Common Statistical Mistakes in the Astronomical Literature

Eric Feigelson

Penn State University

Harvard-Smithsonian Center for Astrophysics 2014

The problem

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.

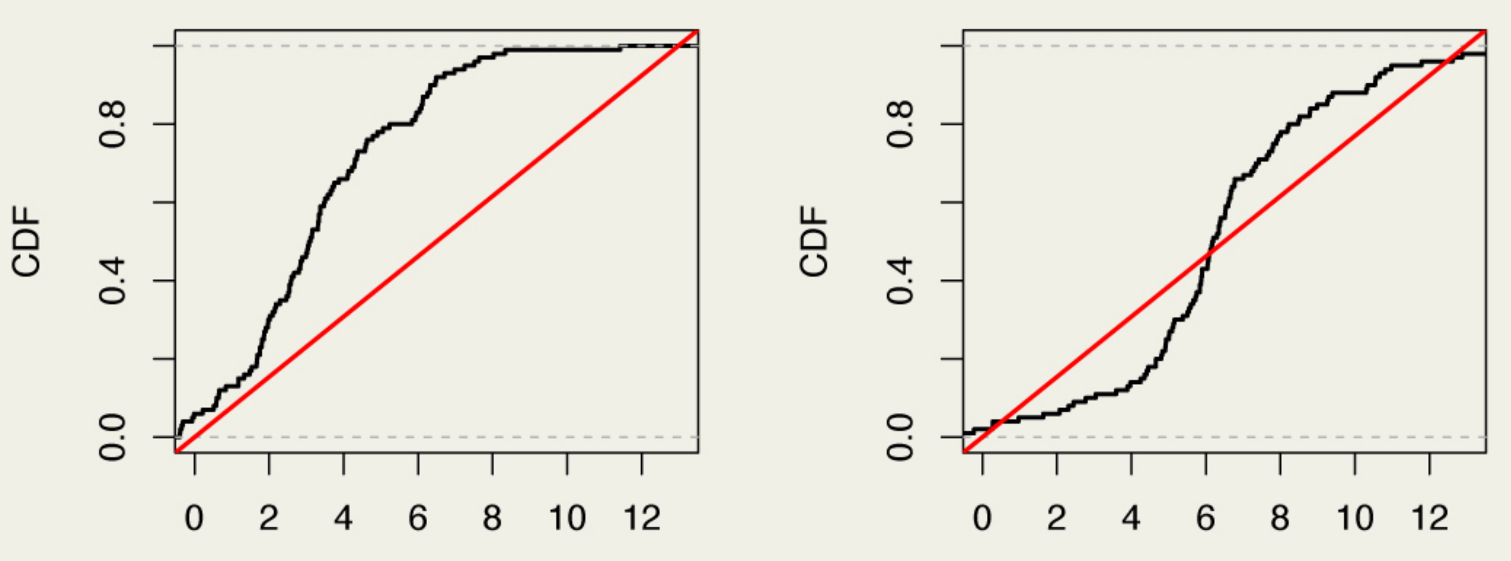
Misuse of the Kolmogorov-Smirnov test

The KS test is used in ~500 astronomical papers/yr, but often incorrectly or with less efficiency than an alternative test

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.

$$M_{KS} = \sqrt{n}\max_{x} |\widehat{F}_n(x) - F_0(x)|,$$

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



Kolmogorov-Smirnov test (continued)

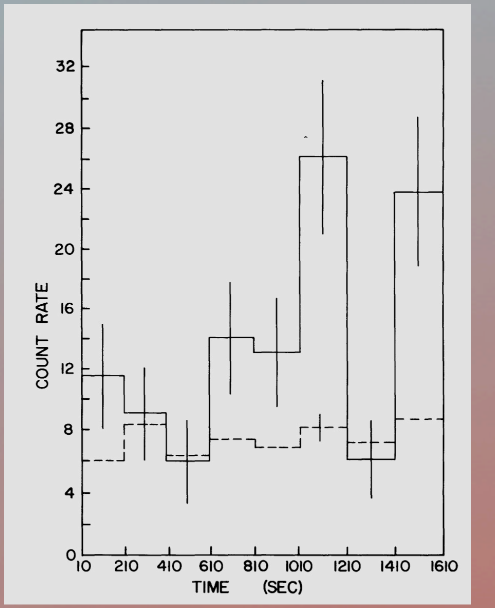
The 1-sample KS test (data vs. model comparison) is distribution-free only when the model is not derived from the dataset. In this case, probabilities must be calculated for each problem using bootstrap resampling.

