

Astrostatistics: Past, Present and Future

Eric Feigelson
Penn State University

Summer School in Statistics for Astronomy
June 2021

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 !!)*

- “... briefly, and in its most concrete form, the object of statistical methods is the reduction of data”
(R. A. Fisher, 1922)
- “Statistics is the mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.”
(Wikipedia, 2014)
- “Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.”
(Wikipedia, 2015)
- “A statistical inference carries us from observations to conclusions about the populations sampled”
(D. R. Cox, 1958)

Does statistics relate to scientific models?

The pessimists ...

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

(Box & Draper 1987)

“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’ *De Revolutionibus*, quoted by C. R. Rao in *Statistics and Truth*)

"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 positivists ...

“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:

- model-independent exploration of the data
- careful statement of the scientific problem
- definition of model(s) 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

- 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. Very confusing.
- Some statistical procedures are based on mathematical proofs which determine the applicability of established results. It is perilous to violate mathematical truths! Some issues are debated among statisticians, or have no known solution.
- Scientific inferences should not depend on arbitrary choices in methodology & variable scale. Prefer nonparametric & scale-invariant methods when disciplinary knowledge is vague. Try multiple methods.
- It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. 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 to 18th century

Best estimate of the length of a year from discrepant data?

- Middle of range: Hipparchos (4th century B.C.)
- Observe only once! (medieval)
- Mean: Brahe (16th c), Galileo (17th c), Simpson (18th c)
- Median w/ bootstrap (21th 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 (~1800-1820)
- Prominent astronomers contribute to least-squares theory (~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

(improving)

Many astronomical studies are still confined to a narrow suite of familiar statistical methods:

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

Even traditional methods are sometimes misused!

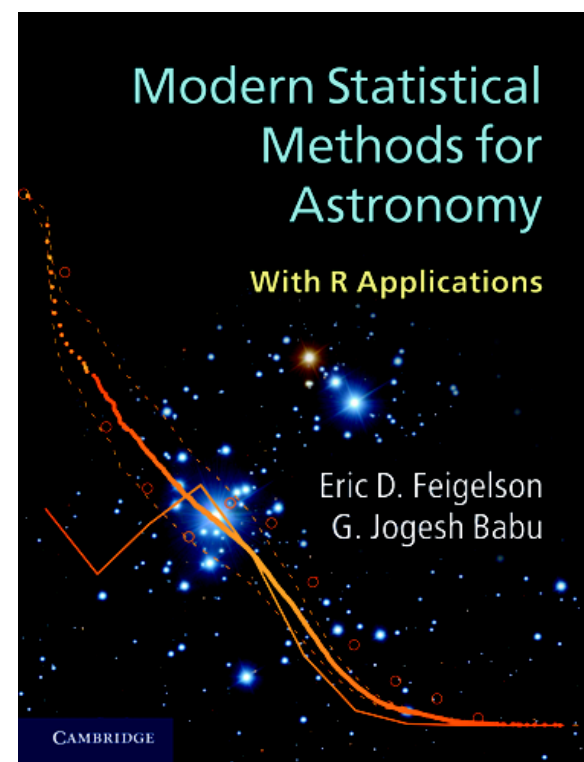
Under-utilized methodology:

- 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















*Winner 2012 PROSE Award for
Best Astronomy & Cosmology Book*

Astrostatistics is difficult due to the breadth of methods needed ...

Cosmology 

Statistics

Galaxy clustering		Spatial point processes, clustering
Galaxy morphology		Regression, mixture models
Galaxy luminosity fn		Gamma distribution
Power law relationships		Pareto distribution
Weak lensing morphology		Geostatistics, density estimation
Strong lensing morphology		Shape statistics
Strong lensing timing		Time series with lag
Faint source detection		False Discovery Rate
Multiepoch survey lightcurves		Multivariate classification
CMB spatial analysis		Markov fields, ICA, etc
Λ CDM parameters		Bayesian inference & model selection
Comparing data & simulation		Uncertainty Quantification

