Draft version October 8, 2024
Typeset using LaTeX **twocolumn** style in AASTeX63

Six Maxims of Statistical Acumen for Astronomical Data Analysis

Hyungsuk Tak [1] Yang Chen [2] Vinay L. Kashyap [3] Kaisey S. Mandel [4,5] Xiao-Li Meng [6] Aneta Siemiginowska [3] and David A. van Dyk [7]

[1] *Departments of Statistics, Astronomy & Astrophysics, Institute for Computational and Data Sciences, Pennsylvania State University University Park, Pennsylvania 16802, USA*
[2] *Department of Statistics, University of Michigan Ann Arbor, Michigan 48109, USA*
[3] *Harvard Smithsonian Center for Astrophysics Cambridge, Massachusetts 02138, USA*
[4] *Institute of Astronomy and Kavli Institute for Cosmology, University of Cambridge, Cambridge CB3 0HA, UK*
[5] *Statistical Laboratory, DPMMS, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK*
[6] *Department of Statistics, Harvard University Cambridge, Massachusetts 02138, USA*
[7] *Statistics Section, Department of Mathematics, Imperial College London London SW7 2AZ, UK*

(Received June 1, 2019; Revised January 10, 2019; Accepted October 8, 2024)

Submitted to ApJS

ABSTRACT

The acquisition of complex astronomical data is accelerating, especially with newer telescopes producing ever more large-scale surveys. The increased quantity, complexity, and variety of astronomical data demand a parallel increase in skill and sophistication in developing, deciding, and deploying statistical methods. Understanding limitations and appreciating nuances in statistical and machine learning methods and the reasoning behind them is essential for improving data-analytic proficiency and acumen. Aiming to facilitate such improvement in astronomy, we delineate cautionary tales in statistics via six maxims, with examples drawn from the astronomical literature. Inspired by the significant quality improvement in business and manufacturing processes by the routine adoption of Six Sigma, we hope the routine reflection on these Six Maxims will improve the quality of both data analysis and scientific findings in astronomy.

Keywords: Astrostatistcs (1882) – Astronomy Data Analysis (1858)

1. INTRODUCTION

Although data science aims to address the myriad challenges arising from the entire life cycle of data (Wing 2019), there are a number of unique, or at least unusual, characteristics of astronomical data that demarcate the statistical challenges in astronomy and affect our approach to analyzing astronomical data. First, astronomical observations are not obtained from designed experiments in the traditional sense. There are no experimental settings that the astrophysical researcher compares by controlling conditions, such as treatment versus placebo. Consequently, observations are not exactly repeatable in the sense that observational conditions, instrumental properties, or the astrophysical phenomena themselves are changing over time. For instance, we cannot observe the same supernova explosion multiple times under different controlled conditions. Second, calibration (e.g., of instruments) is a crucial step of the observation process because it allows us to connect the observed signals to the underlying physics (see, e.g., Guainazzi et al. 2015). Unfortunately, calibration is never exact and thus adds uncertainty to the final analysis (e.g., Lee et al. 2011). Third, sparsity is inevitable even with big data, in a sense that researchers are always interested in the most distant objects and the

Corresponding author: Hyungsuk Tak (tak@psu.edu)

faintest signals which are studied using small subsets of the full data. For instance, new classes of astronomical phenomena are rarely first discovered as bright sources, but are often among the most interesting scientifically. Fourth, observed astronomical objects are at different stages of their life cycles and evolve on different time scales, all of which are much larger than we can observe. This characteristic can help us understand long time scales via a population study, but homogeneity and completeness of the data may become an issue. Finally, measurement error uncertainties are heteroscedastic and are often given as constants along with the data (Feigelson & Babu 1998; Feigelson et al. 2021).

Taken together, these unusual characteristics can lead to challenges, especially when using off-the-shelf data-analytic tools to analyze astronomical data (Siemiginowska et al. 2019). This is because underlying assumptions of standard statistical methods do not typically take account of these features. For example, even a simple linear regression model requires extra modeling assumptions to account for selection effects and measurement errors for astronomical data analysis (Kelly 2007). Thus, it is important to check these underlying assumptions on a case-by-case basis to ensure sound astronomical data analysis, especially when deploying methods that were developed outside of astronomy.

It is worth pointing out an important distinction in the jargon between the astronomical and statistical literature. A "model" in astrophysics refers to a parsimonious mathematical representation of expected (or predicted) signal from a physical process that generates emission that is eventually detected via telescopes. This could be the blackbody energy spectrum, or the pulse profile of a pulsar, or the number density of a population of sources in a globular cluster projected onto the sky, etc. In contrast, a "model" in statistics is a stochastic representation of the data-generating process that accounts for discrepancies between the astrophysical model and the data. This stochastic representation is indexed in that it is specified up to a set of unknown model parameters that are fit to the data. It reflects systemic adjustments (including observational constraints and instrument effects), selection effects, stochastic components such as Poisson and Gaussian errors, and anything else that effects the distribution of the data. To take a simple example, the choice of the Poisson($g(\theta)$) model to represent photon counts forms a statistical model, whereas the astrophysical model stipulates the functional form of $g(\theta)$, e.g., $g(\theta)$ could be a power-law in energy E, i.e., norm $\times E^{-\alpha}$, with model parameters $\theta = \{\text{norm}, \alpha\}$. The physical model is designed to describe a physical process without necessarily representing the stochastic aspects of data generation that lead to uncertainty in parameter estimation. Uncertainty quantification, on the other hand, is at the heart of the statistical model which aims to represent data and its variability as fully as possible. A statistical model also describes the hierarchical connections between the various processes that translate incoming photons to observed electronic signals (e.g., van Dyk et al. 2001).

As an illustration of the particular difficulties in handling astronomical data, consider the estimation of the time delay between the multiple images of a strongly gravitationally lensed time-varying source. In estimating the time delay between gravitationally lensed light curves of Q0957+561 (Hainline et al. 2012), Tak et al. (2018b) adopt a damped random walk statistical model (also known as a continuous-time auto-regressive model of order one or an Ornstein-Uhlenbeck process) as a data generating process. This model reveals multiple modes in the posterior distribution of the time delay parameter as illustrated in the top panel of Figure 1. The height of the mode near 400 days is much less than the mode near 1100 days. However, it turns out that the highest mode near 1100 days is spurious, caused by model misspecification. The modes near 1100 days disappear when the astronomical model additionally incorporates polynomial regression to account for the effect of microlensing (Tak et al. 2017b) that is known to be present in the data (Hainline et al. 2012); see the bottom panel of Figure 1. Consequently, the mode near 400 days becomes prominent, in agreement with some previous analyses of this quasar (Schild 1990; Shalyapin et al. 2008).