The KS test is distribution-free (i.e. probabilities can be used for hypothesis testing) only in 1-dimension. Multi-dimensional KS tests are based on arbitrary ordering; probabilities 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 (e.g. regression using midpoint or centroid?)
  - √N errors not accurate for sparse counts
  - Kernel density estimation (e.g. Gaussian convolution) is recommended for nonparametric smoothing
  - Anderson-Darling test recommended for goodness-of-fit



- Binned χ^2-type regression estimators can be replaced by unbiased, unbinned maximum likelihood estimators

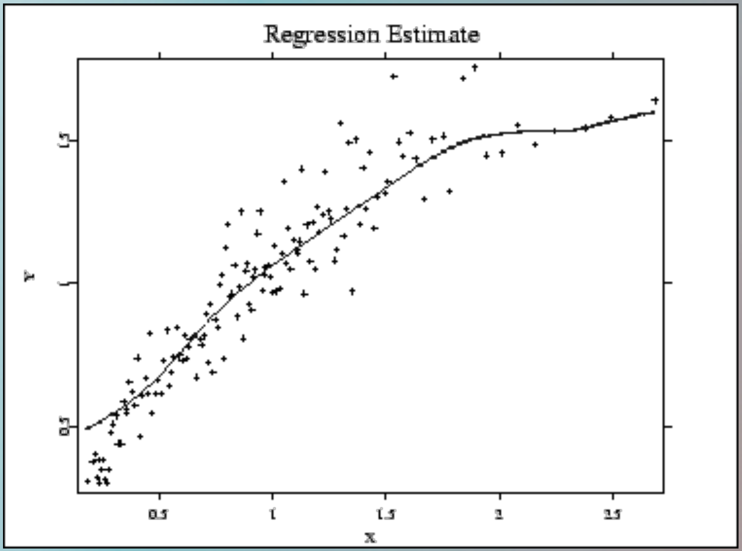
- Inference from histograms particularly inaccurate for asymmetrical distributions; e.g. power law (Pareto)

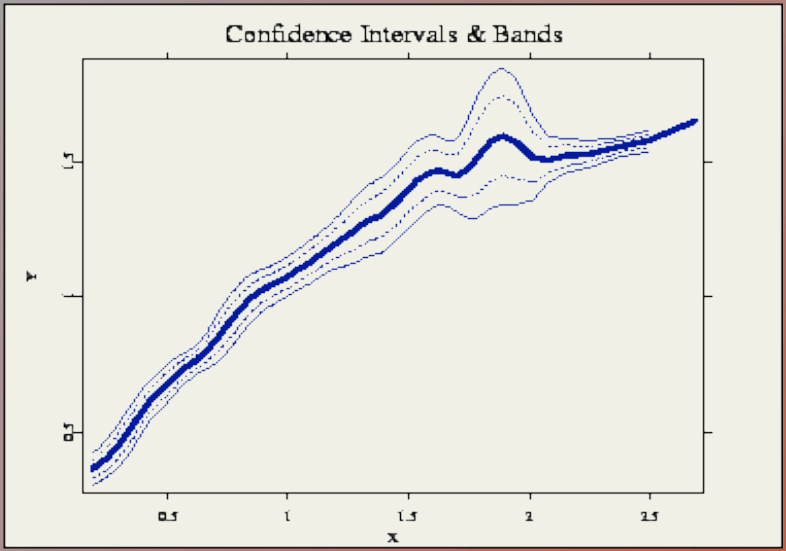
- Misuse of 2-sample comparisons with arbitrary split of subsamples for continuous data (use nonparametric correlation measures)

Binned statistics (continued)

Astronomers often use histograms because they give plausible √N-type confidence intervals along the distribution.

However, 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 closely related to modeling with Gaussian Processes.



