3rd INPE Advanced School in Astrophysics: Astrostatistics

The Frequentist Approach to Astrostatistics

Eric Feigelson
(Center for Astrostatistics, Penn State)

Lecture 1: Introduction to statistics and astrostatistics

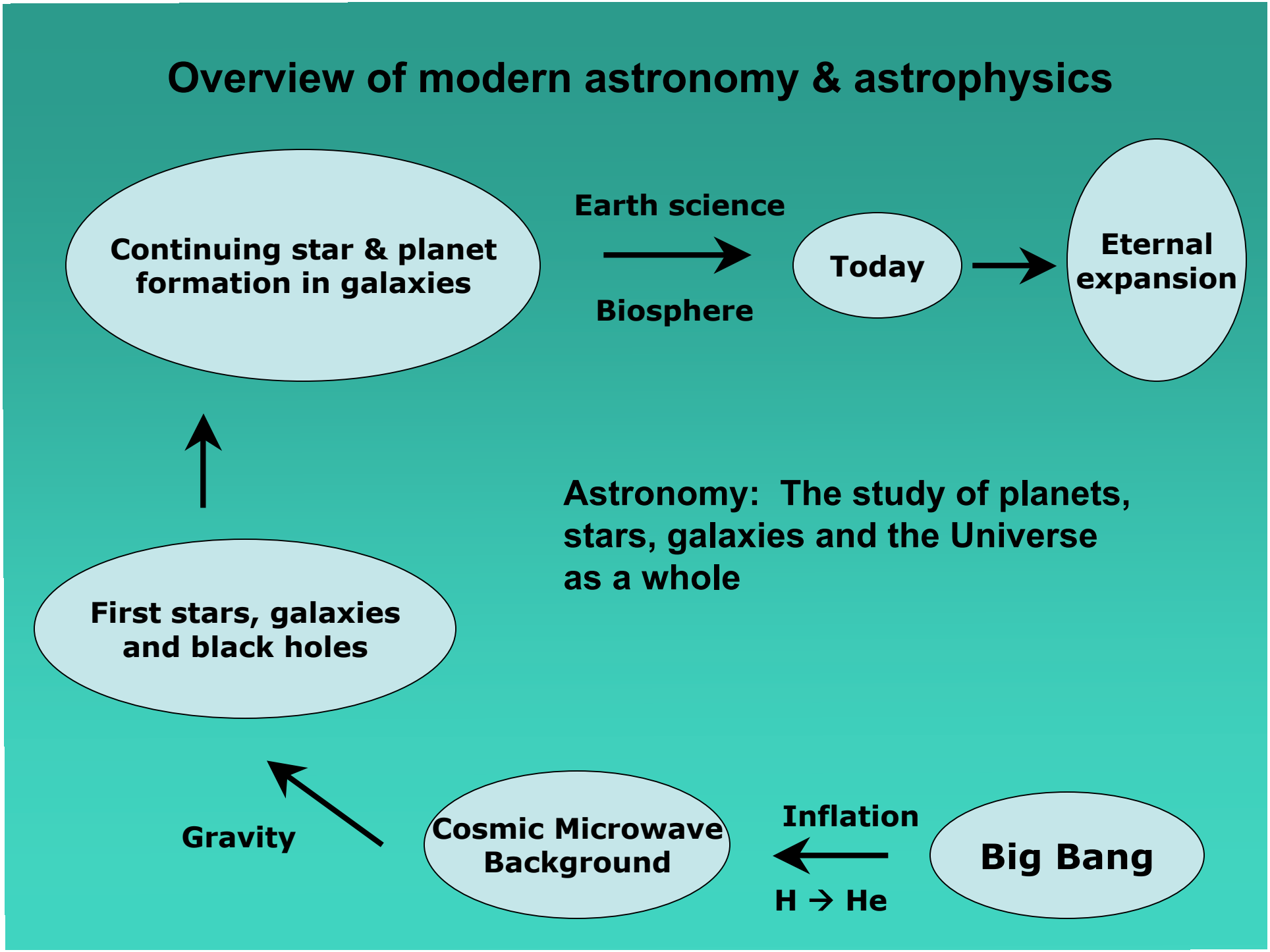
- The role of statistics in scientific research
- Scope of modern statistics
- Astrostatistics: past, present and future
- Resources for astrostatistics

What is astronomy?

