Towards Improved Statistical Practices in Astronomical Studies

Eric Feigelson

with G. Jogesh Babu

Center for Astrostatistics, Penn State University

Statistical Scientific Editor, AAS Journals

2nd East Asian Workshops on Astrostatistics
Nanjing & Guiyang
July 2018

The underlying situation

Astronomers are well-trained in the mathematics underlying physics, but not in applied fields associated with statistical methodology.

Consequently, many astronomers use a narrow suite of familiar statistical methods that are often non-optimal, and sometimes incorrectly applied, for a wide range of data and science analysis challenges.

This talk highlights some common problems in recent astronomical studies, and encourages use of improved methodology.

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:

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities ← *easiest step with R*
- judicious scientific evaluation of the results

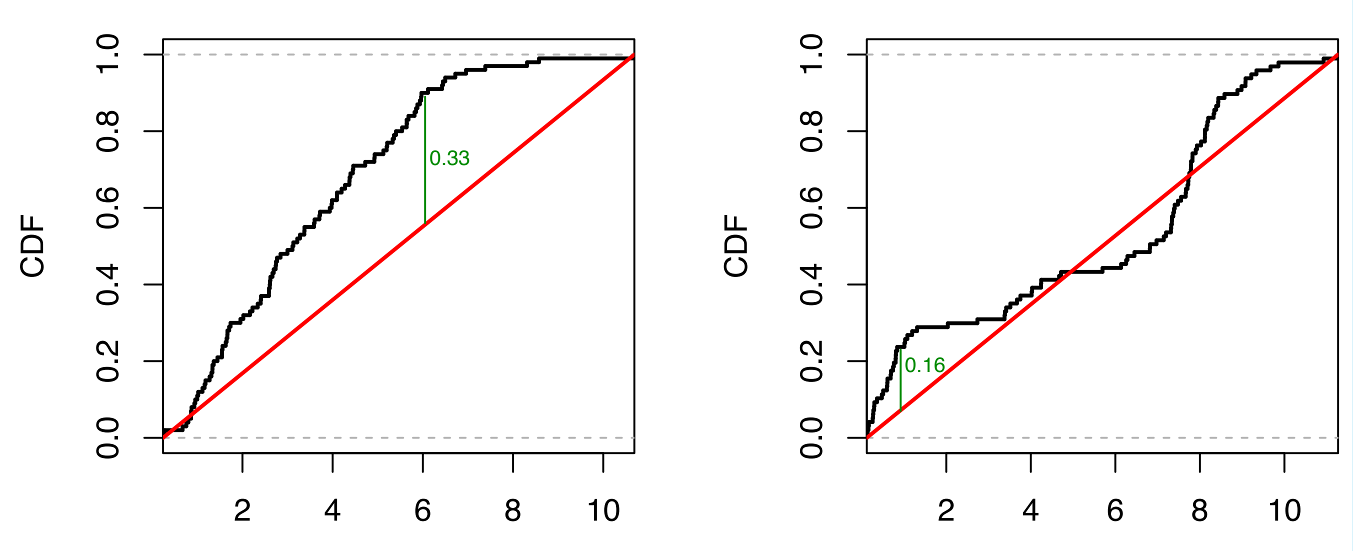
Astronomers often do not adequately pursue each step

Misuse of the Kolmogorov-Smirnov test

The KS test is used in ~500 astronomical papers/yr, but often incorrectly or with less efficiency than an alternative test. Three problems are identified:

1. The KS statistic efficiently detects differences in global shapes, but not small scale effects or differences near the tails. The Anderson-Darling statistic (tail-weighted Cramer-von Mises statistic) is more sensitive.

$$A^2_{AD,n} = n\sum_{i=1}^{n} \frac{[i/n - F_0(X_i)]^2}{F_0(X_i)(1 - F_0(X_i)}.$$



Kolmogorov-Smirnov test (continued)

2. The 1-sample KS test (data vs. model comparison) is distribution-free only when the model is not derived from the dataset. In this case, probabilities must be calculated for each problem using bootstrap resampling.

3. The KS test is distribution-free only in 1-dimension. Multi-dimensional KS tests are based on arbitrary ordering; probabilities can be obtained from bootstrap resampling.

See the viral page
Beware the Kolmogorov-Smirnov test!
at http://asaip.psu.edu

Overuse of binned statistics

- Histograms are good for visualization, poor for inference
  - Arbitrary bin width, binning algorithm, zero point
  - Loss of information within bin
  - √N errors not accurate for sparse counts

  Kernel density estimation (e.g. Gaussian convolution) and non-parametric local regressions with confidence bands are recommended for density estimation (smoothing)

- Binned estimators can be replaced by unbiased, unbinned maximum likelihood estimators. Chi-square tests & regressions based on arbitrarily binned data give unreliable probabilities because the degrees of freedom are not known.