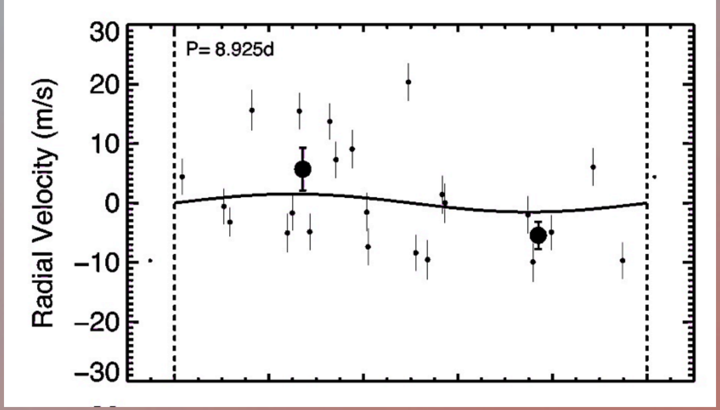


K. Takezawa, *Introduction to Nonparametric Regression (2005)*

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

- A χ^2-like statistic, defined as the sum of squared residuals divided by the square of the measurement errors, is often minimized to obtain a best fit model. This statistic applies only when restrictive assumptions apply: the errors are Gaussian and account for all of the scatter of the data about the model.

- Parameter estimation and parameter confidence intervals may be biased (even completely incorrect) if the model is misspecified (i.e., 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. 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. An important study is B.C. Kelly (ApJ 2007) who presents a likelihood that includes heteroscedastic measurement errors as a component of the variance, and that can also treat censoring and truncation.

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?

- 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. 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.

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

- Regressions with binary (Yes/No) or categorical response variables should use *logistic regression*.

Improper use of the likelihood ratio test

The LRT is a classical likelihood-based test to compare the relative validity of two models for a single dataset. Particularly used for ***nested models,*** where the null model is a special case (fewer parameters) of the alternative model. The log-LRT statistic, *assuming the null model is true,* is asymptotically distributed as the chi-squared statistic. Theory is based on Neyman-Pearson lemma (1933) and Wilks' theorem (1938) within the theory of maximum likelihood estimation.

The principal problem is that astronomers frequently use the LRT to test the existence or non-existence of a faint feature (e.g. source or spectral line above background). The LRT is not (even asymptotically) chi-squared distributed near a boundary of the parameter space. Alternative procedures can be used (e.g. Bayes factors).

Statistics, handle with care: Detecting multiple model components with the Likelihood Ratio Test

Protassov, van Dyk, Connors, Kashyap, Siemiginowska [ICHASC]

Astrophys. J. 571, 545-559, 2002

Over-reliance on 3σ

Many astronomers have a nearly-religious belief that a 3σ (P=99.7%) criterion decides whether a scientific result or discovery is reliable.

- The choice of P=99.7% is arbitrary (Wall QJRAS 1979, Shewart 1939) Social science, biomedical, etc communities choose P=0.95 (Fisher 1925) Particle physicists choose 5σ significance (e.g. Higgs boson 2012). Manufacturing business leaders (e.g. Motorola, General Electric, 1990s--) and statistical process control experts adopt Six Sigma doctrine.
- The quantitative relationship between the standard deviation σ and the probability P=99.7% relies on the assumption that the Gaussian distribution *accurately*, out to the tails of the distribution, applies to the situation under study. There is often no basis for assuming this.
- Bayesian hypothesis testing has non-quantitative valuations of model testing.

