Statistical challenges in modern astronomy

Eric Feigelson (Astro & Astrophys)
&
Jogesh Babu (Stat)

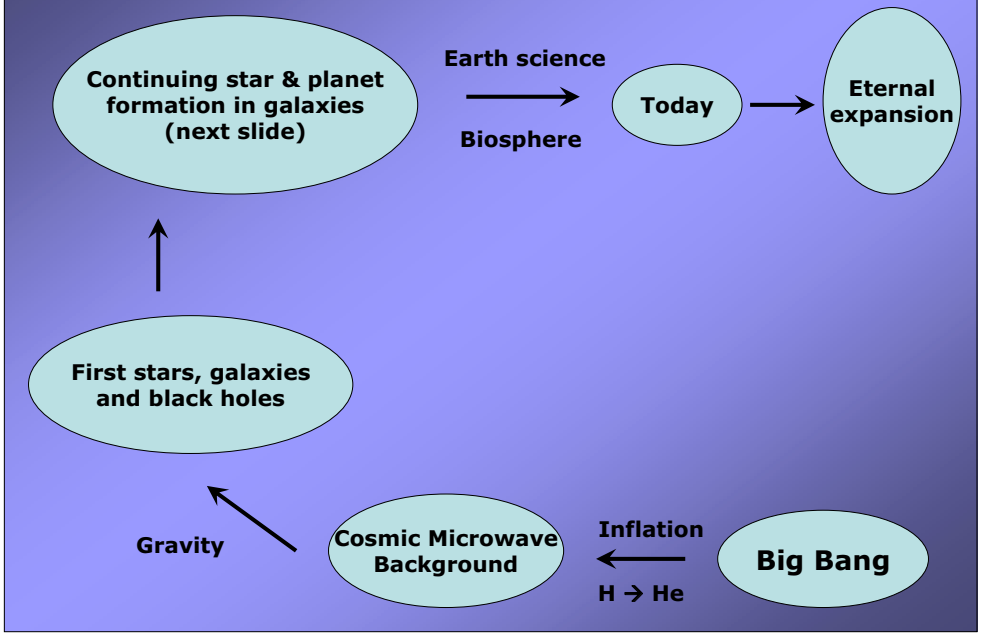
Penn State University

What is astronomy?

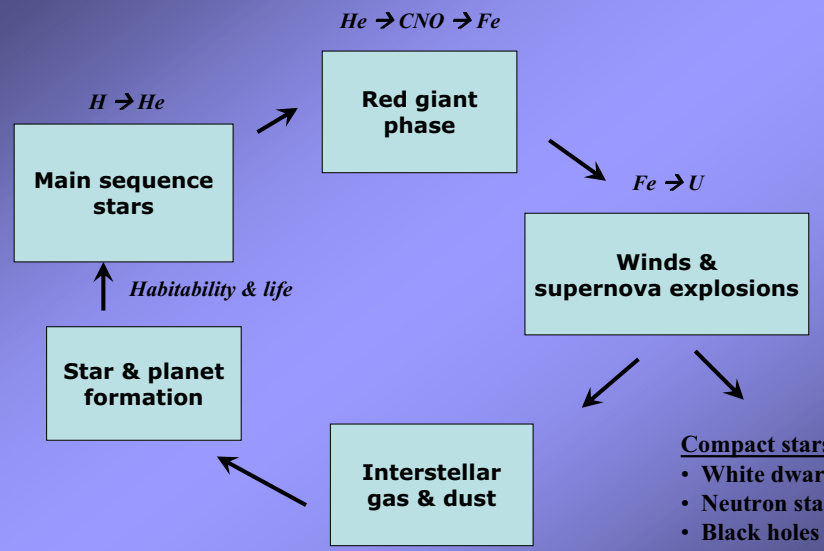
Astronomy (astro = star, nomen= name in Greek) is the observational study of matter beyond Earth -- planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations. The perspective is rooted from our viewpoint on or near Earth using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

1. "... statistics refers to the methodology for the collection, presentation, and analysis of data, and for the uses of such data." (applied stat textbook)