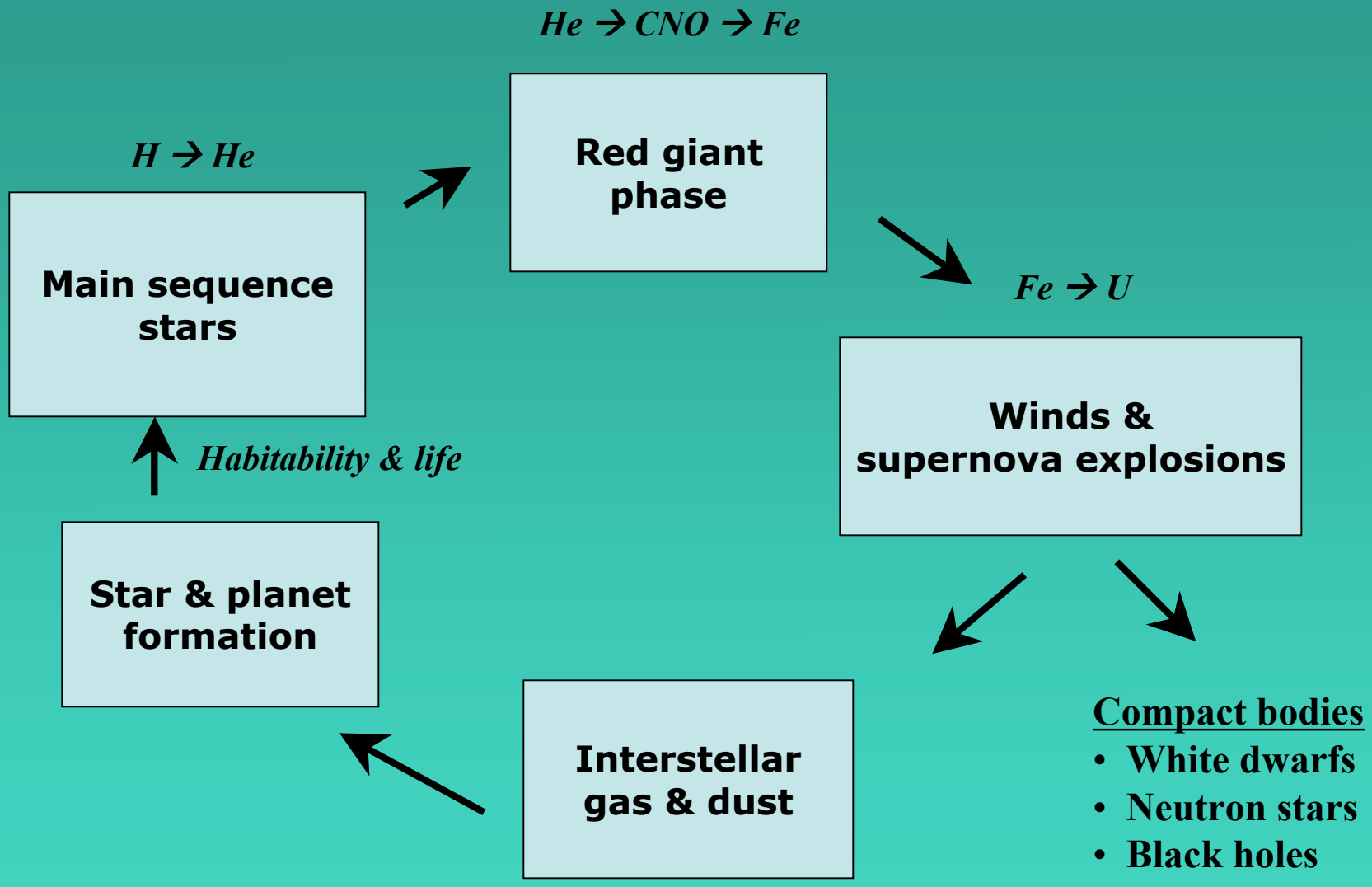
Astronomy (astro = star, nomen = name in Greek) is the observational study of matter beyond Earth – planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations. The perspective is rooted from our viewpoint on or near Earth using 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.

