R
A software environment for comprehensive statistical analysis of astronomical data

Eric Feigelson

Center for Astrostatistics

Penn State University

ADASS XXI 2011 Paris

Brief history of statistical computing

1960s – c2003: Statistical analysis developed by academic statisticians, biometricians, etc. but implementation relegated to commercial companies (SAS, BMDP, Statistica, Stata, Minitab, etc).

1980s: John Chambers (ATT, USA)) develops **S** system, C-like command line interface.

1990s: Ross Ihaka & Robert Gentleman (Univ Auckland NZ) mimic **S** in an open source system, **R**. Expands to ~15 Core Team members, GNU GPL release.

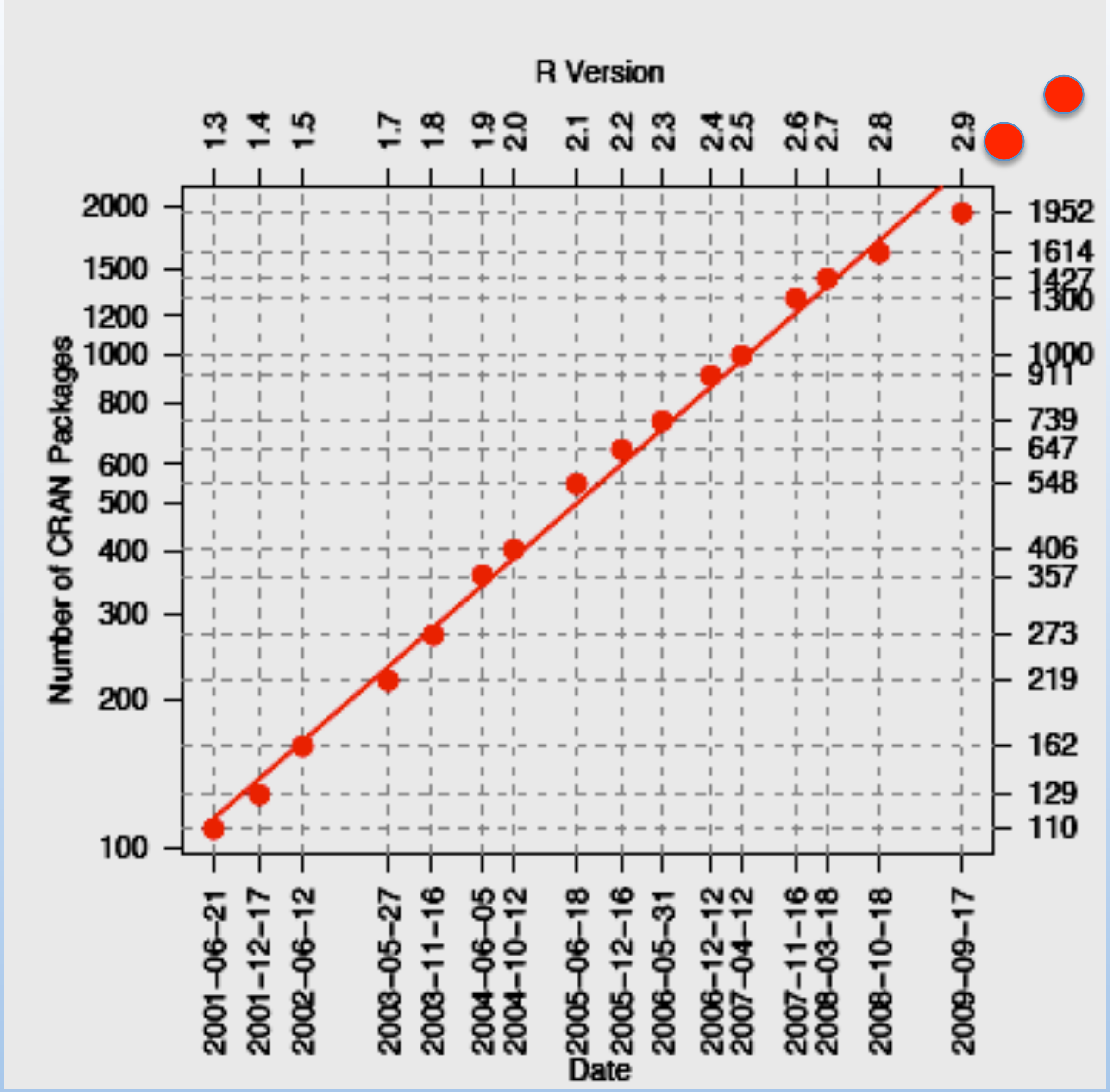
2000s: **Comprehensive R Analysis Network (CRAN)** for user-provided specialized packages grows exponentially. ~20 early packages incorporated into base-R.