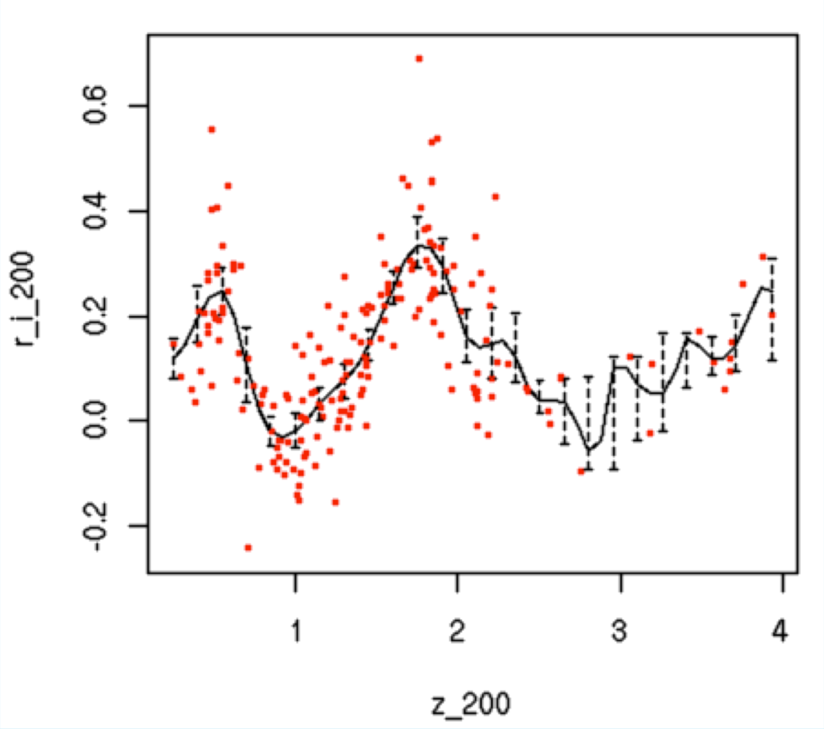
- Inference from histograms are particularly inaccurate for asymmetrical distributions;
  e.g. slope estimates of power law (Pareto) distributions (use MLE instead)

- Poor use of 2-sample comparisons when continuous data is arbitrarily split into
  subsamples for continuous data (use nonparametric correlation measures)

Local regressions

Statisticians have recently developed *local regression* models that give heteroscedastic confidence intervals from spline-type regressions, often using bootstrap resampling in windows. In 2-3 dimensions, geostatistics have developed *kriging* regression models that give maps and variograms from (un)evenly sampled data points. Kriging is synonymous with Gaussian Processes regression.

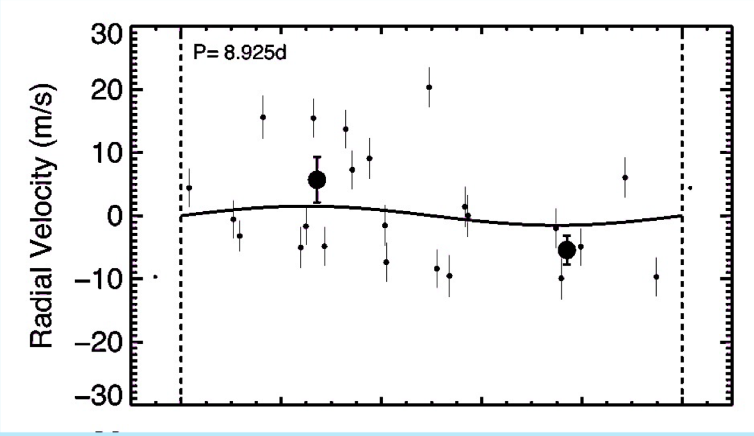
Nadaraya-Watson local regression estimator with bootstrap confidence intervals for a small sample of Sloan quasars



Problems with Regression I
Improper use of minimum χ^2 fitting

A χ^2-like statistic for binned data, defined as the sum of squared residuals divided by the square of heteroscedastic measurement errors, is often minimized to obtain a best fit model. Problems with this method include:

- Parameter estimation and parameter confidence intervals may be biased (even completely incorrect) if the model is misspecified (minimum χ^2>1.0 or <1.0). This occurs when the measurement errors are incorrectly specified, or are not fully responsible for the variance of the response variable.



- The statistic may not be χ^2-distributed or the degrees of freedom may be ill-determined. The theorems underlying the χ^2 test apply only to restricted situations; e.g. the bins must be established before the data are acquired begins (multinomial experiment), not chosen later.

Alternatives to minimum χ^2 fitting include: unweighted least squares regression; unbinned maximum likelihood estimation; and Bayesian inference. See B.C. Kelly (ApJ 2007) for a flexible likelihood framework with heteroscedastic measurement errors as a component of the variance.