Overview of modern astronomy & astrophysics



Lifecycle of the stars



What is statistics?
(No consensus !!)

Statistics characterizes and generalizes data

- "The first task of a statistician is cross-examination of data" (R. A. Fisher, 1949)
- "Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data" (Wikipedia, 2009.5)
- "[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?

"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)

"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 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 to scientific data is not a straightforward, 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.

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 issues are debated among statisticians. Many 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 can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. We are scientists first! Statistics is only a tool. We should be knowledgeable in our use of statistics and judicious in its interpretation.

Statistics: Some basic definitions

- Statistical inference
  - Seeking quantitative insight & interpretation of a dataset
- Hypothesis testing
  - To what confidence is a dataset consistent with a previously stated hypothesis?
- Estimation
  - Seeking the quantitative characteristics of a functional model designed to explain a dataset. An estimator seeks to approximate the unknown parameters based on the data
- Probability distribution
  - A parametric functional family describing the behavior of a parent distribution of a dataset (e.g. Gaussian = normal)
- Nonparametric statistics
  - Inference based directly on the dataset without parametric models Independent & identically distributed (iid) data point
  - A sample of similarly but independently acquired quantitative measurements.

Some basic definitions (cont.)

- Frequentist statistics
  - Suite of classical inference methods based on simple probability distributions. Hypotheses are fixed while data vary.
- Bayesian statistics
  - Suite of inference methods based on Bayes' Theorem based on likelihoods and prior distributions. Data are fixed while hypotheses vary.
- L_1 and L_2 methods
  - 19th century frequentist methods for estimation based on minimizing the absolute or squared deviations between a sample and a model
- Maximum likelihood methods
  - 20th century frequentist methods for parametric estimation based on the likelihood that a dataset fits the model (often like L_2)
- Gibbs sampling, MCMC, …
  - New computational methods for integrations over hypothesis space in Bayesian statistics

Some basic definitions (cont.)

- Robust methods
  - Statistical procedures that are insensitive to data outliers
- Model selection & validation
  - Procedures for estimating the goodness-of-fit and choice of parametric model. (Nested vs. non-nested models, model misspecification)
- Statistical power, efficiency & bias
  - Mathematical evaluation of the effectiveness of a statistical procedure to achieve its desired goals
- Two-sample & k-sample tests
  - Statistical tests giving probabilities that k samples are drawn from the same parent sample
- Independent & identically distributed (iid) data point
  - A sample of similarly but independently acquired quantitative measurements.
- Heteroscedasticity
  - A failure of iid due to differently weighted data points

Some fields of applied statistics

- Multivariate analysis
  - Establishing the structure of a table of rows & columns
  - Analysis of variance, regression, principal component analysis, discriminant analysis, factor analysis
- Multivariate classification
  - Dividing a multivariate dataset into distinct classes
- Correlation & regression
  - Establishing the relationships between variables in a sample
- Time series analysis
  - Studying data measured along a time-like axis
- Spatial analysis
  - Studying point or continuous processes in 2-3-dimensions
- Survival analysis
  - Studying data subject to censoring (e.g. upper limits)
- Data mining
  - Studying structures in mega-datasets
- Biometrics, econometrics, psychometrics, chemometrics, quality assurance, geostatistics, astrostatistics, …, …

Astronomy & statistics: A glorious history

Hipparchus (4th c. BC): **Average via midrange of observations**

Galileo (1572): **Average via mean of observations**

Halley (1693): **Foundations of actuarial science**

Legendre (1805): **Cometary orbits via least squares regression**

Gauss (1809): **Normal distribution of errors in planetary orbits**

Quetelet (1835): **Statistics applied to human affairs**

But the fields diverged in the late 19-20th centuries,
astronomy → astrophysics (EM, QM)
statistics → social sciences & industries

Astronomy & the birth of statistics

Hipparchus (4th c. BC): **Average via midrange of observations; error propagation in estimate of length of a year**

Middle Ages: **Chose `best' of discrepant observations, or only observe once**

Brahe (1580s): **Control of random & systemic observational error**

Kepler (1609 & 1619): **Trial-and-error parameter estimation and model selection of nonlinear models for Mars' orbit from unevenly spaced bivariate time series**

Halley (1693): **Foundations of actuarial science**

Legendre & Laplace (c1805): **Cometary orbits via least squares regression**

Gauss (1809): **Normal distribution of errors in celestial mechanics**

Many astronomers (1800s): **Contributed to least squares theory**

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, Today, 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 fitting (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 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

  Babu & Feigelson ADASS 2006

Do we need statistics in astronomy today?

