Statistical data analysis

“The best thing about R is that it was written by statisticians. The worst thing about R …”

Bo Cowgill, Google

1. Tutorial - Introduction to R

What is R?

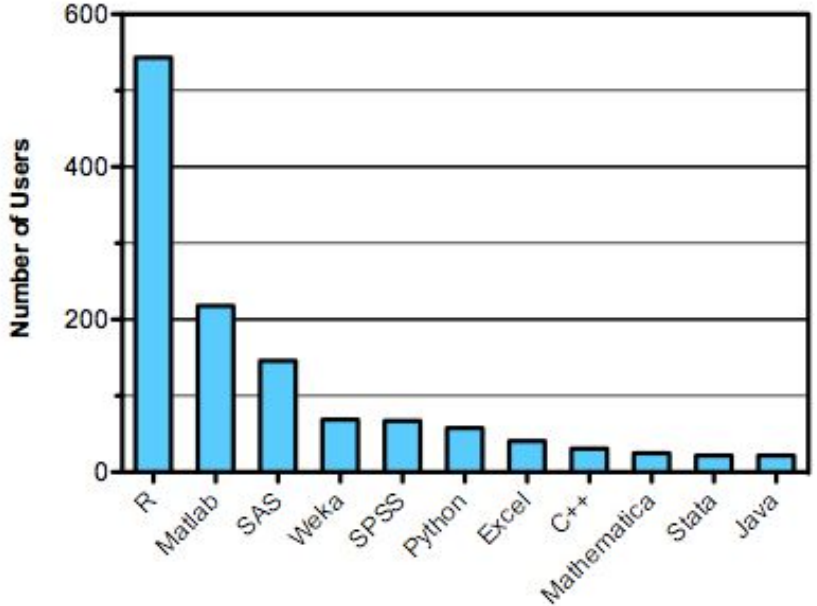
- Open source statistical language and software environment .
- Available freely under the GNU public license.
- De facto standard for statistical research.
- The core of R is an interpreted computer language.
- Developed for the Unix-like, Windows and Mac families of operating systems.
- R has a command line interface, but there are several graphical front-ends available (RStudio, RKWard, Rattle, Red-R, ...).

http://www.r-project.org

… and Why R?

- Widely used among statisticians and data miners for developing statistical software and data analysis.
- A large number of statistical procedures (linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing).
- Very active community and package contributions (CRAN).
- Very little programming language knowledge necessary.
- About **2 million users worldwide** in 2009 in the article in The New York Times (http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/?_php=true&_type=blogs&_r=0).

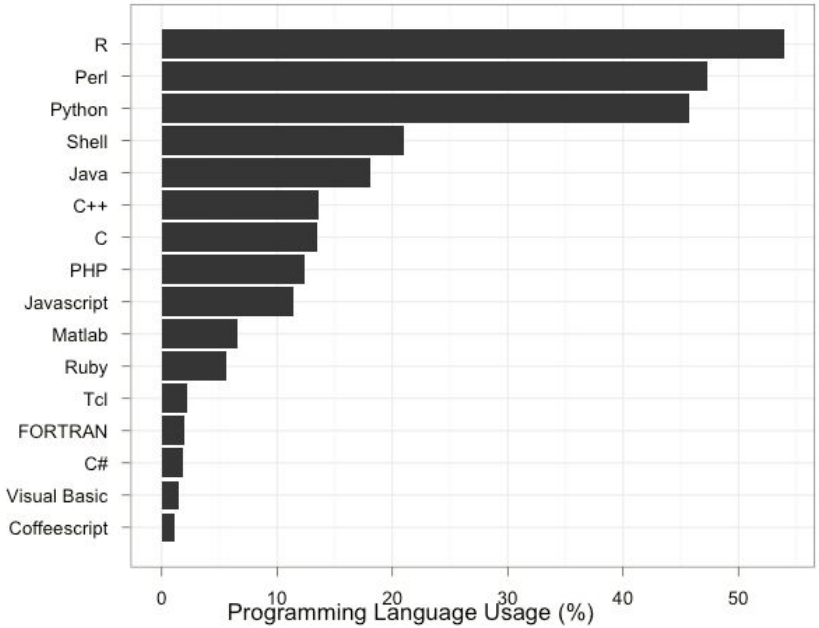
… and Why R?