By ~2005, **R/CRAN** is the dominant public statistical software system for the development and promulgation of new statistical methodology. Used extensively by statisticians and many user communities (genomics, econometrics, ecology). Estimated 2M users (2010)

The R statistical computing environment

- Ri integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards. Quality control is limited.
- Fully programmable C-like language, similar to IDL. Specializes in vector or matrix inputs.
- Easily downloaded from http://www.r-project.org for Windows, Mac or linux
- Many resources: R help files (3500p for base **R**), on-line tutorials, >100 books, *Use R!* conferences, *The R Journal* & *J. Stat. Software*
- >3300 user-provided add-on **CRAN** packages, >50,000 statistical functions
- Difficulties: Finding what you want, and understanding what you find. Improved education in statistics addresses the latter problem.

Growth of CRAN contributed packages



Oct 1, 2011 count:

3,320 packages

Aspects of the social organization and trajectory of the R project, J. Fox, ***The R Journal***, 1/2, 5 (2009)

Some functionalities of R

arithmetic & linear algebra
bootstrap resampling
empirical distribution tests
exploratory data analysis
generalized linear modeling
graphics
robust statistics
linear programming
local and ridge regression
maximum likelihood estimation
multivariate analysis
multivariate clustering
neural networks
smoothing
spatial point processes
statistical distributions
statistical tests
survival analysis
time series analysis

Selected methods in Comprehensive R Archive Network (CRAN)

Bayesian computation & MCMC, classification & regression trees, genetic algorithms, geostatistical modeling, hidden Markov models, irregular time series, kernel-based machine learning, least-angle & lasso regression, likelihood ratios, map projections, mixture models & model-based clustering, nonlinear least squares, multidimensional analysis, multimodality test, multivariate time series, multivariate outlier detection, neural networks, non-linear time series analysis, nonparametric multiple comparisons, omnibus tests for normality, orientation data, parallel coordinates plots, partial least squares, periodic autoregression analysis, principal curve fits, projection pursuit, quantile regression, random fields, random forest classification, ridge regression, robust regression, self-organizing maps, shape analysis, space-time ecological analysis, spatial analyisis & kriging, spline regressions (MARS, BRUTO), tessellations, three-dimensional visualization, wavelet toolbox

Selected CRAN Task Views
(http://cran.r-project.org/web/views)

Task Views provide brief overviews of CRAN packages by topic & functionality. Maintained be expert volunteers, updated ~monthly

Bayesian	~100 packages
ChemPhys	~70 packages
Cluster	~110 packages
Graphics & gR	~40 packages
HighPerformanceComputing	~60 packages
Machine Learning	~60 packages
Medical imaging	~15 packages
Robust	~20 packages
Spatial	~110 packages
Survival	~140 packages
TimeSeries	~90 packages

Interfaces: BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML
(This is very important for ADASS scientists. R scripts can ingest subroutines from these languages. Two-way communication for C, Fortran, Python & Ruby: you can ingest R functions in your legacy codes.)

I/O: ASCII, binary, bitmap, cgi, FITS, ftp, gzip, HTML, SOAP, URL

Graphics & emulators: Grace, GRASS, Gtk, Matlab, OpenGL, Tcl/Tk, Xgobi

Math packages: GSL, Isoda, LAPACK, PVM

Text processor: LaTeX

Since c.2005, R has been the premier public-domain statistical computing package, growing exponentially.

Some features of R

- Designed for individual use on workstation, exploring data interactively with advanced methodology and graphics. Can be used for automated pipeline analysis. Very similar experience to IDL.

- **R** objects placed into `classes': *numeric, character, logical, vector, matrix, factor, data.frame, list,* and dozens of others designed by **CRAN** packages. *plot, print, summary* functions are adapted to class objects. The *list* class allows a hierarchical structure of heterogeneous objects (like IDL *sav* file).

- Extensive graphics based on SVG, RGTK2, JGD, and other GUIs. See graphics gallery at http://www.oga-lab.net/RGM2.

- Uni- or bi-directional interfaces to other languages: BUGS, C, C++, Fortran, Java, JavaScript, Matlab, Python, Perl, Xlisp, Ruby.

- Only one astronomy **CRAN** package to date: FITSio (limited functionality)

Computational aspects of R

R scripts can be very compact

IDL: temp = mags(where(vels le 200. and vels gt 100, n))
upper_quartile = temp((sort(temp))(ceil(n*0.75)))

R: upper_quartile <- quantile(mags[vels>100. & vels<200.], probs=0.75)

Vector/matrix functionalities are fast (like C); e.g. a million random numbers generated in 0.1 sec, a million-element FFT in 0.3 sec.