- Are these stars/galaxies/sources an unbiased sample of the vast underlying population?
- When should these objects be divided into 2/3/… classes?
- What is the intrinsic relationship between two properties of a class (especially with confounding variables)?
- Can we answer such questions in the presence of observations with measurement errors & flux limits?

Do we need statistics in astronomy today?

- Are these stars/galaxies/sources an unbiased sample of the vast underlying population? Sampling
- When should these objects be divided into 2/3/… classes? Multivariate classification
- What is the intrinsic relationship between two properties of a class (especially with confounding variables)? Multivariate regression
- Can we answer such questions in the presence of observations with measurement errors & flux limits? Censoring, truncation & measurement errors

- When is a blip in a spectrum, image or datastream a real signal? Statistical inference
- How do we model the vast range of variable objects (extrasolar planets, BH accretion, GRBs, …)? Time series analysis
- How do we model the 2-6-dimensional points representing galaxies in the Universe or photons in a detector? Spatial point processes & image processing
- How do we model continuous structures (CMB fluctuations, interstellar/intergalactic media)? Density estimation, regression

Modern statistics is vast and modern astronomy needs this breadth. We encounters problems in: image analysis, time series analysis, model selection, regression, nonparametrics, spatial point processes, multivariate analysis, survival analysis, …, … Hundreds of modern texts and monographs are available in the mathematics library in these fields.

Astronomers also need broad, reliable statistical software. Historically, commercial stat packages have dominated, and astronomers have not purchased them (largest: SAS).

Recently, the first major public-domain statistical software package has emerged: **R (**http://r-project.org**).** Similar to IDL, **R** (and its 1800+ add-on packages in **CRAN**) provide a huge range of built-in statistical functionalities. **R** tutorials with astronomical examples available at http://astrostatistics.psu.edu

How often do astronomers need statistics?

(a bibliometric measure)

Of ~15,000 refereed papers annually:

1% (now 2%?) have `*statistics'* in title or keywords

5% have `*statistics'* in abstract

10% treat variable objects

5-10% (est) analyze data tables

5-10% (est) fit parametric models

We are making mistakes!

- Six unweighted bivariate **least squares linear regression** are used interchangeably with wrong confidence intervals (Feigelson & Babu 1992)
- The **likelihood ratio test** for comparing two parametric models cannot be applied when a parameter is near zero (Protassov, van Dyk et al. 2002)
- Probabilities from the 1-sample **Kolmogorov-Smirnov test** comparing a univariate dataset to its best-fit model are incorrect (Lilliefors 1969; Babu & Feigelson ADASS XV 2006)
- The **Anderson-Darling test** is often more sensitive than the K-S test, and there is no valid 2-dimensional K-S test (Stephens 1974; Simpson 1951)
- **Power-law models** (= Pareto distribution) should not be fit to binned data, use the MLE on the original events (Crawford et al. 1970)

We are making progress!

- Growth of a vanguard of astronomers using and developing state-of-the-art statistical methods. Particularly evident in Bayesian methodology.
- Growth of cross-disciplinary research collaborations in astrostatistics: CHASC (Harvard, UC Irvine), iNCA (Carnegie-Mellon), Astro/Info (Cornell), CASt (Penn State), Berkeley, Michigan, Duke/SAMSI, Stanford …
- Growth of conference series (SCMA, ADA, PhysStat, SAMSI) and monographs for adanced astrostatistics
- Leadership in statistics is very enthusiastic about astronomy
- Week-long Summer School in U.S., India and Brazil are training ~10% of the world's astronomy graduate students annually in statistical inference and the **R** public-domain code.
- Number of astronomical papers with "Methods: statistical" keyword has recently doubled to ~500/yr