No magical difference between a 2.9σ and a 3.1σ result

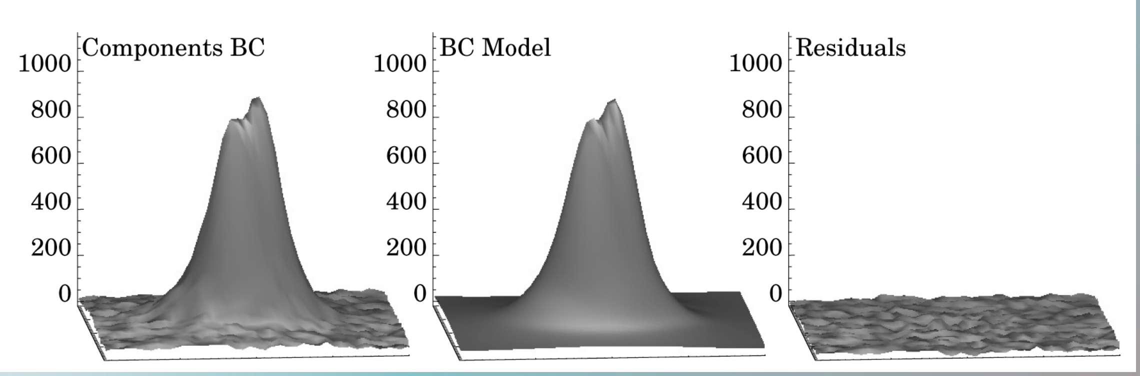
Overuse of Bayesian inference

When uninformative flat priors are used, Bayes' Theorem gives the mean of the likelihood function averaged over (often arbitrarily) chosen parameter ranges. Scientifically uninteresting structure in the likelihood will affect the result.

When no ancillary information is available and the likelihood structure is simple, statisticians recommend that scientists choose the mode (i.e. the `best fit' model) of the likelihood function. This is maximum likelihood estimation (MLE) that has dominated statistical model fitting since Fisher (1922).

When scientific prior information is available, or scientifically important multimodal structure is present in the likelihood and posterior distributions, then Bayesian inference is recommended.

Bayesian inference (continued)

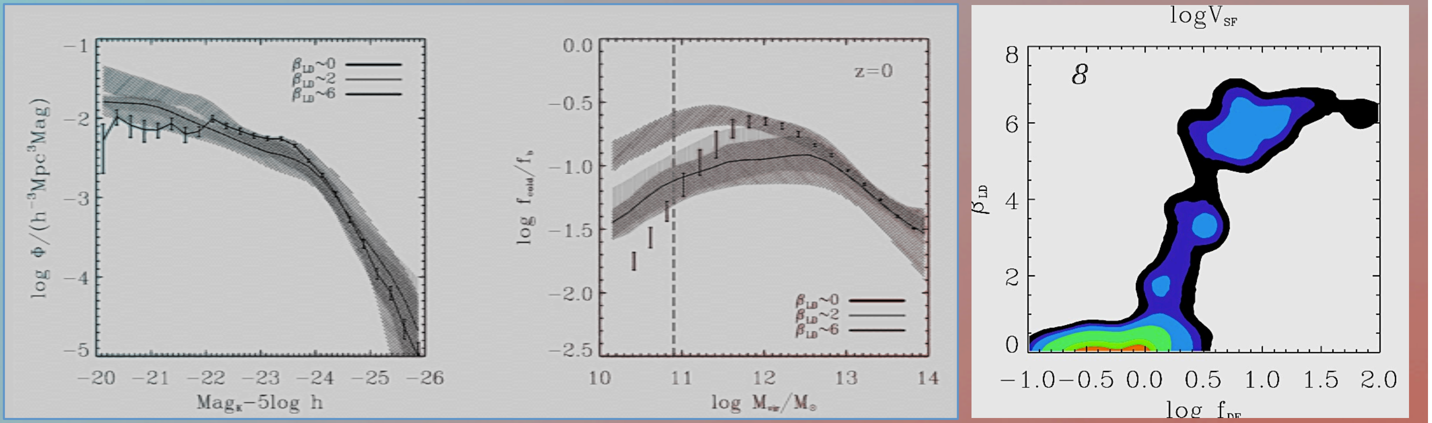


A normal mixture model fitted by minimizing χ^2 with model selection using the BIC.