This example points out several important aspects in astronomical data analysis. First, different model fits on the same data can reveal completely different possibilities, e.g., for the time delay of Q0957+561. All of these possibilities are worth proper investigation in the context of available scientific knowledge, in an attempt to determine which are simply the result of model misspecification and which are new scientific discoveries. Second, blindly making inference based on the highest mode of the posterior distribution or likelihood function (or smallest loss function in machine learning methods) can be misleading, as illustrated in the top panel of Figure 1. Thus it is essential to check whether the model captures important characteristics of the data sufficiently well before drawing any conclusions. Lastly, it is the story embedded in the data that can provide insight for improved modeling of physical phenomena, such as microlensing. The better the statistical and astronomical models reflect the data, the better the quality of what the data reveal to us.

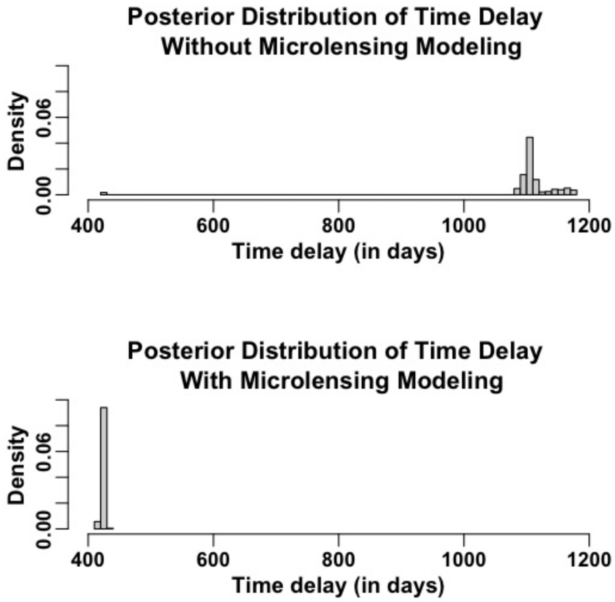

Figure 1. The posterior distributions obtained under two different models for the time delay given the same data of two gravitationally lensed light curves of Q0957+561 (Hainline et al. 2012). The model producing the posterior distribution in the top panel does not account for the effect of microlensing, producing multiple modes (Tak et al. 2018b), even though the tiny mode near 400 days is of scientific interest. The modes near 1100 days do not appear in the bottom panel where the model additionally accounts for the microlensing effect via polynomial regression (Tak et al. 2017b). It turns out that the modes near 1100 days in the top panel are spurious caused by model mis-specification.

In what follows, we discuss several issues that arise in astronomical data analyses in light of the unique or unusual features of astronomical data. We formulate our observations into the following six maxims, each of which is in the sprint of George Box's well-known aphorism "all models are wrong but some are useful" (Box & Draper 1987):

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 can be given interpretations, but some are more compelling.
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.

While we believe that the statement of each of the maxims is new, the ideas that underlie them are not. Rather, the maxims are merely concise statements that we hope capture a sense of the reasoning that defines statistics as a discipline and that is the culmination of the work of generations of data-facing researchers. Our aim is to encourage researchers to carefully consider their (possibly implicit) modelling and statistical assumptions and how these assumptions may affect scientific findings. We hope that by keeping the maxims in mind as part of their daily data-analysis routine, researchers will improve the quality of both data modeling and scientific findings in astronomy.

2. ALL DATA HAVE STORIES, BUT SOME ARE MISTOLD.

In this section, we explain several issues in modeling astronomical data, such as sampling mechanisms, selection effects, preprocessing, and calibration, and discuss possible solutions to improve the quality of astronomical data analysis.

2.1. *Sampling mechanism*

Statisticians typically assume that the data are measurements of a statistical sample that is representative of the larger class of objects under study. For example, we might have a sample of white dwarf stars from the Milky Way Galaxy and measure the metallicity of each or we might have a sample of exoplanets and measure the mass of each. Formally, statisticians may assume that we have obtained a probability sample from the larger class or population. (In a probability sample, all objects in the population have a known non-zero probability of being included in the sample.) Unfortunately, such a sample is nearly impossible to obtain in astronomy. While it is true that measurements of the properties of individual objects have become more accurate, this accuracy does not translate into a more representative sample of objects. In fact, none of the so-called all-sky surveys have uniform coverage as they all provide preferential or deeper coverage of specific parts of the sky. For example, Sloan Digital Sky Survey (SDSS) is targeted at the northern celestial hemisphere, the Rubin Observatory has lower coverage in the northern hemisphere, and space-based observatories like TESS and eROSITA have deeper coverage towards the poles. Likewise, narrowly focused pencil-beam surveys like the *Hubble* Deep Field or *Chandra* Deep Field surveys have varying sensitivity across the field of view due to the detector or telescope responses.

Modeling such data without paying attention to the exact nature of the sampling mechanism and how well they represent the population of interest can result in biased inferences (Kelly 2007, Section 5). Astronomers are generally aware of adverse selection effects introduced by the Eddington or Malmquist biases (Landy & Szalay 1992; Teerikorpi 2015), but we caution that the systematics of any survey or measurement must be carefully considered on a case-by-case basis.

A well-known example occurs in Hubble (1929), where systematically high peculiar velocities in the local Galactic neighborhood initially led to a large overestimate of the eponymous Hubble constant, H_0. Indeed, the importance of modeling systematic uncertainty is apparent throughout the history of the measurement of H_0. Figure 2 shows that early estimates, from the mid-to-late twentieth century, were either around 50 km s^{-1} Mpc^{-1} denoted by the dashed horizontal line (e.g., Sandage & Tammann 1975) or around 100 km s^{-1} Mpc^{-1} visualized by a dotted horizontal line (e.g., de Vaucouleurs & Bollinger 1979; de Vaucouleurs & Peters 1986). The half length of each vertical bar around the point estimate represents its 1σ uncertainty.