Recent resurgence in astrostatistics

- Improved access to statistical software. R/CRAN public-domain statistical software environment with thousands of functions. Increasing capability in Python.
- Papers with sophisticated methodology in astronomical literature greatly increases
- Short training courses (Penn State, India, Brazil, Greece, China, Italy, France, Germany, Spain, Sweden, LSST, IAU/AAS/CASCA/... meetings)
- Cross-disciplinary research collaborations (Harvard/ICHASC, Carnegie-Mellon, Penn State, NASA-Ames/Stanford, CEA-Saclay/Stanford, Cornell, UC-Berkeley, Michigan, Imperial College London, Swinburne, Texas A&M, JPL, LANL, ...)
- Cross-disciplinary conferences (*Statistical Challenges in Modern Astronomy, Astronomical Data Analysis 1991-2021, PhysStat, SAMSI 2006/2012, Astroinformatics 2012--*, *IAU Symposia 2014--*, *IEEE Symposia 2018--*)
- Scholarly society working groups and a new integrated Web portal <http://asaip.psu.edu> serving: Int'l Astrostatistical Assn, Int'l AstrInformatics Assn, Int'l Stat Institute SIGAstro, Int'l Astro Union Commission B3, Amer Astro Soc Working Group, Amer Stat Assn Interest Group, LSST Science Collaboration, IEEE Astro Data Miner Task Force)

A new imperative: Large-scale surveys & megadatasets

Huge imaging, spectroscopic & multivariate datasets are emerging from specialized survey projects & telescopes:

- 10^9 - 10^{10} -object photometric catalogs x 10^0 - 10^3 epochs from 2MASS, SDSS, VISTA, CRTS/ZTF, Pan-STARRS, DES, **LSST** ...
- 10^6 - 10^8 - galaxy redshift catalogs from SDSS, LAMOST, ...
- 10^9 star astrometric catalog from Gaia
- Spectral-image datacubes (VLA, ALMA, IFUs)
- Radio interferometer data streams (e.g. 30 Tflops processor for LOFAR)

The Virtual Observatory is an international effort to federate many distributed on-line astronomical databases.

Powerful statistical tools are needed to derive scientific insights from TBy-PBy-EBy databases

New resources in astrostatistics

Textbooks

Modern Statistical Methods for Astronomy with R Applications,
Feigelson & Babu, 2012

Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data,
Ivecic, Connolly, VanderPlas & Gray, 2014

Societies (join one!)

Intl Astrostatistics Assn (2010)

AAS Working Group in Astrophysics & Astrostatistics (2013)

ASA Interest Group in Astrostatistics (2014)

IAU Commissions B1–B2–B3 & WG/TDA (2015)

IEEE Task Force on Astro Data Mining (2016)

Intl Astroinformatics Assn (c2018)