Radigan et al. 2013

A complex galaxy star formation model fitted to diverse datasets with complex posteriors

Lu et al. 2013



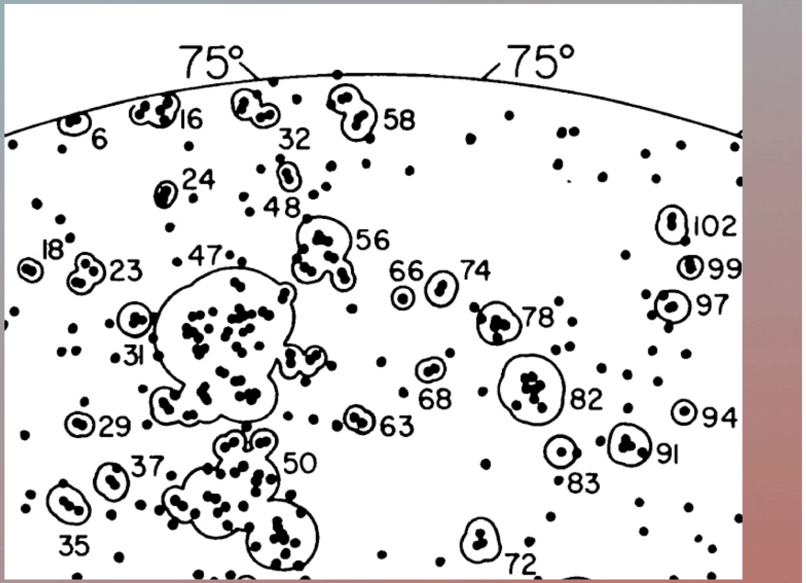
Computation of MLEs via the Expectation-Maximization Algorithm (EM Algorithm) if often simpler than computation of Bayesian model averages using Markov chain Monte Carlo (MCMC) techniques.

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.

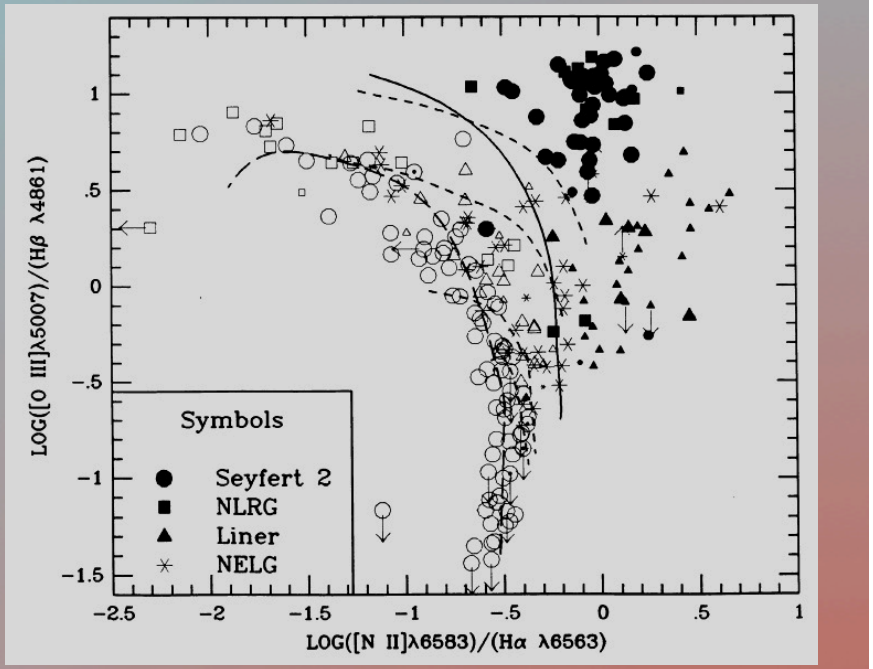


Turner & Gott 1976

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

Multivariate clustering II
Arbitrary choice of cluster boundaries

Boundaries between classes are often constructed by eye based on low-dimensional projections. These are *decision trees* when boundaries are parallel to variable axes.



Veilleux & Osterbrock 1987

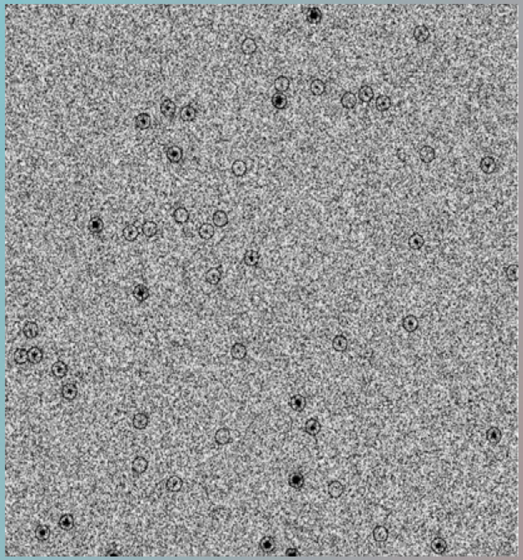
Formal methods for constructing unsupervised, deterministic decision trees were developed during the 1970-2000s by Leo Breiman and others:

