Six Maxims of Statistical Acumen for Astronomical Data Analysis

Hyungsuk (Tak) Tak

Department of Statistics
Department of Astronomy & Astrophysics
Institute for Computational & Data Sciences
The Pennsylvania State University

Sep 4, 2024

Joint work with Yang Chen (Univ. Michigan), Vinay Kashyap (Harvard-Smithsonian CfA), Kaisey Mandel (Univ. Cambridge), Xiao-Li Meng (Harvard), Aneta Siemiginowska (Harvard-Smithsonian CfA), and David van Dyk (Imperial College London).

Background

The Statistical Editor of the American Astronomical Society contacted Jogesh Babu, David van Dyk, and me on **Jan 21, 2019**, asking us to write an article about standard guidelines for more principled data analyses in the AAS publications.

What are unique features of the data in astronomy?

1. Data are not obtained from designed experiments.
2. Calibration of instruments is crucial in observation process.
3. Sparsity is inevitable even with big data.
4. Astronomical objects evolve on different time scales.
5. Measurement error is heteroscedastic.
6. Etc.

Standard statistical methods, like linear reg., necessitate more care.

Six Maxims

Six Maxims in the sprint of George Box's famous aphorism,

"All models are wrong, but some are useful."

1. All data have stories, but some are mistold.
2. All assumptions are meant to be helpful, but some can be harmful.
3. All prior distributions are informative, even those that are uniform.
4. All models are subject to interpretation, but some are less contrived.
5. All statistical tests have thresholds, but some are mis-set.
6. All model checks consider variations of the data, but some variants are more relevant than others.

Maxim 1

"All data have stories, but some are mistold."

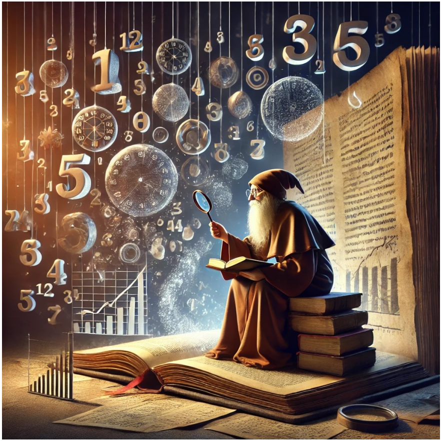


Image Credit: ChatGPT.

Some stories behind the data are mistold.

Knowing **the story behind the data** helps improve modeling for more reliable inferences and correct potential model mis-specification.

- Sampling mechanism
- Selection effect
- Pre-processing
- Calibration

Some stories behind the data are mistold.

1. Sampling mechanism
   - Astronomical data are not a random (representative) sample. Non-uniform coverage. Possibly a biased inference (Kelly, 2007).
   - Be careful about the systematics of any survey. E.g., systematically high peculiar velocities in early H_0 estimates.

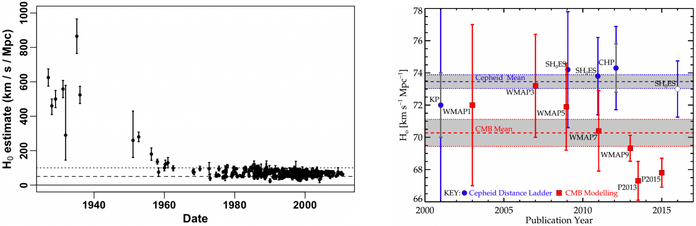


Image Credit: John P. Huchra (left) and Beacon+ (2016) (right).

It is important to iteratively correct and improve existing models because any analysis remains vulnerable to imperfect knowledge of the story behind its data

Some stories behind the data are mistold.

2. Selection effects
    - Astronomical data are often obtained intentionally and purposefully for specific research projects.
    - Once publicly available, people may use them as if they were randomly and uniformly selected, possibly unaware of the danger of selection effect in the original data.
    - Examples include matching/merging multiple catalog data sets, often used as training data.
    - Stratified classification for matching between training and test sets (Autenrieth+, 2024) or importance re-weighting (Izbicki+, 2017).

Some stories behind the data are mistold.

3. Preprocessing

- Most astronomical data are pre-processed via multi-stage software pipelines specific to a given telescope.

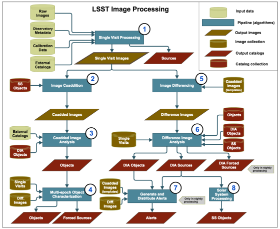


Image Credit: Jurić+(2023).

- Whenever possible, pre-processing procedures should be accounted for within the overall statistical model as much as possible, e.g., cross identification (Budavári & Szalay, 2008; Portillo+, 2017).

Some stories behind the data are mistold.

