Astrostatistics: Past, Present and Future

Eric Feigelson

Dept. of Astronomy & Astrophysics

Center for Astrostatistics

Penn State University

edf@astro.psu.edu

IMAV 2013 Valparaiso

Outline

Statistics and Astronomy

Astrostatistics: Past

Astrostatistics: Present

Astrostatistics: Future --- IMAV 2013

What is astronomy?

Astronomy (astro = star, nomen = name in Greek) 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. The perspective is rooted from our viewpoint on or near Earth using telescopes or robotic probes.

Astrophysics (astro = star, physis = nature) 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 gravity, electromagnetism, quantum mechanics, plasma physics, chemistry, and so forth – 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 the study of the collection, organization, analysis, interpretation, and presentation of data. There is also a discipline called mathematical statistics that studies statistics mathematically." (Wikipedia, 2012)

- "[Statistics is] the study of algorithms for data analysis" (R. Beran)

- "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 …

“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)

“`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, 2005)

My personal conclusions
(X-ray astronomer with 25 yrs statistical experience)

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 simple mechanical enterprise. It requires: careful statement of the problem, model formulation, choice of statistical method(s), calculation of statistical quantities, and judicious evaluation of the result. Astronomers do not adequately pursue each of these steps.

Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (hundreds of books, jargon problem!). Some procedures are rooted in theorems, while others are debated inconclusively among statisticians.

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.

Astronomers should be knowledgeable in 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 closely 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:

- Six unweighted bivariate least squares fits are used interchangeably, often with wrong confidence intervals

  Feigelson & Babu ApJ 1992
- Use of the likelihood ratio test for comparing two models is often inconsistent with asymptotic statistical theory

  Protassov et al. ApJ 2002
- K-S goodness-of-fit probabilities are inapplicable when the model is derived from the data, K-S less sensitive than Anderson-Darling test

  Babu & Feigelson ADASS 2006

Missed or under-used methodology:

- modeling (MLE, EM Algorithm, BIC, bootstrap)
- multivariate analysis & classification (MANOVA, 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)
- data mining (Random Forests, Support Vector Machines)
- statistical computing (R)

Advertisement ….

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

A new imperative: Large-scale surveys, megadatasets & the Virtual Observatory

Huge, uniform, multivariate databases are emerging from specialized survey projects & telescopes:

- 10^9-object photometric catalogs from USNO, 2MASS, SDSS, VISTA, …
- $10^{6\text{-}8}$- galaxy redshift catalogs from 2dF, SDSS, LAMOST
- $10^{6\text{-}7}$-source radio/infrared/X-ray catalogs (WISE, Herschel, eROSITA)
- Spectral-image (3D) databases (VLA, ALMA, IFUs)
- $10^{9\text{-}10}$-object x 1000 epochs (3D) surveys (PTF, Pan-STARRS, VISTA, LSST, …)

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

Powerful statistical tools are needed to derive scientific insights from tera/peta/exa-byte dababases extracted using VO technologies

Recent resurgence in astrostatistics

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

- Improved access to statistical software. R/CRAN public-domain statistical software environment with thousands of functions. Increasing capability in Python

- Short R training courses (Penn State, India, Brazil, Spain, USA, Greece, China, Chile)

- Cross-disciplinary research collaborations (Harvard, Carnegie-Mellon, Penn State, NASA-Ames/Stanford, CEA-Saclay/Stanford, Cornell, UC-Berkeley, Michigan, Imperial College London, LSST Statistics & Informatics Science Collaboration, Valparaiso…)

- Cross-disciplinary conferences (*Statistical Challenges in Modern Astronomy, Astronomical Data Analysis, PhysStat*, SAMSI 2012, *Astroinformatics 2012, …*)

- Textbooks:
    - *Practical Statistics for Astronomers, Wall/Jenkins, 2nd ed 2012*
    - *Modern Statistical Methods for Astronomy with R Applications, Feigelson/Babu, 2012*
    - *Machine Learning and Data Mining in Astronomy, Ivezic et al., 2013)*

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

The Astrostatistics & Astroinformatics Portal

Astrostatistics:
The Present into the Future

The First International Meeting in Astrostatistics, Valparaiso

IMAV 2013

New technology instruments

- Adaptive optics (Andres Guesalaga) Stochastic time series from atmospheric turbulence+ image deconvolution
- Radio interferometry (Pablo Roman) Measured Fourier component → optimized images; data cubes → science (3D image segmentation & characterization)
- Optical amplitude interferometry (Paul Nunez) Image recovery from Fourier amplitudes for bright stars
- New Chilean instruments (Chris Smith) Dark Energy Survey & Large Synoptic Survey Telescope

Wide-field multi-epoch surveys

- Variable star classification (Susana Eyeramendy) VVV: Photometry for millions of variable stars. Hierarchical LASSO classifier
- Variable star characterization (Pablo Estevez) Correntropy periodograms

Era of precision cosmology

- Cosmological parameters (Antonella Cid) Nonlinear hierarchical model, several datasets, prior constraints. Lagrangian multipliers, Bayesian model selection
- Foreground vs. cosmic microwave emission (Jerome Bobin) Independent Component Analysis
- Simulations of galaxy populations (Nelson Padilla) Boundary between astrophysical models and astronomical surveys

Various astronomical problems

- Henry Boffin: Binary star mass ratios
- Veronica Motta: Quasar accretion disk sizes via microlensing
- Andres Jordan: Extrasolar planet spectral time series
- Nestor Espinoza: Modeling stellar noise to characterize planetary transits

Mathematics for astronomy

- Rolando Biscay: Functional Data Analysis with many possible applications: classification of stellar spectra, time series, galaxy morphologies, …

Organizational developments

- Astroinformatics Laboratory (Eduardo Vera) The Center for Mathematical Modeling + Natl Lab High Performance Computing, provides powerful resources for Chile
- Astrostatistics & Astroinformatics Portal and ISI, IAU, AAS & LSST working groups

Training in statistics & informatics tools

- Mathematical statistics (Jogesh Babu) Well-established principles & mathematics underlying modern statistics (probability, MLE, bootstrap, …)
- Virtual Observatory (Amalia Bayo) Tools to access and analyze multi-observatorydata: TOPCAT, VOSA, …
- R & CRAN software (Eric Feigelson) Statistical methodology implemented in a large, coherent, free software system

IMAV 2013 is a window into contemporary interface between statistics and astronomy.

Astronomy → Astrophysics advances with new instrumentation, new discoveries, new interpretations, new understanding. There is increasing recognition that improved data analysis and science analysis are crucial elements of this process.

”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.”

Thanks to

Michel Cure & colleagues

for organizing this stimulating meeting!