*An important under-utilized
resource is the public-domain*

R

statistical software environment

A brief history of statistical computing

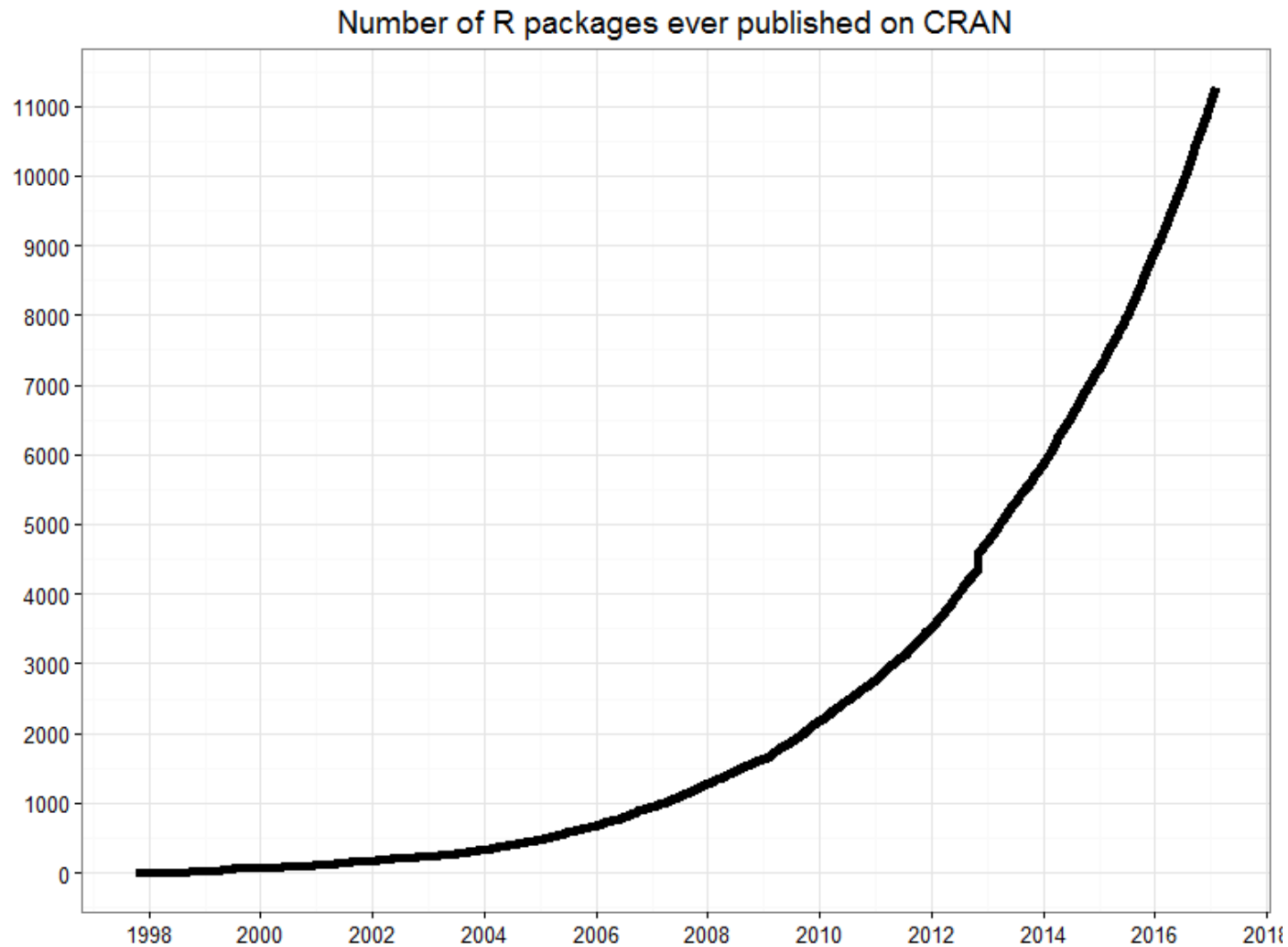
1960s – c2000: Statistical analysis developed by academic statisticians, but implementation relegated to commercial companies for mainframes and PCs (SAS, BMDP, Statistica, Stata, Minitab, etc).

1980s: John Chambers (ATT, USA) develops S system, written in C with a command line interface.

1990s: Ross Ihaka & Robert Gentleman (Univ Auckland NZ) mimic S in an open source system, R. R Core Development Team expands, GNU GPL release.

Early–2000s: Comprehensive R Analysis Network (CRAN) of specialized packages grows exponentially. Important packages incorporated into base-R.

Growth of CRAN contributed packages



May 13 2021:
17,591 packages
(growing ~5/day)

~150,000
functions ?