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

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

- 10^9-object catalogs from USNO, 2MASS & SDSS opt/IR surveys
- 10^7- galaxy redshift catalogs from 2dF & SDSS
- 10^{6-7}-source radio/infrared/X-ray catalogs
- Spectral databases: 10^5 SDSS quasars, 10^4 stellar radial velocities, 10^3 Spitzer protoplanetary disks, …, …
- Huge image databases, growing datacubes (EVLA/ALMA, IFUs)
- Planned Large-aperture Synoptic Survey Telescope will generate ~10 PBy video, ~10^{10} object catalogs

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

Powerful statistical tools are needed to derive scientific insights from extracted VO datasets

VOStat provides a Web-based interface to elementary methods in **R**

Some methodological challenges for astrostatistics in the 2010s

- Simultaneous treatment of measurement errors and censoring (esp. multivariate). Progress by Kelly, Loredo, …
- Statistical inference and visualization with very-large-N datasets too large for computer memories
- A user-friendly cookbook for construction of likelihoods & Bayesian computation of astronomical problems
- Links between astrophysical theory and non-Fourier decompositions (e.g. wavelets, ARMA, loess)
- Rich families of time series models to treat aperiodic accretion and explosive phenomena

Some resources for learning & using statistics

All astronomers should be familiar with basic statistics for physical scientists at the level of:

- Data Reduction and Error Analysis for the Physical Sciences, Bevington & Robinson (2003)
- Practical Statistics for Astronomers, Wall & Jenkins (2003)
- Statistical Data Analysis, Cowan (1998)

Useful broad-scope volumes in statistics:

- A First Course in Probability, Ross (8th ed, 2008, undergrad)
- Probability and Statistical Inference, Hogg & Tanis (8th ed, 2009, undergrad)
- Statistical Models, Davison (2003, advanced with R code)
- Principles of Statistical Inference, Cox (2006, discursive)
- Statistical Methods in Experimental Physics, James (2nd ed, 2006)

A selection of many topical volumes (see astrostatistics.psu.edu/Bibliographies):

- Monte Carlo Statistical Methods, Robert & Casella (2004)
- Introduction to Modern Nonparametric Statistics, Higgins (2004)
- Applied Multivariate Statistical Analysis, Johnson & Wichern (2002)
- An R and S-PLUS Companion to Multivariate Analysis, Everitt (2005)
- The Analysis of Time Series: An Introduction, Chatfield (2003)
- A Wavelet Tour of Signal Processing, Mallat & Mallat (1999)
- Applied Measurement Error in Nonlinear Models, Carroll, Ruppert & Stefanski (2006)
- Astronomical Image and Data Analysis, Starck & Murtagh (2006)
- Statistical Analysis of Spatial Point Patterns, Diggle (2001)
- Visual Data Mining: Techniques and Tools for Data Visualization and Mining, Soukup & Davidson (2002)
- Elements of Statistical Learning: Data Mining, Inference, and Prediction, Hastie, Tibshirani & Friedman (2001)
- Survival Analysis: A Self-Learning Text, Kleinbaum & Klein (2005)
- Model Selection and Multimodel Inference, Burnham & Anderson (2002)

A few Bayesian resources

- Bayesian Computation with **R**, Albert (2009)
- Data Analysis: A Bayesian Tutorial, Sivia & Skilling (2006)
- Bayesian Logical Data Analysis for the Physical Sciences, Gregory (2005)
- Bayesian Data Analysis, Gelman, Carlin, Stern & Rubin (2003)
- Probability Theory, Jaynes (2003)
- Statistical decision theory and Bayesian analysis, Berger (1985)

- "The promise of Bayesian inferences for astrophysics", Loredo in Statistical Challenges in Modern Astronomy (1992)
- BIPS:Bayesian Inference for the Physical Sciences, http://www.astro.cornell.edu/staff/loredo.bayes

A vision of the future ….

Astrostatistics in 2029

Dozens of young astronomers have M.S. degrees in statistics and computer science. Science based on petabyte/exabyte datasets use modern methods effectively and thoughtfully. Astronomical papers reference statistics monographs.

Major problems in Bayesian cosmology are resolved and methodologies for heteroscedastic measurement errors are established. Astronomers regularly use hundreds of methods coded in **P**, the successor to **Q** and **R**.

Thirty well-funded cross-disciplinary research groups in astrostatistics and astroinformatics on three continents push frontiers of astrostatistical methodology. ***Statistical Challenges in Modern Astronomy*** and **Astro-Informatics** conferences are held annually.