Software used in data analysis competitions in 2011 (checked in 2016).

http://r4stats.com/articles/popularity/

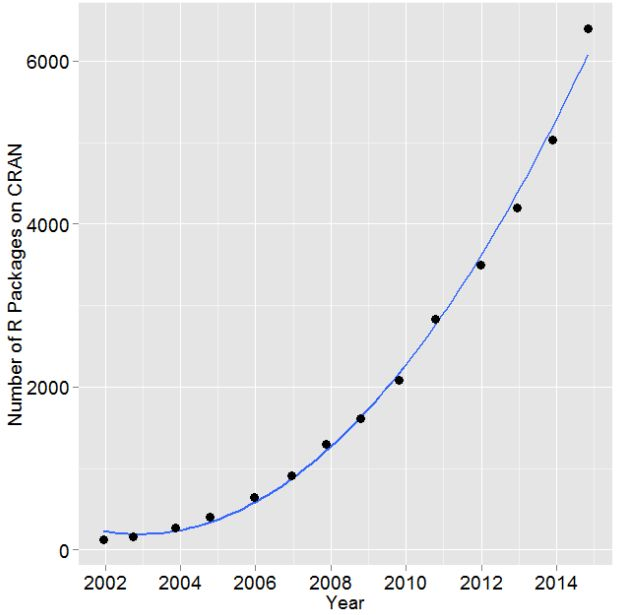
… and Why R?



R in bioinformatics (2012).

http://bioinfsurvey.org/analysis/programming_languages/

… and Why R?



Number of R packages available on its main distribution site for the last version released in each year.

http://r4stats.com/articles/popularity/

R & other programming languages

1. Calling C, C++ and Fortran from R
   a. for computationally intensive tasks.

2. Calling R from C, C++, Java, .Net or Python

parad.py

```
import rpy2.robjects as robjects

robjects.r('set.seed(112)')
x = robjects.r.rnorm(10000000,0,1)
y = robjects.r.rnorm(10000000,0,1)
res = robjects.r['head']
print(res(x.ro/y))
```

R & other programming languages

Calling C++ code with OpenMP from R

parad.cpp

```
#include <Rcpp.h>
#include <cstdlib>
#include <iostream>
#include <omp.h>
using namespace std;
RcppExport SEXP parad(SEXP x, SEXP y){
    int i,n,max;
    Rcpp::NumericVector vector1(x);
    Rcpp::NumericVector vector2(y);
    n=vector2.size();
    Rcpp::NumericVector product(n);
    max=omp_get_max_threads();
    omp_set_num_threads(max);

   #pragma omp parallel for
   for(i=0;i<n;i++){
   product[i]=vector1[i]/vector2[i];
   }
   return(product);
}
```

compilar parametrs

```
$ export PKG_LIBS='`Rscript -e "Rcpp:::LdFlags()"` -fopenmp
-lgomp'
$ export PKG_CXXFLAGS='`Rscript -e "Rcpp:::CxxFlags()"`
-fopenmp'
$ R CMD SHLIB parad.cpp
```

parad.R

```
library(Rcpp)

dyn.load('parad.so')
set.seed(112)
x=rnorm(10000000,0,1)
y=rnorm(10000000,0,1)
head(.Call('parad',x,y))
identical(.Call('parad',x,y),x/y)
```

R vs Matlab

- Freely available (Open source)
- Free packages are stored in the Comprehensive R Archive Network (**CRAN**)
- Very active community
- Many packages for **Symbolic data analysis** (symbolicDA, RSDA) and **factor analysis** available in CRAN.
- Bioconductor - an open source software framework for biologists and bioinformatics

R is great for data analysis and statistics.

- Not Free
- Some toolboxes can be expensive
- Specially developed libraries for matrix operations (**LAPACK**)
- Official releases and updates twice a year
- Excels in parallel computing
- **Simulink** - environment for modeling, simulating and analyzing multidomain dynamic systems

Matlab is great for numerical computing.

Where to learn R?

- An Introduction to R
  - https://cran.r-project.org/doc/manuals/R-intro.pdf
- R style guide:
  - https://google.github.io/styleguide/Rguide.xml
- For Matlab users:
  - http://www.math.umaine.edu/~hiebeler/comp/matlabR.html
- R reference Card
  - http://mirrors.nic.cz/R/doc/contrib/Short-refcard.pdf
- R reference Card for data mining
  - http://mirrors.nic.cz/R/doc/contrib/YanchangZhao-refcard-data-mining.pdf