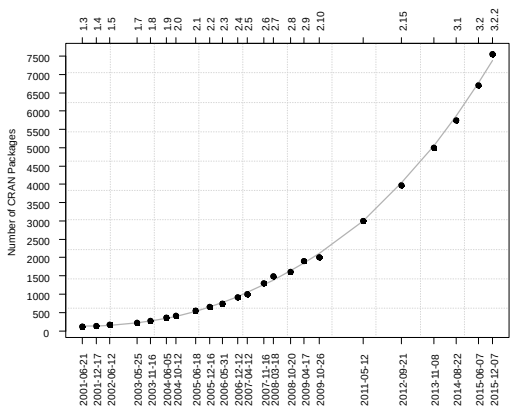
- new website and logo in 2015:

the old one

the new one

- R Consortium formed in June 2015
- current version: R 3.2.2 (released August 2015)
- (3.2.3 about to be released)

16



17

Other Related Developments

- Bioconductor started 2001
- Revolution Analytics founded in 2007, acquired by Microsoft in 2015
- RStudio founded in 2008
- New York Times article about R in January 2009
- "big data" (Google, Oracle, IBM, Intel, Microsoft, …)
- "data science" ('hacking' skills core component)
- open science, reproducible research

18

Why is it called R?

- **R**oss Ihaka and **R**obert Gentleman
- pun/play on the name of the S language
- like computer scientists, statisticians are geeks

- Data Scientist: The Sexiest Job of the 21st Century

19

Everybody Loves Videos!

20

Why Use R?

- IT'S F&CKING FREE (as in beer), YOU DUMMY!
- and it is free (as in free speech) & open source
- and its capabilities are at least as good as that of proprietary software (often better)
- most comprehensive coverage of methods
- huge/active/enthusiastic user community
- the 'lingua franca' of statistic(ian)s
- forces you to adopt a scripting approach
- cross-platform

21

22

Tooth Growth Data

- sample: 60 guinea pigs
- outcome: length of odontoblasts (teeth)
- treatment: Vitamin C supplementation (0.5, 1, or 2 mg/day delivered either via orange juice or a solution with ascorbic acid)

23

Edgar Anderson's Iris Data

- sepal and petal length and width for 50 samples for 3 different iris species

Iris Setosa

Iris Versicolor

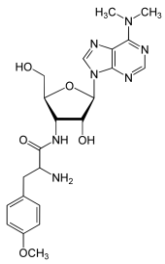
Iris Virginica

http://en.wikipedia.org/wiki/Iris_flower_data_set

24

Puromycin Data

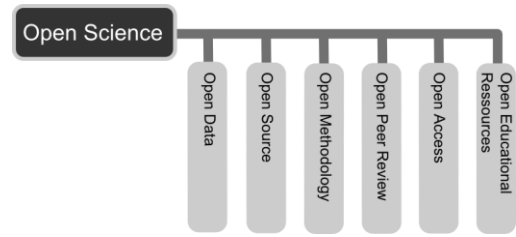
- Puromycin is an antibiotic that is a protein synthesis inhibitor
- study examined the velocity of an enzymatic reaction as a function of substrate concentration with and without the enzyme treated with Puromycin



https://en.wikipedia.org/wiki/Puromycin

25

Open Science / Reproducible Research



26

rmarkdown & knitr

- to create dynamic and fully reproducible documents/presentations/reports with R
- in essence: you write a single document that includes the text and analysis code that is then rendered into a desired output format
- more details:
  - http://rmarkdown.rstudio.com/
  - http://yihui.name/knitr/
  - Gandrud (2013): Reproducible Research with R & RStudio (website; source code)

27

My Recommendations

- learn R if you plan on staying in research
- knowing SPSS, Stata, and SAS is also useful
- learn Python is you want to be 'data scientist'
- keep your eyes on Julia
- embrace learning new tools and statistics

Some Interesting Developments

- Calling R from SPSS
- R Integration in JMP
- R Interface in SAS/IML Studio
- SAS University Edition
- Python Interfaces for R

29

Some R Resources

- R and CRAN (manuals, contributed documentation)
- use RStudio (unless you already use vim, Emacs, Notepad++, Sublime Text, Atom, WinEdt, …)
- many books / hard to give recommendations (search your favorite bookseller and look at reviews)
  - Field et al. (2012): Discovering Statistics Using R
  - Dalgaard (2008): Introductory Statistics with R
  - Muenchen (2011): R for SAS and SPSS Users
  - Springer Use R! Series, Chapman & Hall/CRC The R Series
- courses (lots: Coursera, DataCamp, Code School, Udemy, Udacity, statistics.com, my own, …)

30

Free Software and Beyond …

- free software is a movement
- open access and sharing of knowledge as a general philosophy
  - Free Software Foundation
  - Foundation for Open Access Statistics
  - Open Access Journals
  - Open Educational Resources
  - Wikipedia
  - OpenCola
  - …

31

U Haz Questionz?