Introduction to Astrostatistics and R

Eric Feigelson
Penn State University

Cosmology on the Beach, 2014
Lecture #1
My credentials

Professor of Astronomy & Astrophysics and of Statistics
Assoc Director, Center for Astrostatistics at Penn State
Scientific Editor (methodology), Astrophysical Journal
Chair, IAU Working Group in Astrostatistics & Astroinformatics
Councils, Intl Astrostatistics Assn, AAS/WGAA, LSST/ISSC
Lead author, MSMA textbook (2012 PROSE Award)
Lead editor, Astrostatistics & Astroinformatics Portal

Observational astronomer on small-scale structure at z<0.000001
Outline

Introduction to astrostatistics
- Role of statistics in astronomy
- History of astrostatistics
- Status of astrostatistics today

Introduction to R
- History of statistical computing
- The R language & CRAN packages
- Sample R script
What is astronomy?

**Astronomy** is the observational study of matter beyond Earth: planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations.

**Astrophysics** is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an indirect, inferential intellectual effort based on the assumption that physics – gravity, electromagnetism, quantum mechanics, etc – apply universally to distant cosmic phenomena.
What is statistics?

(No consensus !!)

Statistics characterizes and generalizes data

- “… briefly, and in its most concrete form, the object of statistical methods is the reduction of data” (R. A. Fisher, 1922)

- “Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.” (Wikipedia, 2013)

- “A statistical inference carries us from observations to conclusions about the populations sampled” (D. R. Cox, 1958)
Does statistics relate to scientific models?

\textit{The pessimists} ...

“There is no need for these hypotheses to be true, or even to be at all like the truth; rather … they should yield calculations which agree with observations” (Osiander’ s Preface to Copernicus’ \textit{De Revolutionibus}, quoted by C. R. Rao)

“Essentially, all models are wrong, but some are useful.”

(Box & Draper 1987)

"The object [of statistical inference] is to provide ideas and methods for the critical analysis and, as far as feasible, the interpretation of empirical data ... The extremely challenging issues of scientific inference may be regarded as those of synthesising very different kinds of conclusions if possible into a coherent whole or theory ... The use, if any, in the process of simple quantitative notions of probability and their numerical assessment is unclear." (D. R. Cox, 2006)
The optimist ...

“The goal of science is to unlock nature’s secrets. ... Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. ... Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical inference.”

(P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, 2005)
**Recommended steps in the statistical analysis of scientific data**

The application of statistics can reliably quantify information embedded in scientific data and help adjudicate the relevance of theoretical models. But this is not a straightforward, mechanical enterprise. It requires:

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities (easiest step with R)
- judicious scientific evaluation of the results

_Astronomers often do not adequately pursue each step_
Some further comments

• Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (jargon problem!), and there are usually several ways to proceed.

• Some statistical procedures are based on mathematical proofs which determine the applicability of established results; it is easy to ignore these limits and emerge with unreliable results. It is perilous to violate mathematical truths! Some issues are debated among statisticians, or have no known solution.

• It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. P-values are not necessarily useful … we are scientists first! Statistics is only a tool towards understanding nature from incomplete information.

We should be knowledgeable in our use of statistics and judicious in its interpretation
Astronomy & Statistics: A glorious past

For most of western history, the astronomers were the statisticians!

Ancient Greeks – 18th century
What is the best estimate of the length of a year from discrepant data?
• Middle of range: Hipparcos (4th century B.C.)
• Observe only once! (medieval)
• Mean: Brahe (16th c), Galileo (17th c), Simpson (18th c)
• Median (20th c)

19th century
Discrepant observations of planets/moons/comets used to estimate orbital parameters using Newtonian celestial mechanics
• Legendre, Laplace & Gauss develop least-squares regression and normal error theory (c.1800-1820)
• Prominent astronomers contribute to least-squares theory (c.1850-1900)
The lost century of astrostatistics....

In the late-19th and 20th centuries, statistics moved towards human sciences (demography, economics, psychology, medicine, politics) and industrial applications (agriculture, mining, manufacturing).

During this time, astronomy recognized the power of modern physics: electromagnetism, thermodynamics, quantum mechanics, relativity. Astronomy & physics were wedded into astrophysics.

Thus, astronomers and statisticians substantially broke contact; e.g. the curriculum of astronomers heavily involved physics but little statistics. Statisticians today know little modern astronomy.
The state of astrostatistics today
*(not good!)*

The typical astronomical study uses:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression for model fits (Legendre 1805, Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

Even traditional methods are often misused:

*see Friday’s lecture*

`Common statistical mistakes in the astronomical literature’
Under-utilized methodology:

- modeling (MLE, EM Algorithm, BIC, bootstrap)
- multivariate classification (LDA, SVM, CART, RFs)
- time series (autoregressive models, state space models)
- spatial point processes (Ripley’s K, kriging)
- nondetections (survival analysis)
- image analysis (computer vision methods, False Detection Rate)
- statistical computing (R)

Advertisement ....