R's growing importance in data science

Consultants



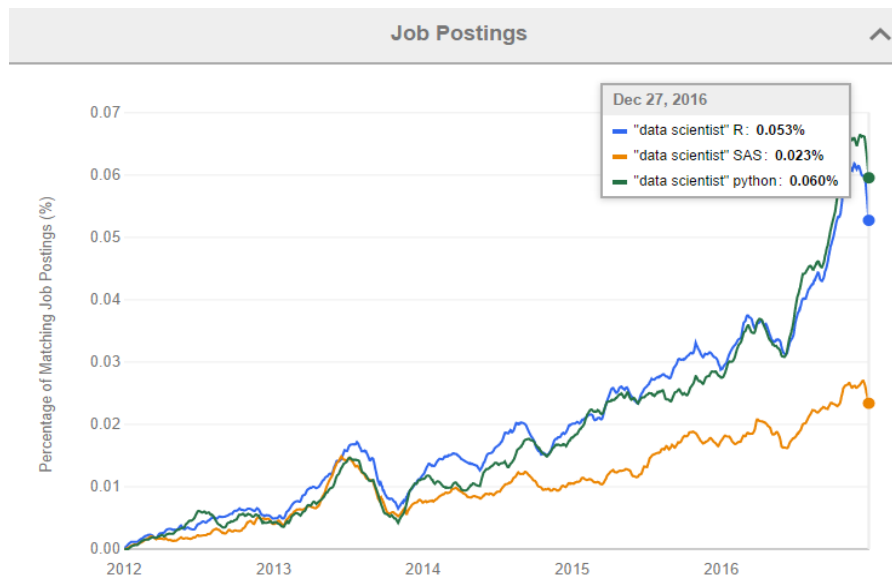
Academics



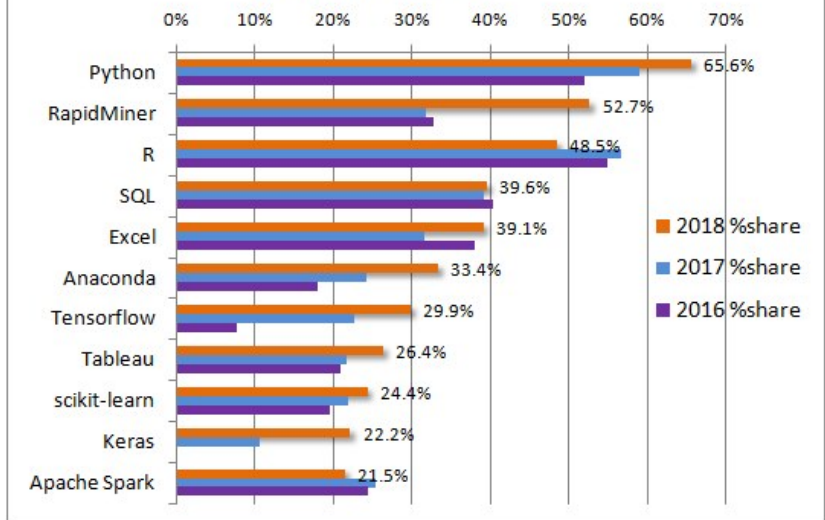
Rexer Analytics Data Miner Survey 2017

“The biggest change in tool adoption we’ve seen over [2007–17] has been the dramatic growth in the use of R.

Job trends for Data Scientist 2012–18



KDnuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018



The R statistical computing environment

- R integrates data manipulation, graphics and extensive statistical analysis. Uniform documentation and coding standards. But quality control is limited for community-provided CRAN packages.
- Fully programmable C-like language, similar to IDL. Specializes in vector/matrix inputs.
- Easy download from <http://www.r-project.org> for Windows, Mac or linux. On-the-fly installation of CRAN packages. Quick communication with C, Fortran, C++, Python. Emulator of Matlab.
- ~17,000 user-provided add-on **CRAN** packages, ~150,000 statistical functions.

- Many resources: R help files, CRAN Task Views and vignette files, on-line tutorials, >200 books, hundreds of blogs (r-bloggers.com), *Use R!* conferences, galleries, companies, *The R Journal* & *J. Stat. Software*, etc.

Principal steps for using R in astronomical research:

- ***Knowing what you want*** [education, consulting, thought]
- ***Finding what you want*** [Google, Rseek, Rdocumentation]
- ***Writing R scripts*** [R Help files, books, StackOverflow]
- ***Understanding the results*** [education, consulting, thought]

With R, astronomers can spend more time thinking, learning & exercising methodology, and less time coding

Some functionalities of base 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)

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 analysis & kriging, spline regressions, tessellations, three-dimensional visualization, wavelet toolbox

CRAN Task Views

(<http://cran.r-project.org/web/views>)

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

- Bayesian ~110 packages
- Chem/Phys ~75 packages (incl. 20 for astronomy)
- Cluster/Mixture ~100 packages
- Graphics ~40 packages
- HighPerfComp ~75 packages
- Machine Learning ~70 packages
- Medical imaging ~20 packages
- Robust ~50 packages
- Spatial ~135 packages
- Survival ~200 packages
- TimeSeries ~170 packages

***Since c.2010, R has been the
world's premier
statistical computing package***

**Data scientists recommend both Python and R
Usage of both is growing rapidly**

(<https://asaip.psu.edu/forums/software-forum/195790576>)

A vision of astrostatistics by 2030 ...

- Astronomy graduate curriculum has 1 year of statistical and computational methodology
- Some astronomers have M.S. in statistics or computer science
- Astrostatistics and astroinformatics is a well-funded, cross-disciplinary research field
- Astronomers regularly use many methods from R/CRAN
- Many astronomers take advantage of information education and cross-disciplinary conferences in astrostatistics.