More recently, significant improvements in instrumentation and techniques have led to better understanding of the systematics involved and have narrowed the range of the measured values of H_0 further. However, a statistically significant discrepancy remains among the estimates, raising a question regarding the validity of the standard cosmological model (Verde et al. 2019; Efstathiou 2020; Riess et al. 2021). For example, Figure 3 (excerpted from Figure 1 of Beaton et al. (2016)) illustrates the tension between estimates of H_0 from the so-called late-Universe measurements calibrated by the Cepheid distance scale (in blue), and early-Universe measurements obtained by the cosmic microwave background (in red). Feeney et al. (2018) show that the Bayesian evidence of the standard cosmological model is about seven times smaller than that of an extended cosmological model, that includes an additive deviation from the standard cosmological model. The corresponding Bayes factor between the two cosmological models is 0.15 ± 0.01 given the Planck 2015 XIII data (Planck Collaboration et al. 2016) and the distance-ladder data of Riess et al. (2016) with extra supernova outliers being considered.

The role of systematics is clear in recent work describing the tension among competing estimates of H_0 owing to the extraordinary efforts of the astronomical and cosmological communities to pin down H_0. How does a researcher with fewer resources recognize similar effects in their analysis and remedy them? We posit that this requires an iterative process that implements corrections and appropriately incorporates model complexity in follow-up analyses. Still, it is important to recognize that any analysis remains vulnerable to imperfect knowledge of the story behind its data.

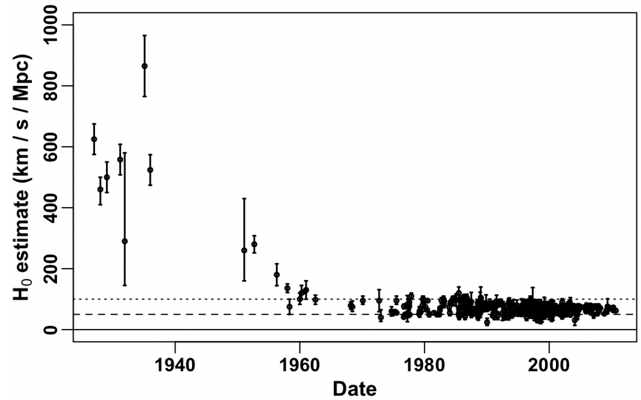

Figure 2. A history of Hubble constant estimates made between 1920 and 2008. The dashed horizontal line indicates the H_0 estimate at 50 km s^{-1} Mpc^{-1} and the dotted horizontal line represents the estimate at 100 km s^{-1} Mpc^{-1}. The half lengths of vertical bars around dots represent 1σ uncertainty. This figure is generated using the estimates of H_0 compiled by John P. Huchra from the literature as part of the NASA/HST Key Project on the Extragalactic Distance Scale (https://lweb.cfa.harvard.edu/~dfabricant/huchra/hubble.plot.dat).

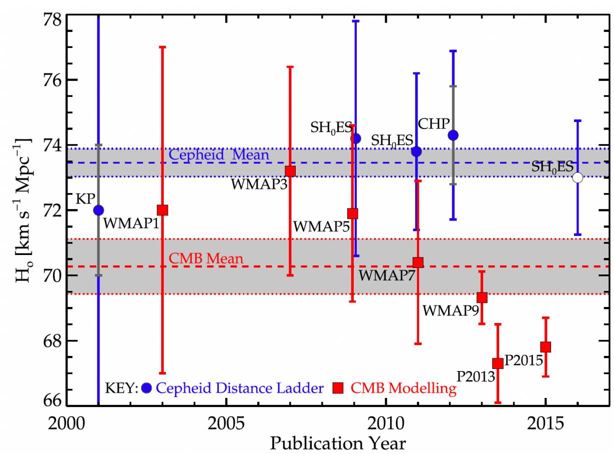

Figure 3. The tension between the early- and late-Universe measurements of H_0. The estimates in blue are computed by standard candle method calibrated by Cepheid and those in red are obtained by the cosmic microwave background measurements under the standard cosmology (ΛCDM). This figure is excerpted from Figure 1 of Beaton et al. (2016) by permission of the AAS.

2.2. *Selection effect*

In addition to non-uniform coverage, astronomical data are often obtained intentionally and purposefully for specific research projects. When such astronomical data become public through various archives, other researchers may download them and use them as if they were randomly and uniformly selected, possibly unaware of the danger of selection effect in their sample. Likewise, when the contents of different surveys are examined together, their individual characteristics can affect the overall interpretation in complex ways. This is well-appreciated when different catalogs are matched (e.g., Budavári & Szalay 2008; Rots et al. 2017a,b), but less so when population studies are carried out. Catalog data are often used as training sets when applying machine learning methods, even though such training sets may not represent the population of interest well due to the non-uniform coverage and selection effects within the catalog. For example, the Chandra Source Catalog (CSC; Evans et al. 2010) provides a selection of fields, each observed for individual scientific reasons. Such "samples of convenience" are not probabilistic and not representative of the population, in contrast to flux-limited all-sky surveys like the ROSAT All-Sky Survey (RASS; Voges 1993; Voges et al. 1999; Boller et al. 2016). An example of how to deal with these effects is provided by Revsbech et al. (2017) and Autenrieth et al. (2024) who reduce the effect of a biased training set in classifying Type Ia supernovae via stratification. It forms training sets that are more representative of the corresponding strata within the test set. Similarly, Izbicki et al. (2017) propose a non-parametric density estimator for photometric redshift that accounts for selection bias in a non-representative training set by importance reweighting of the training set.

2.3. *Preprocessing*

Most astronomical data are pre-processed via multi-stage software pipelines specific to a given telescope. As illustrated in Figure 4, in each stage of the pre-processing hierarchy, one astronomer's inference is passed down to be used as an input by the next astronomer. Another inference is then made with the previous inference being treated as the data. Unfortunately, this pre-processing is often ignored even though the pre-processing steps can reveal evidence of potential systematic errors. For example, in the case of solar flares databases, precision of recorded flare intensities, complex detection/missing characteristics, temperature effects, incompleteness in matched features may all cause systematic errors in the data (Ryan et al. 2012; Aggarwal et al. 2018).