4. Calibration
   - The process of characterizing uncertainties and bias corrections induced by instruments, enabling the translation of measured signals into physically meaningful units.
   - Details of calibration can have a dramatic impact on the quality of the data (Villanueva et al., 2021).
   - Whenever available, information about calibration uncertainty must be incorporated in to the analysis (Lee+, 2011; Xu+, 2014; Chen+, 2019; Marshall, 2021).

“Knowing **the story behind the data** helps improve modeling for more reliable inferences and correct potential model mis-specification.”

Maxim 2

“All assumptions are meant to be helpful, but some can be harmful.”

Image Credit: ChatGPT.

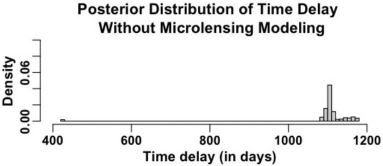
Some assumptions can be harmful.

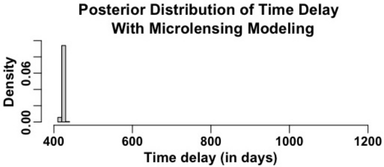
Popular statistical models were developed for specific purposes or motivated by particular problems.

- May not account for unusual features of astronomical data.
- Even linear regression needs correction / extra modeling for astronomical data analysis (non-Normality, heteroscedasticity, etc.).
- Checking the assumptions of popular models is often facilitated by well-defined model checking procedures, such as residual analysis (Tanaka+, 1995; Bulbul+, 2014; Reeves+, 2009; Mandel+, 2017).

Some assumptions can be harmful.

Model checking in light of the knowledge of domain science can reveal evidence of potential model misspecification. For example, in the time delay estimation of quasar Q0957 + 561 (Tak+, 2017, 2018),





The story behind the data (about microlensing) in Hainline+ (2012).

Some assumptions can be harmful.

The χ^2-minimization is often used on (Poisson) count data, but it bases on a Gaussian approximation. This approximation becomes inaccurate if

- the estimated variance of the approximate Gaussian distribution is quite different from that of the observed (or average) count (Feigelson & Babu, 2012),
- the underlying Poisson assumption is not appropriate (e.g., due to overdispersion), or
- counts in some bins are too small,

Building a model directly on the count data is better (Hilbe, 2014; Kelly+, 2012), which also facilitates the use of information criteria.

Some assumptions can be harmful.

The central limit theorem bases various asymptotic results (such as MLE, likelihood ratio), and requires regularity conditions and large data size.

- Regularity conditions are crucial, for example, number of parameters and that of data in calibration (Chen+, 2019), a non-nested model or a nested model but on the boundary in likelihood ratio test (Protassov+, 2002)
- Large data size is necessary. More parameters need more data.
- Alternative in goodness-of-fit test when data size is not large (Chen+, 2024).

Maxim 3

“All prior distributions are informative, even those that are uniform.”

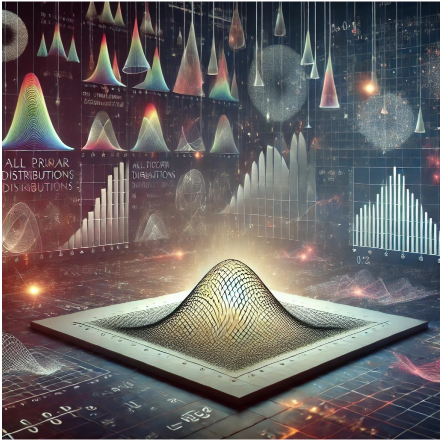


Image Credit: ChatGPT.

Uniform priors are informative.

It is difficult to find an article that conducts a Bayesian analysis without using uniform priors (often uniform on the logarithmic scale).

Is the interpretation of Bayesian inference more straightforward than that of frequentist (hypothetical repeated sampling scenario)?

Interpretation of Bayesian inference depends on that of prior distribution, e.g., the result of "log(Flux) $\sim$ Unif" is interpreted as if "Flux $\sim$ Unif".

Uniform priors are informative.

In some sense, uniform priors lessen the burden to become a Bayesian.

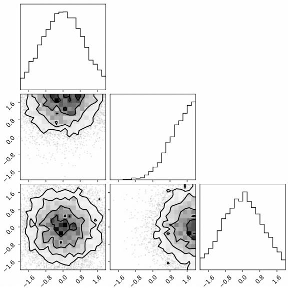
- Relatively easier to conduct Bayesian analysis to do more than MLE.
- Provides a way to incorporate scientific knowledge into the bounds.

However, the bounds of the uniform priors are HARD bounds that completely excludes values of parameters outside.

Lindley (1985) warns against assigning a probability of zero to events that are not logically impossible in what is often referred to as Cromwell's rule, "I beseech you [to] think it possible that you may be mistaken."

Uniform priors are informative.

It is not uncommon to see examples in the literature!



A search for articles that include the word ‘Bayesian’, published in MNRAS in June 2024 yields 58 articles, of which 17 displayed corner plots; 7 of these plots (41.2%) showed boundary issues.