**Modern Statistical Methods for Astronomy with R Applications**
E. D. Feigelson & G. J. Babu,
Cambridge Univ Press, 2012
Cosmology

Galaxy clustering
Galaxy morphology
Galaxy luminosity fn
Power law relationships
Weak lensing morphology
Strong lensing morphology
Strong lensing timing
Faint source detection
Multiepoch survey lightcurves
CMB spatial analysis
ΛCDM parameters
Comparing data & simulation

Statistics

Spatial point processes, clustering
Regression, mixture models
Gamma distribution
Pareto distribution
Geostatistics, density estimation
Shape statistics
Time series with lag
False Discovery Rate
Multivariate classification
Markov fields, ICA, etc
Bayesian inference & model selection
under development

Comparing data & simulation

under development
Recent resurgence in astrostatistics

• Improved access to statistical software. R/CRAN public-domain statistical software environment with thousands of functions. Increasing capability in Python (http://www.astro.cornell.edu/staff/loredo/statpy/essentials.html).

• Papers in astronomical literature doubled to ~500/yr in past decade (“Methods: statistical” papers in NASA-Smithsonian Astrophysics Data System)

• Short training courses (Penn State, India, Brazil, Spain, USA, Greece, China, this school)


• Scholarly society working groups and a new integrated Web portal (ISI, IAU, AAS, LSST, http://asaip.psu.edu)

Two advertisements!

Consider attending the first IAU conference on astrostatistics:
Statistical Challenges in 21st Century Cosmology
IAU Symposium 306, Lisbon PT, May 2014

Consider joining ASAIP and its associated organizations (IAA, IAU/WGAA, AAS/WGAA & LSST/ISSC):
http://asaip.psu.edu

Welcome to ASAIP

The Astrostatistics and Astroinformatics Portal (http://asaip.psu.edu) is a new Web site serving the cross-disciplinary communities of astronomers, statisticians and computer scientists. It is intended to foster research into advanced methodologies for astronomical research, and to promulgate such methods into the broader astronomy community. The WWW public is welcome to read materials in ASAIP. Use the navigation bar above, or the search box at the upper right, to find material throughout the ASAIP Web site.
A brief history of statistical computing

1960s – c2003: Statistical analysis developed by academic statisticians, 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.

Early–2000s: Comprehensive R Analysis Network (CRAN) for user–provided specialized packages grows exponentially. Important packages incorporated into base–R.
Growth of CRAN contributed packages

Jan 2014:
~5050 packages
~100,000 functions

The R statistical computing environment

- R integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards. But quality control is limited.

- Fully programmable C-like language, similar to IDL. Specializes in vector/matrix inputs.

- Easy download from [http://www.r-project.org](http://www.r-project.org) for Windows, Mac or Linux. On-the-fly installation of CRAN packages.

- >5000 user-provided add-on CRAN packages, tens of thousands of statistical functions

• Principal difficulty: Finding what you want, and understanding what you find. Google helps the former problem. Improved education in statistics addresses the latter problem.
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
- max 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)

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

CRAN Task Views provide brief overviews of CRAN packages by topic & functionality. Maintained be expert volunteers. Partial list:

- Bayesian ~110 packages
- ChemPhys ~60 packages (incl. 11 for astronomy)
- Cluster ~90 packages
- Graphics ~40 packages
- HighPerformanceComputing ~80 packages
- Machine Learning ~70 packages
- Medical imaging ~25 packages
- Robust ~25 packages
- Spatial ~125 packages
- Survival ~175 packages
- TimeSeries ~140 packages
Interfaces: BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML

*This is very important for astronomers.* R scripts can ingest subroutines from these languages. Packages exist for 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 with >2M users.
Some features of R

- Designed for individual use on workstation, exploring data interactively with advanced methodology and graphics. But it 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 GUls. See graphics gallery at http://www.oga-lab.net/RGM2 and http://gallery.r-enthusiasts.com/
Computational aspects of R

R scripts can be very compact

**IDL:**
\[
\text{temp} = \text{mags} (\text{where}(\text{vels le } 200. \text{ and } \text{vels gt } 100, n))
\]
\[
\text{upper\_quartile} = \text{temp} (\text{(sort}(\text{temp}))(\text{ceil}(n*0.75)))
\]

**R:**
\[
\text{upper\_quartile} \leftarrow \text{quantile} (\text{mags}[\text{vels}>100. \& \text{vels}<200.], \text{probs}=0.75)
\]

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.
\[
\text{for (i in 2:1000000) x[i] = x[i-1] + 1}
\]

The **R** compiler rewritten in 2012 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 originally designed for an individual exploring small datasets, R can be pipelined and can treat megadatsets*
Projects at Penn State to promulgate R in astronomy

- **VOStat** ([http://vostat.org](http://vostat.org)) Web-service to ~50 simple R functions.

- **Summer School in Statistics for Astronomy** Since 2005 (U.S., India, …) teaches established statistical methods to ~10% of world’s astronomy graduate students.


- **Astrostatistics & Astroinformatics Portal** ([http://asaip.psu.edu](http://asaip.psu.edu)) Recent papers, discussion forums, and other resources

- **Infrastructure CRAN packages** CFITSIO, IDL’s astrolib, datasets
A vision of astrostatistics in 2025 …

- Astronomy curriculum has 1 year of statistical inference and methodology

- Young astronomers have M.S. degrees in statistics and computer science

- Astrostatistics is a significant niche cross-disciplinary field involving a few percent of astronomers (cf. astrochemists, instrumentalists). Well-funded cross-disciplinary research groups in astrostatistics and astroinformatics in several nations push frontiers of astrostatistical methodology.

- Astronomical papers reference statistics monographs.

- Astronomers regularly use hundreds of methods coded in P, the successor to Q and R.

- *Statistical Challenges in Modern Astronomy* meetings are held annually with ~250 participants