As another example, catalog data pre-processing is performed via standard pipelines and assumptions. This pre-processing procedure generally affects the catalog quality and reliability; outliers may arise if measurements are not performed in a consistent way; different definitions of upper limits may cause an issue of censoring; an incorrectly implemented pre-processing procedure may introduce systematic error. Thus, it is important to understand how the catalog quantities are derived from the raw data through a chain of pre-processing stages, especially when different catalogs are compared or merged (Budavári & Szalay 2008). Whenever possible, the statistical and systematic errors introduced by the pre-processing procedure should be accounted for within the overall statistical model as much as possible (e.g., Portillo et al. 2017).

2.4. *Calibration*

Calibration is a foundational part of astronomical inference, more so than in any other physical science. While instrument calibration is indeed used extensively in fields like experimental physics and geophysics, it is of particularly critical importance in astronomy[1]. Astronomical data are for the most part obtained through observations of remote sources, with physical quantities inferred by transforming the observed signals from a detector. Each telescope or focal plane instrument has its own specific characteristics that affect this transformation, and considerable effort is put into determining these, and tracking changes to them (see, e.g., Guainazzi et al. 2015; Partridge et al. 2016; Payne et al. 2020). Ground-based photometric optical astronomy still relies on obtaining regular observations of "standard stars" with similar airmass to the target being observed, so even atmospheric variations must be adjusted for. High-

[1] Note that the term "calibration" is interpreted differently in astronomical, compared to statistical, literature. In astronomy, it refers to the process of characterizing uncertainties and bias corrections induced by instruments, enabling the translation of measured signals into physically meaningful units. In statistical literature, however, calibration generally refers to a process of "inverse regression", where measurements of dependent quantities are used to predict corresponding standard measurements, mediated through a known model function. For example, if a functional form $Y = f(X)$ is learned using a training data set, new measurements of a test data set Y_0 are used to predict $X_0 = f^{-1}(Y_0)$. This is mainly motivated by instrumental calibration in chemistry, where high-quality "standard" measurements X are more time-consuming and expensive than "test" measurements Y. The statistical literature includes theories and methods for various linear, non-linear, multivariate, and dynamic approaches to statistical calibration (Osborne 1991; Kubokawa & Robert 1994; Brown 1994; Oman & Srivastava 1996; Rivers 2014; Brown 2018). Methods are divided into those designed to handle the case where both standard and test measurements have appreciable error, known as comparative calibration (Kelly 2007; Schafer & Purdy 1996) and methods where the standard measurements are assumed to be perfect or nearly perfect, known as absolute calibration.

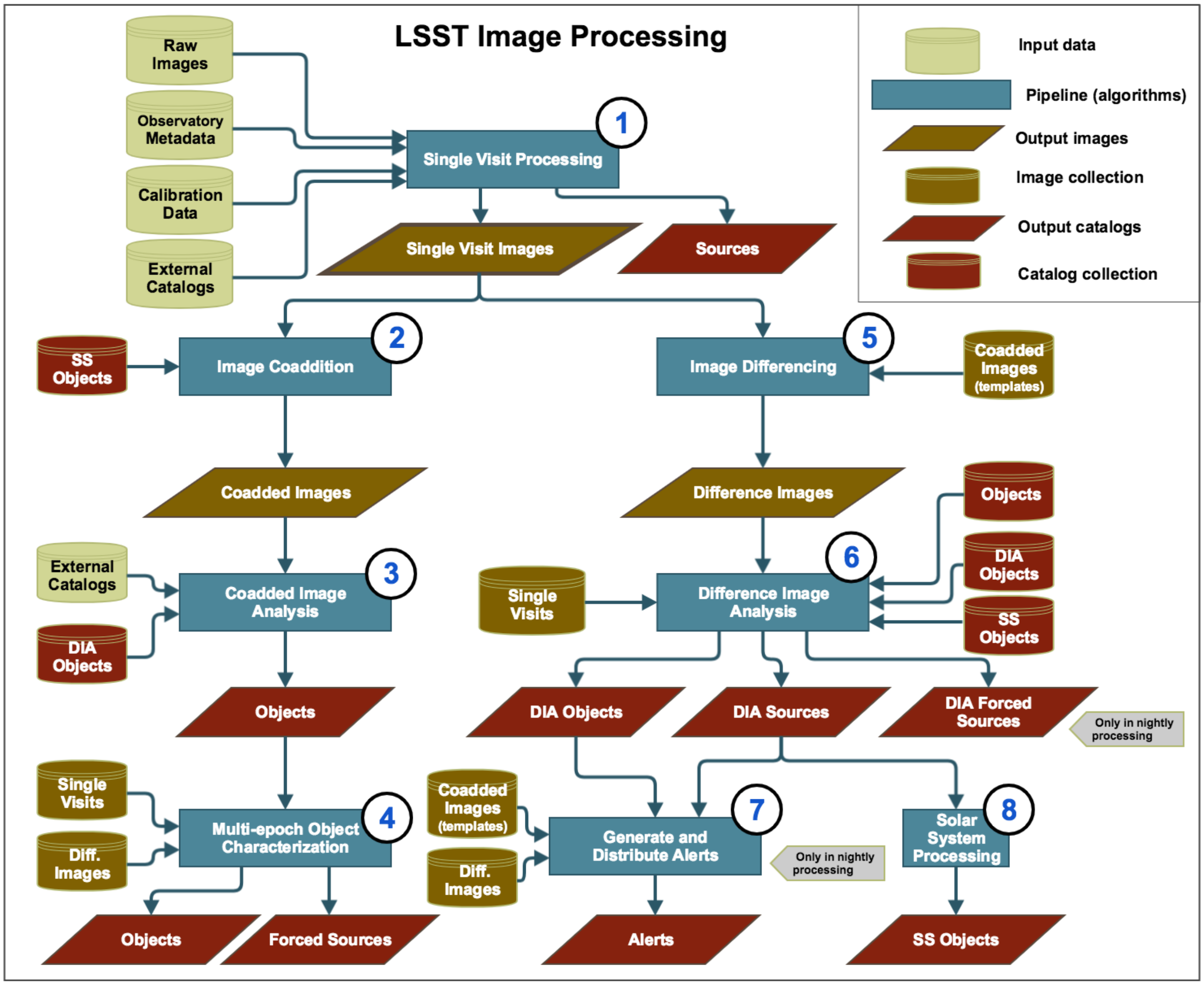

Figure 4. A diagram illustrating an example of a data reduction pipeline, excerpted with permission from Figure 2 of the Rubin Observatory LSST Data Products Definitions Document (Jurić et al. 2023). Each of the numbered boxes expands into another flow diagram. While the details in this figure are not important in the current context, it illustrates the complexity of the data pre-processing.

energy astronomical telescopes construct and store detailed tabular models of the response of a detector to a monochromatic photon[2], and every mission measures and stores its sensitivity (also called effective area in high-energy astronomy) as a function of photon wavelength. As noted by Villanueva et al. (2021), the details of calibration can have a dramatic impact on the quality of the data.

