Astrostatistics and

High Energy Astrophysics

Eric Feigelson
Center for Astrostatistics
Penn State University

HEAD 2008

What is astrostatistics?

What is astronomy?

The properties of planets, stars, galaxies and the Universe, and the processes that govern them

What is statistics?

- "The first task of a statistician is cross-examination of data" (R. A. Fisher)
- "[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)
- "Some statistical models are helpful in a given context, and some are not" (T. Speed, addressing astronomers)
- "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 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)

My conclusion:
The application of statistics to high-energy astronomical data is not a straightforward, mechanical enterprise. It requires careful statement of the problem, model formulation, choice of statistical method(s), and judicious evaluation of the result.

We are making mistakes!

- 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 should not be fit to binned data, use the MLE on the original events (Crawford et al. 1970)

We use a unnecessarily narrow suite of statistical methods

Modern statistics is vast. HEA encounters problems in: image analysis, time series analysis, model selection, regression, nonparametrics, spatial point processes, multivariate analysis, survival analysis, …, … Dozens of monographs are published each year in these fields.

The software situation is much improved: **R** has emerged as the premier public-domain statistical software package. Similar to IDL in style, but with a huge range of built-in statistical functionalities.

See http://r-project.org and tutorials at http://astrostatistics.psu.edu

We are making progress!

- Growth of research collaborations in astrostatistics: California-Harvard Astrostatistics Collaboration (van Dyk et al) specifically oriented towards HEA. Also groups at CMU, Berkeley, Michigan Penn State, Cornell, SAMSI.

- Growth of conference series (SCMA, ADA, PhysStat, SAMSI) and monographs (Starck/Murtagh, Gregory, Lupton, Wall/Jenkins) for advanced statistical treatment of astronomical data

- Week-long Summer School in Statistics for Astronomers held at Penn State since 2005. In steady state, we are training ~10% of world's astronomy graduate students.

Some contemporary issues

- Treatment of measurement errors. See Bayesian approach to regression by Kelly (ApJ 2007)
- Hardness ratios at low count rates. Surprisingly tricky! See review by Brown et al. (Stat Sci 2001) and Bayesian solution by Park et al. (ApJ 2006)
- Upper limits in Poisson background. Surprisingly tricky! See review by Cowan (SCMA IV 2007)
- Increased use of numerical bootstrap and MCMC numerical confidence intervals (Babu 1984; Ptak, this session)