2. "[Statistics is] the study of algorithms for data analysis." (Beran 2003)

3. "A statistical inference carries us from observations to conclusions about the populations sampled." (Cox 1958)

4. "Uncertain knowledge + Knowledge of the amount of uncertainty in it = Usable knowledge" (Rao 1997, Statistics & Truth)

Statistics and science

6. "In statistical inference experimental or observational data are modeled as the observed values of random variables, to provide a framework from which inductive conclusions may be drawn about the mechanism giving rise to the data." (Univ Cambridge text, 2005)

7. "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." (Gregory, astronomer, 2005)

8. "[Statistical models] … can guide our thinking, lead us to propose courses of action and so on, and if used sensible, and with an open mind, and if checked frequently with reality, might help us learn something that is true. ... What we do works (when it does) because it can be seen to work, not because it is based on true or even good models of reality." (Terry Speed, SCMA conference

9. "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, Principles of Statistical Inference, 2006)

10. "Statistics has become the primary mode of quantitative thinking in literally dozens of fields, from economics to biomedical research. The statistical tide continues to roll in, now lapping at the previously unreachable shores of the hard sciences. ... Yes, confidence intervals apply as well to neutrino masses as to disease rates, and raise the same interpretive questions, too. (Bradley Efron, ASA Presidential Address, 2004)

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

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

How often do astronomers need statistics?

(a bibliometric measure)

Of ~15,000 refereed papers annually:

1% 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

The state of astrostatistics today

The typical astronomical study uses:

- 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 often misused:

- Six unweighted bivariate least squares fits are used interchangeably in H_o studies with wrong confidence intervals
  Feigelson & Babu ApJ 1992
- Likelihood ratio test (F test) usage typically inconsistent with asymptotic statistical theory
  Protassov et al. ApJ 2002
- K–S g.o.f. probabilities are inapplicable when the model is derived from the data
  Babu & Feigelson ADASS 2006

A new imperative: Virtual Observatory

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

- **10^9-object catalogs from USNO, 2MASS & SDSS opt/IR surveys**
- **10^6- galaxy redshift catalogs from 2dF & SDSS**
- **10^5-source radio/infrared/X-ray catalogs**
- **10^{3-4}-samples of well-characterized stars & galaxies with dozens of measured properties**
- **Many on-line collections of 10^2-10^6 images & spectra**
- **Planned Large-aperture Synoptic Survey Telescope will generate ~10 Pby**

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
(NSF FRG involving PSU/CMU/Caltech)

But astrostatistics is an emerging discipline

- **We organize cross-disciplinary conferences at Penn State *Statistical Challenges in Modern Astronomy (1991/96, 2001/06)***
- **Fionn Murtagh & Jean-Luc Starck run methodological meetings & write monographs**
- **We organize Summer Schools at Penn State and astrostatistics workshops at SAMSI**
- **Powerful astro-stat collaborations appearing in the 1990s:**
  - **Harvard/Smithsonian (David van Dyk, Chandra scientists, students)**
  - **CMU/Pitt = PICA (Larry Wasserman, Chris Genovese, …)**
  - **NASA-ARC/Stanford (Jeffrey Scargle)**
  - **Penn State CASt (Jogesh Babu, Eric Feigelson)**
  - **Efron/Petrosian, Berger/Jeffreys/Loredo/Connors, Stark/GONG, …**

Some methodological challenges for astrostatistics in the 2000s

- Simultaneous treatment of measurement errors and censoring (esp. multivariate)
- 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 wavelet coefficients (spatial & temporal)
- Rich families of time series models to treat accretion and explosive phenomena

Structural challenges for astrostatistics

Cross-training of astronomers & statisticians

- New curriculum, summer schools & textbook
- Effective statistical consulting

Enthusiasm for astro-stat collaborative research

- Recognition within communities & agencies
- More funding (astrostat gets <0.1% of astro+stat)

Implementation software

- StatCodes Web metasite (www.astro.psu.edu/statcodes)
- Standardized in R & VOStat (www.r-project.org)

Inreach & outreach

- PSU's Center for Astrostatistics