It is important to understand, however, that the available calibration products are not perfect. They are the result of measurements carried out in controlled conditions, and thus include measurement errors, as well as systematics that manifest themselves once the instruments are deployed (often in harsh space environments where the chances of radiation damage is high). Differences in calibration between different instruments must be weighed when different data streams are considered together. Where available, calibration uncertainty information must be folded in to the analysis (Lee et al. 2011; Xu et al. 2014). More recently, efforts have been made to derive corrections to effective areas and to source flux estimates based on simultaneous observa-

[2] See OGIP Calibration Memo CAL/GEN/92-002 and addendum (https://heasarc.gsfc.nasa.gov/docs/heasarc/caldb/docs/memos/cal_gen_92_002/cal_gen_92_002.html, https://heasarc.gsfc.nasa.gov/docs/heasarc/caldb/docs/memos/cal_gen_92_002a/cal_gen_92_002a.html)

tions of sources with different instruments even in the absence of an absolute reference using multiplicative shrinkage (see, for example, Chen et al. 2019a; Marshall 2021).

There is a common theme in these examples. Knowing the story behind the data allows one to correct for potential model mis-specification, while not knowing the story may leave one oblivious to the same issue. Understanding their data, including limitations in the data collection process and potential selection effect, enables researchers to make appropriate corrections themselves. In this way, being attentive to the story behind their data enables researchers to make more reliable inferences regarding their populations and sources of interest.

3. ALL ASSUMPTIONS ARE MEANT TO BE HELPFUL, BUT SOME CAN BE HARMFUL.

Popular statistical models were developed for specific purposes or motivated by particular problems. Some well known models, such as Gaussian linear regression, can be applicable in a wide variety of settings across various disciplines including astronomy and astrophysics with little difficulty. Another example is survival analysis which is one of the most popular data analyses in bio-medical sciences. In fact, classical survival analysis is not directly applicable to astronomical data because censoring in astronomy is due to statistical measurement errors rather than exactly measured failures. Nonetheless, it has been successfully applied to analyzing left- or right-censored data caused by telescope sensitivity in astronomy (Feigelson & Nelson 1985; Isobe et al. 1986).

However, the use of well-known models without careful consideration of their assumptions must be discouraged because standard statistical models do not account for unusual features of astronomical data or models. Even the standard linear regression model has underlying Gaussian assumption, while astronomical data may deviate from Gaussianity with outlying observations, low Poisson counts, background subtraction, error propagation, binned data, and/or heavy-tailed and asymmetric distributions.

As another example, standard statistical models, such as linear regression or auto-regressive moving average models, often assume that measurement errors are homoscedastic and unknown. However, astronomical data often come with one-sigma measurement-error uncertainties that are heteroscedastic and are (assumed to be) completely known (Feigelson et al. 2021). To model these heteroscedastic measurement-error uncertainties, community efforts have been made in various fields of astronomy. Introducing errors in measurement to standard models is a popular idea in statistics (Fuller 1987). The technique was later tailored to heteroscedastic measurement errors in astronomical data in the contexts of (but not limited to) linear regression (Akritas & Bershady 1996; Kelly 2007; Andreon & Hurn 2013; Sereno 2016), damped random walk process (Kelly et al. 2009; Hu & Tak 2020), continuous-time ARMA(p, q) process (Kelly et al. 2014; Meyer et al. 2023), and astronomical object classification (Bovy et al. 2011; Shy et al. 2022).

Checking the assumptions of popular models is often facilitated by well-defined model checking procedures. For example, checking model assumptions via residual analysis is common in regression because it can provide insight into possible improvement for the current model fit. Tanaka et al. (1995) improve a poor continuum fit of spectral data via subsequent residual analysis; Bulbul et al. (2014) and Reeves et al. (2009) detect emission lines and absorption lines via residuals; Mandel et al. (2017) compare conventional and proposed models for the color-magnitude relation of Type Ia supernova by checking their Hubble residuals to see which model is better supported by the data.

Residual analysis often provides hints that can be used to improve model assumptions. For instance, when a model for light curves such as a damped random-walk model relies on a Gaussian measurement error assumption (Kelly et al. 2009), a residual analysis may reveal some evidence against the Gaussian assumption in the presence of outliers. In an effort to improve residual analysis, Tak et al. (2019) and Wang & Taylor (2022) derive a heavy-tailed version of the damped random-walk model that is still able to constrain the same model parameters in a robust manner.

Besides standard model checking procedures, investigating the fitted model in light of the knowledge of domain science is also crucial as it can reveal evidence of potential model misspecification. For instance, the sensitivity or dependence of model fits on the starting values of optimizers or Monte Carlo samplers is not necessarily an indication of a numerical problem. It could instead point to a multi-modal outcome on a non-convex surface of the parameter space, providing several distinct model fits at different modes. Considering that a model describes a data generation process, it also implies that distinct sets of parameter values for a given model could have generated the same observed data, even though each set may not be equally likely to have generated the observed data.

Alternatively, a multi-modal likelihood function may indicate that the model is misspecified or is not elaborate enough to describe the data, either of which can lead to an unrelaible fit. Mode-based estimates, such as maximum likelihood estimates and posterior modes, aim to compute the parameter values corresponding to the particular model within the posited class that best matches the data (under a criteria determined, e.g., by the likelihood or posterior). Even the best match within the posited class of models, however, may not be very good if the class is not sufficiently rich. Using the fully Bayesian posterior distribution with a misspecified model can also lead to unreliable results, as emphasized in Figure 1, where multimodality disappears when we additionally model microlensing. We emphasize that a well specified model is key to any model-based method, and thus it is critical for researchers to check the fit of their posited model in light of domain science knowledge, instead of blindly proceeding with the highest mode or other computed summary as the best model fit.