Some **R** functions are much slower; e.g.

for (i in 2:1000000) x[i] = x[i-1] + 1

The **R** compiler is now being rewritten from `parse tree' to `byte code' (similar to Java & Python) leading to several-fold speedup.

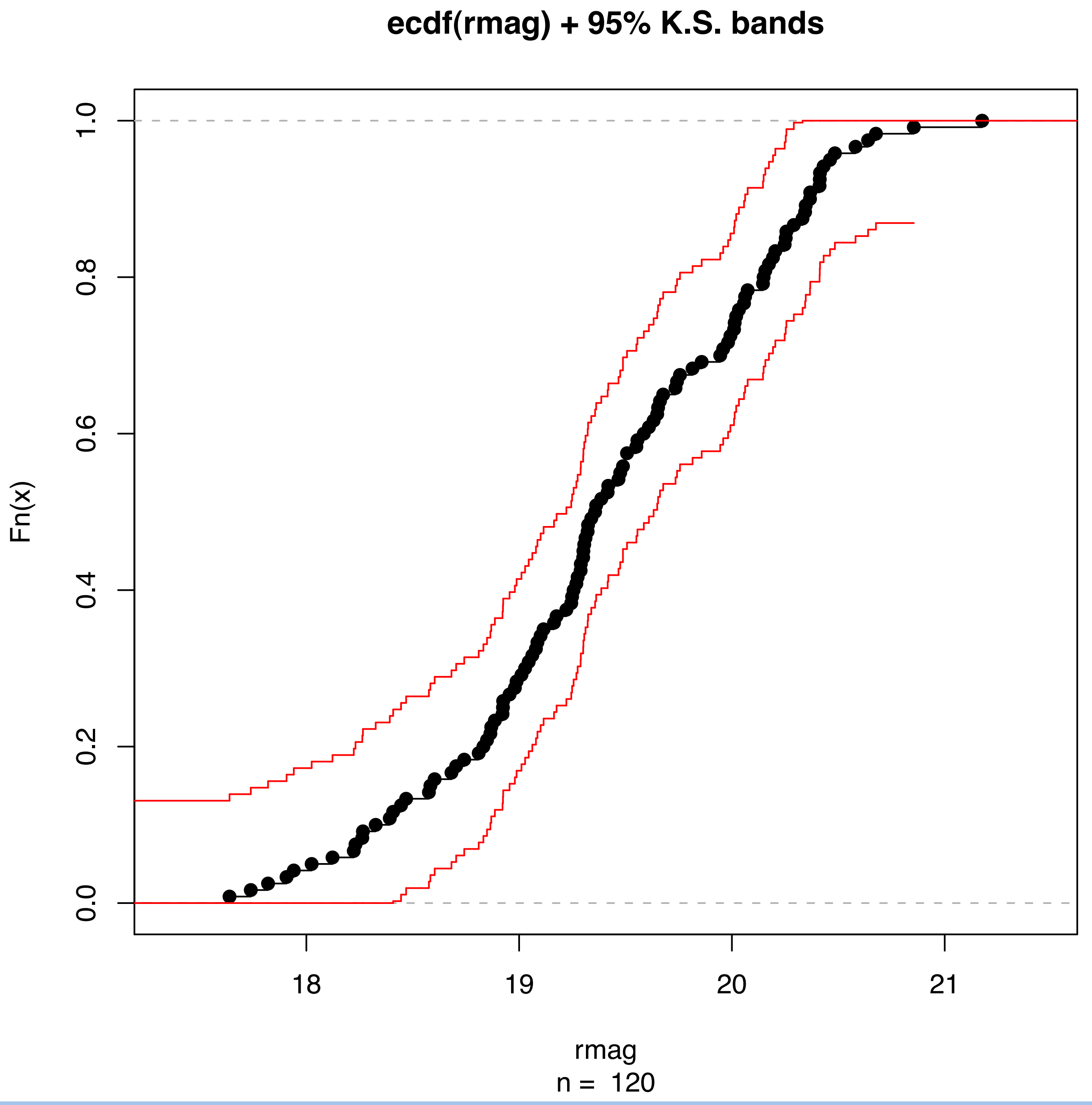
Several dozen **CRAN** packages are devoted to high-performance computing, parallelization, data streams, grid computing, GPUs, (PVM, MPI, NWS, Hadoop, etc). See **CRAN** HPC Task View.