Problems with Regression II
Inadequate residual analysis

Detailed study of residuals between data and a best-fit model gives critical insight into the quality of the fit:

- How much of the original variance is reduced by the model? Examine the *adjusted* R^2 or *Mallows* C_p.

- Are the residuals autocorrelated indicating that structure is present outside of the model? Try the autocorrelation function & Durbin-Watson statistic that have known distributions for Gaussian white noise.

- Are the residuals normally distributed? If not consider *quantile regression* to study behavior in more detail.

- Are outliers present? Use *standardized residuals* and *Cook's distance* to quantify the effects of individual points on the model. Use *robust regression* techniques to reduce effects of outliers, if necessary.

Problems with Regression III
Inadequate model selection & goodness-of-fit

Consider carefully whether the model addresses the scientific question and adequately fits the data.

- Is there a scientific basis for choosing the response variable? If not, try *symmetric regression models*.

- Use the Anderson-Darling test to evaluate goodness-of-fit. Validate the model and parameter confidence intervals using *cross-validation* and *bootstrap* techniques.

- Consider elaborating or simplifying the model with more or fewer parameters. Use penalized measures (adjusted R^2, Akaike Information Criterion, Bayesian Information Criterion) for model selection.

Problems with Regression IV
Other issues

- Regression results change when arbitrary variable transformations (e.g. log) are made. Nonparametric tests should precede regression analysis.

- Astronomers tend to view intrinsically multivariate regression problems as a sequence of bivariate problems. This is unnecessary and restrictive: most regression methods are intrinsically multivariate.

- Regressions for variables that are not independent (e.g., B-V vs. V-I diagrams) should be performed with great caution.

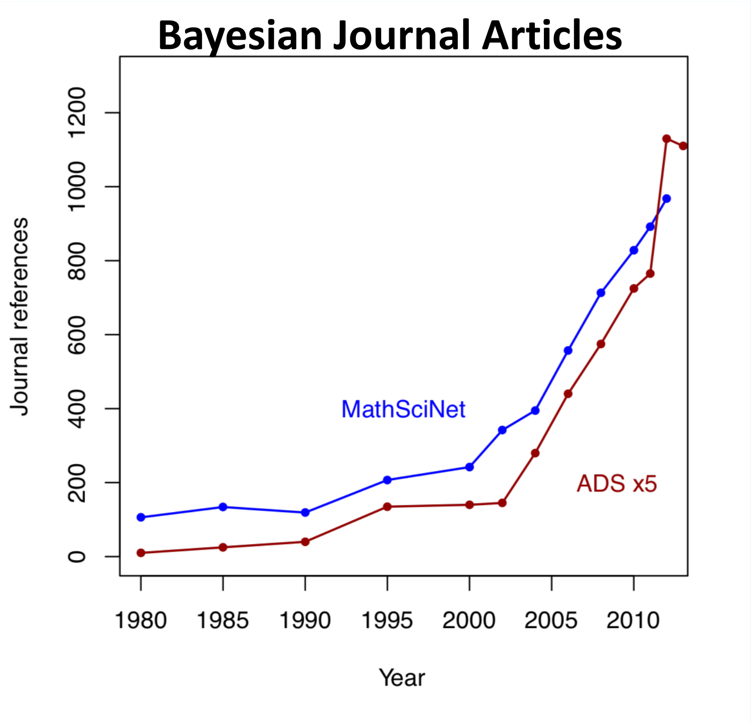
- Problems with Poisson-distributed response variables should use *Poisson regression*.

- Problems with binary (Yes/No) response variables should use *contingency tables* and *logistic regression*.

Overuse of Bayesian inference

Rapid rise in Bayesian analyses since ~2003 in both the astronomical & statistical communities. But some applications might be reconsidered:

When uninformative flat priors and Bayesian model averaging is used, the mean of the likelihood function averaged over (often arbitrarily) chosen parameter ranges. Scientifically uninteresting structure in the likelihood will affect the result.



When the model is used, the result is usually identically to maximum likelihood estimation (MLE) that has dominated statistical model fitting since Fisher (1922).

When relevant scientific prior information is available, Bayesian inference is recommended. It is a scientific, not a statistical, decision whether to weight the data's likelihood with prior information.

Bayesian inference is also recommended for marginalization of nuisance variables and model comparison.