Another popular estimation tool in astronomy is χ^2-minimization that is built on a Gaussian approximation for the measurement error. For example, when the data are binned[3] Poisson counts, a Gaussian approximation to the Poisson counts is required for χ^2-minimization (Cash 1979; Humphrey et al. 2009; Bonamente 2020). Thus, it is important to understand the limitations that might affect the validity or accuracy of this approximation. The method is often misused in the context of astronomical data analysis, for example, when the estimated variance of the approximate Gaussian distribution is quite different from that of the observed (or average) count, which contradicts the validity of Gaussian approximation to Poisson counts (Feigelson & Babu 2012, Chapter 7.4). The approximation itself can be quite misleading when the underlying Poisson assumption is not appropriate, e.g., when the count data are overdispersed. The approximation becomes less accurate when counts of some bins are small. In this case, merging adjacent small bins is one way to improve the accuracy of the approximation while sacrificing the resolution of the data (Greenwood & Nikulin 1996).

Directly building a Poisson model for the counts without using a Gaussian approximation is another possibility. (This has not always been well recognized among astronomers (Hilbe 2014).) For example, Cash (1979) proposes the so-called C-minimization technique, which is operationally the same as finding the maximum likelihood estimate under a Poisson likelihood function. Bonamente (2020) demonstrates that C-minimization outperforms χ^2-minimization with low-count data because the former does not involve a Gaussian approximation. Also, Kelly et al. (2012) adopt Bayesian hierarchical modeling in fitting spectral energy distributions on flux data, instead of using χ^2-minimization. Hierarchical modeling also provides a mechanism for effectively handling overdispersed data (Gelman et al. 2013; Tak et al. 2017a). Another benefit that direct likelihood-based modelling has over χ^2-minimization is that it facilitates the use of information criteria for model selection, such as the Bayesian Information Criterion (Kass & Raftery 1995), which depend directly on the likelihood.

The central limit theorem is the basis for a Gaussian approximation to Poisson counts and generally plays an important role in statistical inference. It stipulates that the distribution of the average of independent observations becomes more Gaussian as the number of observations approaches infinity. The theorem is the basis of many asymptotic results such as the asymptotic normality of maximum likelihood estimates, the asymptotic χ^2 distribution of the likelihood ratio test statistic via Wilk's theorem (Wilks 1938), and the "projection method" for computing error bars (Avni 1976).

To be confident of the applicability of asymptotic results, researchers must check two things: the assumptions required for the results are met and the data set they are analyzing is sufficiently large. First, all asymptotic results depend critically on their own sets of mathematical assumptions known as regularity conditions. Even with an arbitrarily large data set, the central limit theorem itself fails, for example, if the expected value of the square of the averaged observations is not finite (e.g., when averaging ratios) or if the number of parameters increases sufficiently quickly compared to the sample size (e.g., as with instrument calibration Chen et al. 2019b)). The likelihood ratio test that compares the statistical evidence for two posited models is another example where the regularity conditions play a key role. This is because Wilk's theorem only provides the asymptotic distribution of the likelihood ratio test statistic if, among other conditions, the models being compared are nested (i.e., one model is a special case of the other) and the simpler model is not on the boundary of the parameter space describing the more complex model. Protassov et al. (2002) show that the latter condition fails when testing for an added spectral emission line because an *emission* line by definition cannot have a negative normalization but the normalization is zero in the simpler model (with no line).

Second, even if their regularity conditions are met, asymptotic statistical methods are only reliable with sufficiently large data sets. A likelihood function that

[3] We consider the case where the data are intrinsically binned. When this is not the case, Feigelson & Babu (2012) suggest avoiding issues of arbitrary binning by using cumulative distribution function for maximum likelihood estimation.

exhibits multiple significant modes, for example, may be evidence that either the model is misspecified (and a regularity condition is not met) or the data set is not sufficiently large for the asymptotic Gaussian properties of the likelihood to have "kicked in". In practice, it can be difficult to know whether a data set is large enough. Generally, the more fitted parameters, the more data that are required. Goodness-of-fit tests are a particular challenge when the data size is small, because, in effect they compare the posited model with a fully flexible model, i.e., a model with a large number of fitted parameters. The C-statistic[4], for example, is often used for goodness of fit tests in high-energy spectral analysis (Kaastra 2017). If the C-statistic is applied to high-resolution data with many narrow bins, its asymptotic χ^2 distribution is only guaranteed if the expected counts are large in *all* bins. The alternative is to work with fewer larger bins, but this sacrifices the resolution of the data.

When there are insufficient data for asymptotic results, either Bayesian procedures or bootstrap-based methods can be used with small data sets. Unfortunately, both are computational more costly than their asymptotic frequentist counterparts. Higher-order asymptotics, which retain more terms in their functional expansions, sometimes can show advantages in such scenarios. For example, Chen et al. (2024) obtain a computationally efficient and statistically precise procedure for goodness-of-fit tests based on the C-statistic and higher-order asymptotics. The method only involves calculation of moments and works even in low-count settings where the Wilk's theorem (χ^2 asymptotics on likelihood ratio tests) does not apply.

4. ALL PRIOR DISTRIBUTIONS ARE INFORMATIVE, EVEN THOSE THAT ARE UNIFORM.

Although Bayesian analysis has become popular in astronomy (Pierson 2013; Eadie et al. 2023), it is difficult to find an article that conducts a Bayesian analysis without using uniform priors (often uniform on the logarithmic scale). One possible explanation for this popularity may actually be a misunderstanding, namely a perception that the interpretation of Bayesian inference is more straightforward than that of frequentist inference. For example, one might think that a credible interval is a direct statement about the unknown parameters given the data, while confidence intervals need to be interpreted under a hypothetical repeated sampling scenario of the data. However, it is often forgotten that the interpretation of credible intervals hinges on the interpretation of the prior distribution, which can be philosophically as subtle as frequentist's repeated sampling scenarios because prior distributions are chosen by researchers.