- Classification and Regression Trees (CART) for tree construction & pruning
- Boosting & bootstrap aggregation (bagging) for tree improvement
- Random Forests for tree validation

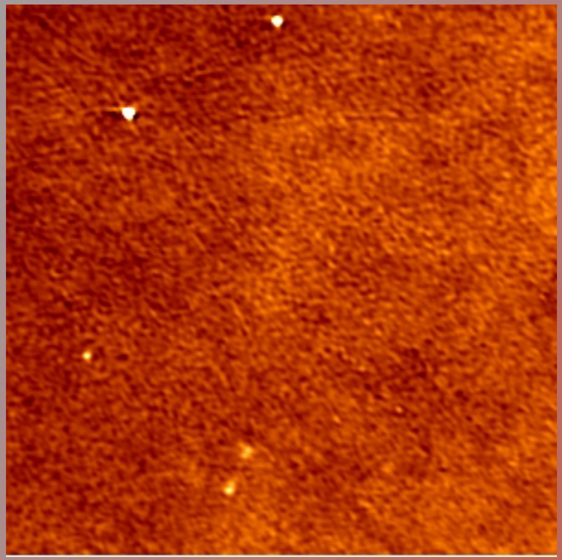
Poor choice of thresholds for multiple hypothesis testing

In faint source detection of large, noisy astronomical images or datacubes, tests of the source existence must be made a myriad times. Traditional ***p***-value approaches (e.g., `3 sigma' or P=0.003) designed to control false negatives (Type I errors) will produce many false positive detections (Type II errors).

An optical image with sources



A channel of a radio datacube



Multiple testing thresholds (continued)

A traditional solution is to control familywise false positive rates applied to the entire multiple testing effort (Bonferroni FWER). However this leads to very conservative thresholds and can miss many true faint sources (too many Type II errors).

A newer approach with widespread interest among statisticians is the distribution-free *False Discovery Rate (FDR)* of Benjamini & Hochberg (1995). Here the scientist chooses two criteria: the significance level of detection, and the desired ratio of Type I to Type II errors (false positives / false negatives). FDR can be very effective.