Suggested standards for Bayesian presentations

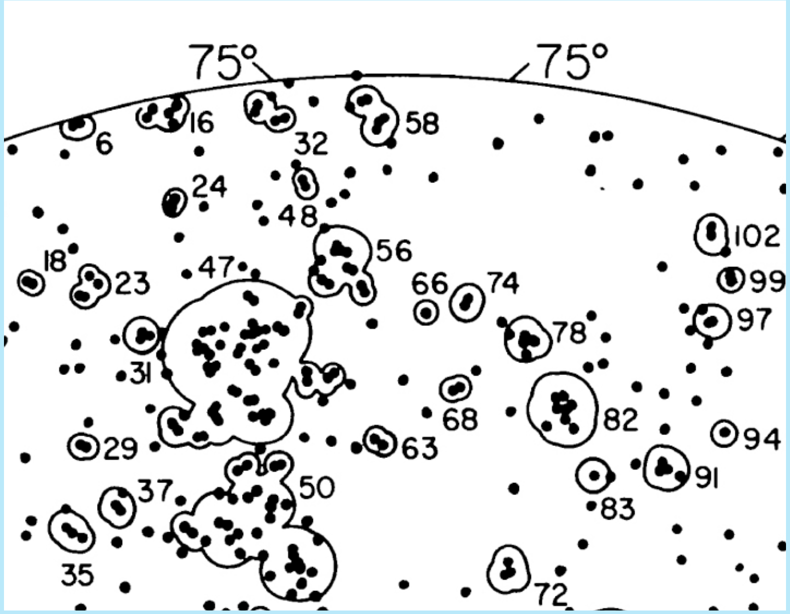
- Give explicit equations for Bayes' Theorem, likelihood, and priors
- Discuss sensitivity of science results to different reasonable prior distributions, including unity (i.e. maximum likelihood estimation)
- Discuss accuracy and convergence of numerical algorithm (e.g. MCMC). Note that in simple cases (unimodal posterior in low dimensions), more efficient optimization algorithms can be used (e.g. EM Algorithm).

Multivariate clustering I
Overuse of `friends-of-friends' algorithm

The FoF (percolation) algorithm is the single linkage agglomerative clustering algorithm (Florek et al. 1951).

Extensive tests show that single linkage tends to give spurious `chaining' of clusters in many situations. Average linkage or Ward's criterion is recommended instead.

FoF may be advised when elongated anisotropic clusters are sought (e.g. filamentary galaxy clustering) but should not be used for general problems of unsupervised multivariate clustering.



Turner & Gott 1976

Useful reference
Everitt et al. *Cluster Analysis*, 5th ed. Wiley, 2011

Multivariate clustering II
Arbitrary choice of cluster boundaries

Astronomers often construct boundaries between classes by eye based on low-dimensional projections. *Decision trees* with boundaries parallel to variable axes are an acceptable clustering technique, but should be performed using objective criteria for optimizing splits, pruning the tree, and testing tree validity with bootstrap resampling.

This is the method of *CART* (Classification and Regression Trees) developed during the 1970-2000s by Leo Breiman and others, culminating in the *Random Forest* technique.

Difficulties with frequency domain time series analylsis

- Use standard techniques of (multi)tapering and smoothing to improve periodogram signal-to-noise

- Beware stating simple false alarm probabilities for periodograms: Fourier, Lomb-Scargle, Box Least Squares, etc. Use simulations to validate peaks and alias structures.

- Beware interpreting peaks in periodograms as periodic behavior. Often aperiodic autocorrelated behaviors produce spurious peaks in finite data sets. Compare with ARMA-type time domain modeling.

Two concluding thoughts：Practical
Avoid reinventing methodology

Astronomers repeatedly begin reinventing procedures that are already well established with theorems, software, and experience.

- If the method reference is an astronomy or physics article, reference instead a statistics textbook or an important seminal paper. Start with Wikipedia, Google, R/CRAN, StackOverflow. For serious study (e.g. PhD dissertation), read topical statistics textbooks.

- Read and learn the history and environment of the methods you want to use. What is the past experience in other fields? What are the alternatives and how are they judged by expert statisticians?

Two concluding thoughts：Conceptual

Astronomy is not statistics!

Astronomical discovery and astrophysical understanding often relies on interpreting quantitative measurements of planets, stars, galaxies and the Universe. In these cases, statistical analysis and evaluation can be very useful. It is silly to ignore the sophistication of modern methodology; this can lead to unwarranted reports of scientific results, and to missed opportunities to infer valid results.

But astronomy and astrophysics also relies on non-quantitative argumentation, often based on insights into how the physical Universe operates. Cause-and-effect sequences are difficult to discern statistically but can be crucial for clarifying alternative explanations. Even in observational astronomy, weaving together strands of unconvincing evidence into scenarios has often led to insights later validated by further study.

Conclusion

While a vanguard of astronomers use and develop advanced methodologies for specific applications, many studies are unnecessarily restricted to a narrow suite of familiar methods.

Astronomers need to become more informed and more involved in statistical methodology, for both data analysis and for science analysis.

Areas of common weakness of statistical analyses in astronomical studies can be identified (this talk). Improvement is often not difficult. Highly capable free software, such as R/CRAN and Python, can be effective in bringing new methodology to advance our science.