For instance, uniform priors are often assumed on a logarithmic scale, that is, $\log(X) \sim \text{Unif}(a, b)$, where a and b are real-valued. One may be tempted to interpret the resulting credible interval as if a non-informative prior were used on the original scale, i.e., on X. A uniform prior on $\log(X)$, however, can be very informative indeed on X, since it corresponds to a power-law prior distribution on X ($d\log(x) = dx/x$) that puts substantial probability mass near the lower bound, e^a. Thus, the posterior distribution depends strongly on whether a uniform prior is assumed on the original or logarithmic scale, and the resulting credible interval needs to be interpreted accordingly. In general, it is a mathematical fact that any prior distribution carries information to be interpreted, as it must specify how likely one state is relative to another; see Section 7 of Craiu et al. (2023) for more discussion.

In some sense, uniform prior distributions and other so-called "non-informative" prior distributions have lessened the burden of subjectivity and prior interpretation for astronomers, making the likelihood (i.e., the data) a dominant source of the posterior variability. In some cases, they also enable a Bayesian inference to be conducted relatively easily for researchers who prefer a Bayesian approach, e.g., to handle nuisance parameters or for uncertainty quantification (Gelman et al. 2017), even when the maximum likelihood estimate is nearly identical to the maximum a posteriori estimate. Moreover, uniform prior distributions provide researchers a way to incorporate scientific knowledge via their boundaries. In most articles, the bounds of uniform priors are clarified to avoid potential posterior impropriety (Tak et al. 2018a).

Even bounded uniform prior distributions, however, must be used with care because the bounds are hard bounds that completely exclude a portion of the parameter space. An issue may occur if the bounds partially or completely exclude important regions of the parameter

[4] The C-statistic is sometimes called the Cash statistic in recognition of Cash (1979) and is defined in different ways by different authors. For example, it is often defined either as -2 times the Poisson log-likelihood function (up to an additive constant, e.g., Eq. (5) of Cash 1979) or as -2 times the log of the likelihood ratio comparing a specific Poisson model with a fully saturated Poisson model (e.g., Eq. (1) of Humphrey et al. 2009). Both definitions are equivalent up to a constant adjustment. The latter is a particular instance of the likelihood ratio test statistic, but with only the alternative likelihood evaluated at its maximum likelihood estimate. We use term C-statistic for the same likelihood ratio test statistic, but follow the standard statistical convention with both likelihoods being evaluated at their respective maximizers; see Section 7 for details.

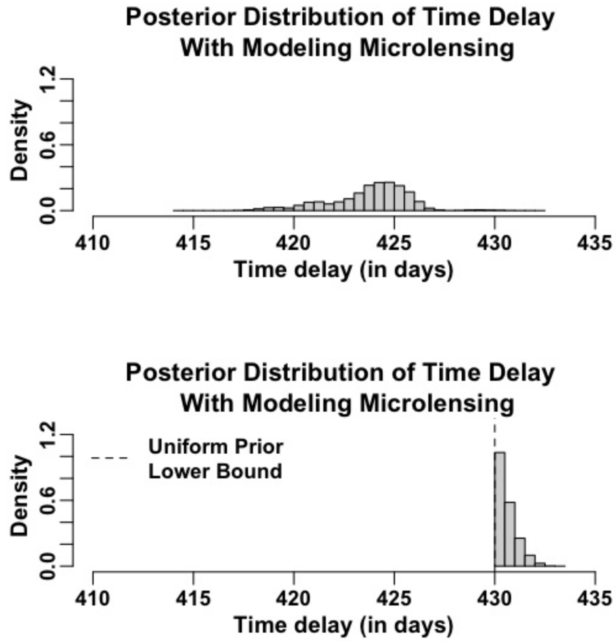

Figure 5. The posterior distribution of the time delay under the microlensing model. The top panel is based on the uniform$(-1178.939, 1178.939)$ prior distribution for the time delay parameter, while the bottom panel is based on the uniform(430, 1178.939) prior distribution. In the bottom panel, the posterior mass accumulates near the lower bound, which may be evidence of mis-specification of the prior distribution.

space a priori. For example, the top panel of Figure 5 magnifies the posterior distribution of the time delay under the microlensing model, previously shown in the second panel of Figure 1. The prior distribution for the time delay adopted in Tak et al. (2017b) is the uniform prior between -1178.939 and 1178.939 days, reflecting the widest range of observation times in the analyzed light curves. As an illustration, let us set up a different lower bound of this uniform prior at 430 days, excluding the modal location at around 425 days. The bottom panel of Figure 5 exhibits the resulting posterior distribution of the time delay. The posterior mass accumulates near the lower bound, as if the posterior mass in the top panel were pushed from the left to the lower bound. This is what happens when the hard bound of a uniform prior zeros out the likelihood beyond the bound. The likelihood cannot overcome this hard bound, regardless of how large the data set is; even one trillion observations would not allow posterior probability beyond the bound. (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[5].)

Substantial posterior probability that is accumulated near the (hard) bounds of a uniform prior distribution may be evidence of mis-specification of the bounded prior distribution. In the astronomical literature, it is not difficult to find examples with substantial posterior mass near the bounds of uniform priors that are set by researchers. This problem can often be identified by inspecting corner plots (pairwise scatter plots with marginal histograms) in published articles, at least when these plots are provided by the authors. A simulated example similar to one we found in the literature survey[6] is shown in Figure 6. When a researchers identifies a boundary issue of this sort, it is critical that they carefully investigate the sensitivity of their results to the bounds of their uniform prior, paying particular attention to the robustness of their scientific conclusions to the choice of bounds. Where there is a natural bound, e.g., where a parameter such as a mass or age must be non-negative or positive, we do not consider the accumulation of posterior mass near this natural bound to be an issue. Therefore, unless there is a strong scientific justification, it is always better to set uniform bounds wide enough not to influence the likelihood.

Besides the boundary issues of uniform priors, a blind use of jointly uniform prior distributions can become a highly informative choice, despite its seemingly non-informative nature (Gelman 1996, p223). For example, when model parameters are constrained such as being in an increasing order (in astronomy, for example, unknown breaking points in multiply broken power laws), a jointly uniform prior on the parameters asymptotically dominates the likelihood function as the number of such model parameters increases. Gelman et al. (2017) provide more examples where uniform prior distributions can result in inferences that do not make sense. A jointly improper uniform prior can also be problematic in high dimensions, even though it results in a proper posterior distribution; Section 4.2 of Gelman et al. (2017) discusses a similar problem that arises when independent Gaussian prior distributions are used in high-dimensional parameter spaces.