While designed for an individual exploring small datasets,
R can be pipelined and can treat megadatsets

Sample R Script

```
# Start a session
setwd('/Users/e5f/Desktop')      # set working directory
getwd()                          # report working directory
system('pwd')                    # report working directory (from operating system)
citation()                       # citation reference for publications using R


# Construct dataset of 120 Sloan quasar r & z band magnitudes
qso <- read.table('SDSS_QSO.dat', head=T, fill=T)
class(qso)                       # data.frame ~ matrix plus column headings
dim(qso)                         # dimension of data.frame
names(qso)                       # column (variable) names
summary(qso)                     # quartiles and mean

rmag <- qso[1:120,7]             # filter on [rows,columns]
zmag <- qso[1:120,11]
zmag                             # print vector on console
```

```
# Make a simple and a better plot: univariate empirical distribution function

par(mfrow=c(2,1))
plot(ecdf(rmag))
plot(ecdf(rmag), cex=0.0, lwd=2, verticals=T, col='darkgreen', xlim=c(17,22),
        ylim=c(0,1), xlab='Sloan magnitude', ylab='Cum. distribution', main='')
plot(ecdf(zmag), cex=0, verticals=T, lwd=2, col='darkred', add=T)
text(18.5,0.4,'z mag', col='darkred')
text(19.5,0.3,'r mag', col='darkgreen')
dev.copy2eps(file='R_test.eps')
par(mfrow=c(1,1))

# Add 95% bootstrap confidence intervals to e.d.f.  (CRAN package sfsmisc)

install.packages('sfsmisc')            # choose a CRAN mirror site and download package
library('sfsmisc')                     # introduce package into this R session
ecdf.ksCI(rmag)                        # this package automatically generates a plot
```


ecdf(rmag) + 95% K.S. bands
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
18
19
20
21
rmag
n = 120

```
#     Some hypothesis tests in R:

#     ks.test, wilcox.test, mood.test (in R) for univariate 2-sample test
#     chisq.test, fisher.test (R) for contingency tables (categorical data)
#     cor ( R) Pearson r, Kendall tau, Spearman rho tests for correlation
#     ad.test (ADGofTest) for univariate Anderson-Darling test (more sensitive than KS)
#     surv2.ks (surv2sample) for univariate 2-sample test with censoring (upper limits)
#     cenken (NADA) for bivariate correlation test with censoring
#     dip (diptest) for Hartigan's test for univariate multimodality
#     grubbs.test (outliers) test for outliers
#     durbin.watson (car) test for serial autocorrelation
#     cramer.test (Cramer) for multivariate 2-sample test with bootstrap resample
#     mshapiro.test (mvnormtest) for multivariate normality test
#     moran.test (sped) test for randomness vs. autocorrelation in 2 or more dimensions
#     kuiper.test, r.test, rao.test, watson.test (CircStats) tests for uniformity of circular data
```

Projects at Penn State to promulgate R in astronomy

- **VOStat, rev 2 (http://vostat.org)** Web-service access to ~50 simple **R** functions. Input files uploaded or from VO/SAMP. R code given. Limited suite of methods, easy ramp to using R.

- **Summer School in Statistics for Astronomy** Since 2005 in U.S., India, Brazil, teaches established statistical methods for 1 week to ~10% of world's astronomy graduate students.

- ***Modern Statistical Methods for Astronomy with R Applications*** (Feigelson & Babu, Cambridge Univ Press, 2012). Textbook based on Summer School but more comprehensive: probability, statistical inference, distributions, nonparametrics, density estimation, regression, multivariate analysis & classification, censoring & truncation, time series analysis, spatial point processes.

- **Astrostatistics & Astroinformatics Portal (http://asaip.psu.edu)** Recent papers, discussion forums, and other resources for cross-disciplinary methodology (astronomers, statisticians, computer scientists). To be used by ISI/AC, LSST/ISSC/ and [planned] IAU/WG, but readable by everyone. To include R forum.

- **Infrastructure CRAN packages** Wrappers for ***CFITSIO*** and translation of IDL's **astrolib** routines. TBD … help needed.

Selected books on R*

Modern Statistical Methods for Astronomy with R Applications
E. Feigelson & G. J. Babu 2012

An Introduction to R (http://www.r-project.org under `Manuals', dozens of tutorials)

Introductory Statistics with R P. Dalgaard, 2nd ed. 2008

R in a nutshell: A desktop quick reference J. Adler 2009

The R Book, M. Crawley 2007

A Handbook of Statistical Analyses Using R, B. S. Everitt & T. Hothorn 2nd ed, 2009

Software for data analysis: Programming with R, J. Chambers 2008

Introductory Time Series with R Cowpertwait & A. V. Metcalfe 2009
(one of dozens in Springer ***Use R!*** series)

ggplot2: Elegant Graphics for Data Analysis H. Wickham 2nd ed, 2009

* New R books arriving 1/month