Uniform priors are informative.

Besides, uniform prior distributions are highly informative in high dimensions (Gelman 1996, p223)

- When model parameters are constrained in increasing order, e.g., breaking points in multiply broken power laws (Gelman+, 2017).
- Jointly improper uniform priors (Gelman+, 2017, Sec. 4).

That is, uniform priors can be more informative than you might imagine.

Maxim 4

“All models are subject to interpretation, but some are less contrived.”

Image Credit: ChatGPT.

Some interpretations are less contrived.

A long-term community effort is the key to building time-tested models with widely accepted astrophysical interpretations.

For example, in AGN's variability study,

1. Damped random walk process with physical interpretations of model parameters (Kelly+, **2009**).
2. More empirical evidence (MacLeod+, **2010**; Kozlowski+, **2010**; Kim+, **2012**; Andrae+, **2013**)
3. Many warnings (Mushotzky+, **2011**; Zu+, **2013**; Graham+, **2014**; Kasliwal+, **2015**; Kozlowski, **2016**).
4. More general stochastic processes with physical interpretations of model parameters (Kelly+, **2014**)
5. More empirical evidence (Moreno+, **2019**; Yu+, **2022**).

Maxim 5

“All statistical tests have thresholds, but some are mis-set.”

Image Credit: ChatGPT.

Some testing thresholds are mis-set.

Issues in *p*-value are less likely in astronomy, possibly due to more conservative thresholds, 3σ (e.g., 0.05 vs 0.0027).

One issue that is less recognized in astronomy is related to multiple hypothesis testing.

- For example, 4.6σ and 5.1σ evidence reported for the first gravitational wave detection (Abbott+, 2016).
- 'Naively' comparing each to the 3σ threshold doubles the family-wise error rate (FWER; probability of committing at least one type I error among the two tests).
- To control a typical 3σ-level FWER, each must be compared to the 3.2σ threshold.

Some testing thresholds are mis-set.

The most popular method to control the FWER in conducting *m* hypothesis tests is **Bonferroni** correction.

- When *p*-values are sorted in increasing order, reject H_0 if '$p_j < 0.0027/m$' to maintain FWER $=$0.0027 (3σ-level).
- Almost impossible to reject when the number of test *m* is large.

A method to overcome this issue is to control the **false discovery rate** (FDR, Benjamini & Hochberg, 1995; Benjamini, 2010).

$$E\left(\frac{\text{false discoveries}}{\text{all discoveries}}\right) = E\left(\frac{\text{false positives}}{\text{all positives}}\right) \leq 0.0027.$$

- Reject H_0 if '$p_j < (0.0027 \times j)/m$' to maintain FDR $=$0.0027.
- Leading to more discoveries (rejections) even when *m* is large.

Maxim 6

“All model checks consider variations of the data, but some variants are more relevant than others.”

Image Credit: ChatGPT.

Some variants are more relevant than others.

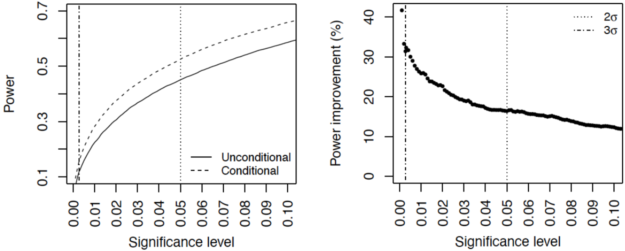
Replicating the data is crucial in both frequentist and Bayesian inferences. How should we generate them to be as realistic as the real data?

> "Always be aware of what you don't know.
> But don't lose track of what you do know."

- Conditioning on what you do know, such as experimental conditions, instrumental effects, exposure time, and sample size.
- For example, $Y_B^{\mathrm{rep}} \leq Y^{\mathrm{obs}}$ as $Y^{\mathrm{obs}} = Y_B^{\mathrm{rep}} + Y_S^{\mathrm{rep}}$ (Roe & Woodroofe, 1999).
- It reduces uncertainty, error bars, and the lengths of confidence intervals, while increasing the testing power.

Some variants are more relevant than others.

Also, conditioning on MLE in goodness-of-fit tests can increase the testing power substantially (Chen+, 2024).



(Left) The testing power with conditioning is uniformly higher.
(Right) With the 3σ threshold, the power improves by 31.4%.

Six Sigma?

https://eastmanbusinessinstitute.com
How Jack Welch Transformed GE with Six Sigma
6σ
Six Sigma
DEFINE
MEASURE
ANALYZE
IMPROVE
CONTROL

Image Credit: Eastman Business Institute (left) and Mike Loughrin (right)

Popular in quality control and business: Just 3.4 defects in one million!

We hope the routine adoption of **Six Maxims** similarly contribute to the quality improvement in astronomical data analysis.

References are in arXiv:2408.16179