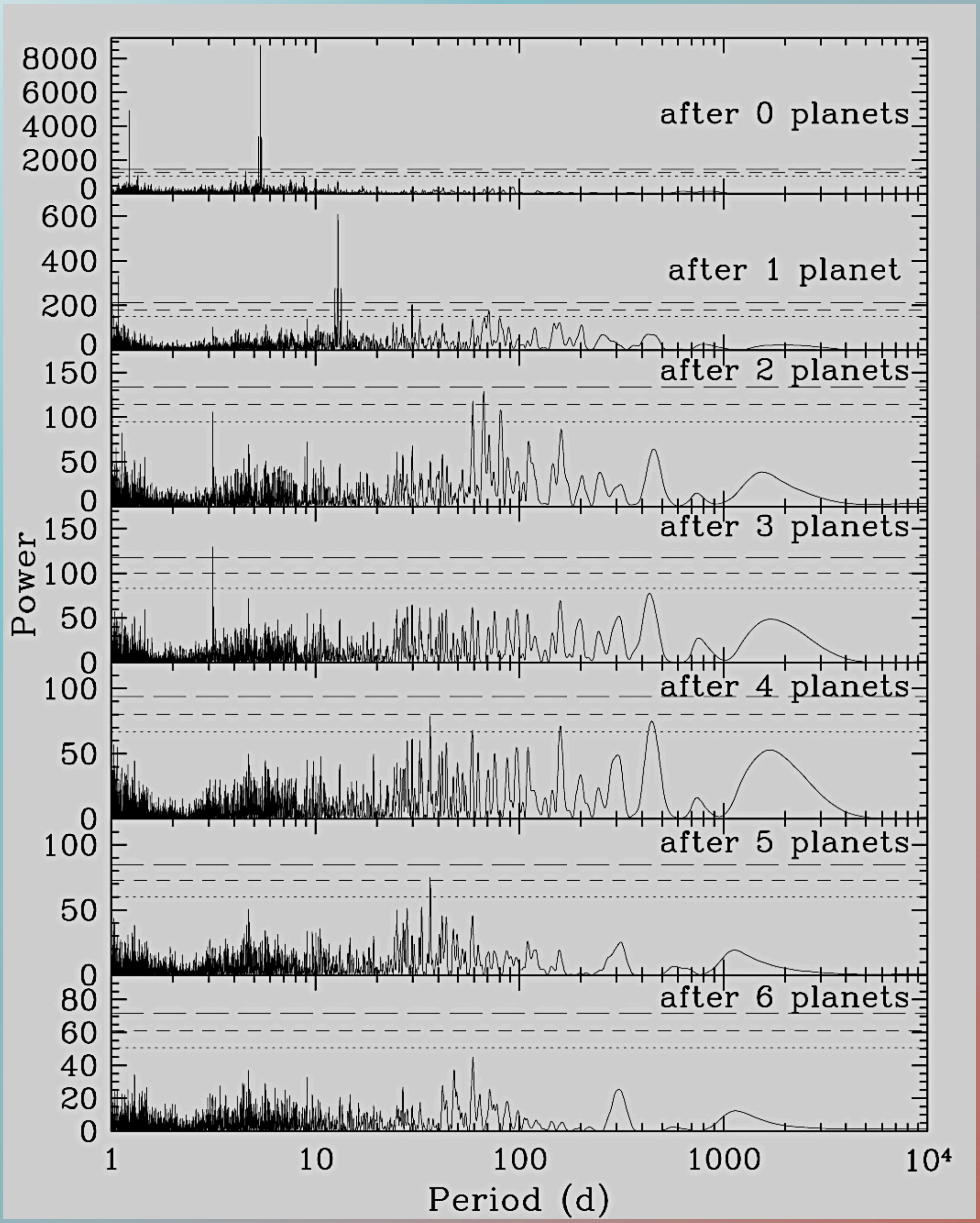
FDR control can give more faint source detection power than FWER but fewer false positives than a threshold uncorrected for the number of tests. Variants are available; e.g. estimating data-dependent optimal thresholds, and treating cases where the tests are not independent (e.g. pixel-level tests with multi-pixel point spread functions).

Uncertain use of Lomb-Scargle periodograms

Generalization of Fourier periodogram for unevenly-spaced data, equivalent to least-squares fitting of sine wave to time series (Scargle 1982). Often performs well, but has limitations:

- The method has an error in its derivation
- Values at high frequency may be biased
- Significance of spectral peak can not be reliable estimated using exponential distribution. Alternative estimates of peak significance has been suggested, but community consensus has not emerged.

In realistic cases, permutation sampling and forward modeling of spectral behavior using the actual observation times are necessary.



GJ 581
Vogt et al. 2010

Other issues

- Improve reporting of hypothesis test probabilities:
  - Insignificant values like P=81% should be reported as P>0.05
  - Significant values like P=2x10^{-13} should be reported as P<<0.0001
  - Avoid high precision like P=0.01492

- Improve methodology references: statistical textbooks rather than astronomical papers
  - Dozens of modern focused texts give authoritative presentations
  - Suggesting books in Wikipedia and in text *Modern Statistical Methods for Astronomy with R Applications* (Feigelson & Babu, 2012)

- Publish code appendices on statistical methods (R, MatLab, Python, etc) to improve reproducibility of scientific results
  - Major journals (e.g. Nature Publ. Group) are raising standards on reproducibility including improved statistical methods description & validation.

Conclusion

While a vanguard of astronomers use and develop advanced methodologies for specific applications, most studies utilize 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 (e.g. regression with model selection and validation, clustering and classification).

Areas of common weakness of statistical analyses in astronomical studies can be identified. Improvement is often not difficult. Highly capable software such as R can be effective in bringing new methodology to bear on astronomical problems.