[5] The reference to Oliver Cromwell refers to a quotation of his: "I beseech you [to] think it possible that you may be mistaken".

[6] For instance, 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 showed boundary issues.

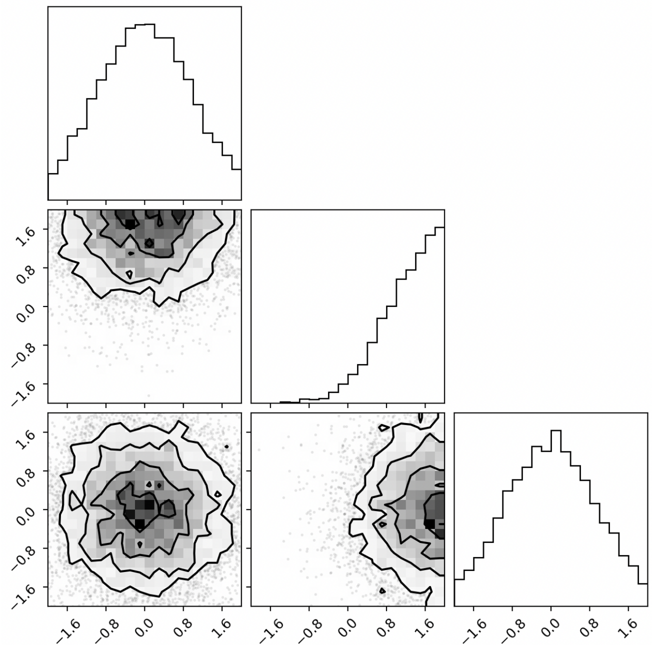

Figure 6. A simulation study where the posterior mass accumulates near the boundary of a uniform prior distribution. The uniform prior distribution of the second parameter is bounded between −0.2 and 0.2, and the posterior mass clearly concentrates on the upper bound. In general, when a uniform prior distribution is used, it is important to check the sensitivity of the resulting posterior distribution to the choice of hard bound on the prior, unless there is a specific scientific justification for the choice of bound.

5. ALL MODELS CAN BE GIVEN INTERPRETATIONS, BUT SOME ARE MORE COMPELLING.

Understanding how the statistical/mathematical interpretation of an empirical data analysis should impact our understanding of astrophysical processes can be challenging. We suggest that it is often best to start with the physics and then consider whether the empirical findings make sense in terms of the physics and/or how we can make sense of them.

Kelly et al. (2009), for instance, carefully investigated a sample of quasar light curves, relating two model parameters of a damped random walk process to physical properties of a quasar. They show that the timescale (short-term variability) of the fitted process is positively (negatively) correlated to both black hole mass and luminosity. This empirical evidence on the relationships between the parameters of the mathematical model and astrophysical properties has since been intensively investigated and is supported by many astronomers (MacLeod et al. 2010; Kozłowski et al. 2010; Kim et al. 2012; Andrae et al. 2013).

For more elaborate model interpretation, the community has also investigated when this interpretation does not hold. Mushotzky et al. (2011) and Zu et al. (2013) show that the damped random-walk process is not suitable for explaining the stochastic variability of quasars when the source variability is on a very short timescale. Also, Graham et al. (2014), and Kasliwal et al. (2015) warn that the process is too simple to explain all types of stochastic variability of quasars. When the underlying true model is not the damped random walk process, Kozłowski (2016) points out that the association between the model parameters and physical properties can be misleading as timescale estimates become biased.

This productive discussion has motivated astronomers to consider the more general and flexible class of models known as continuous-time auto-regressive moving average processes (Kelly et al. 2014). This class encompasses a wide variety of stochastic processes and requires users to select the orders of both the auto-regressive and the moving average components of the model. The Akaike information criterion (Akaike 1974) is a popular model selection criteria that can be used to select these orders (Kelly et al. 2014; Caceres et al. 2019), but may not exhibit well calibrated statistical properties (e.g., Sutherland et al. 2023). Meyer et al. (2023) illustrate an alternative strategy based on the Bayesian evidence. This is well-grounded theoretically, but results should be checked for sensitive to the choice of prior distribution. Both strategies aim to identify an appropriate level of model complexity and thus avoid both under- and overfitting.

Substantial community effort has also been devoted to investigating the applicability and physical interpretation of the resulting power spectral densities via empirical evidence (Moreno et al. 2019; Yu et al. 2022). This class of models is promising as it can be extended to model long-memory auto-correlations (Marquardt 2006; Marquardt & James 2007), although its limited applicability to stationary time series data remains.

This collective community effort is the key to building time-tested models with widely accepted astrophysical interpretations as it is crucial to demonstrate the empirical evidence and effectively warn of cases where the interpretation of the model parameters is fallible.

6. ALL STATISTICAL TESTS HAVE THRESHOLDS, BUT SOME ARE MIS-SET.

Hypothesis testing compares two models for the same data; the two models are called the null and the alternative hypotheses. In the standard frequentist setup, the researcher specifies a test statistic with known (or approximately known) distribution under the null hypothesis. Inconsistency between the observed value of test statistic (computed using the research data) and this null distribution is viewed as evidence that the data is unlikely to have arisen under the null model and thus as evidence in favor of the alternative hypothesis. Inconsistency is typically measured with a p-value, the probability that a value of test statistic as extreme or more extreme than the observed value would arise under the null distribution. The principled usage of this paradigm is crucial in various scientific fields because it is the procedure that provides data-driven evidence for a scientific discovery or anomaly against well-established theories. Unlike in biomedical research, hypothesis testing in astronomy is less likely to suffer from common issues regarding p-values, such as a blind usage of "p-value $<$ 0.05" or p-hacking, for example, collecting more data until the p-value becomes smaller than 0.05 (Wasserstein & Lazar 2016). This is partly because astronomers typically use more conservative thresholds for establishing statistical significance, such as a 3σ level (Vallisneri et al. 2023), making significance more difficult to achieve by simple data manipulation[7]. This 3σ threshold corresponds to type I error rate of α equal to 0.0027 (that is, the probability of rejecting the null when it is correct) for a two-sided test when a test statistic is distributed as $N(0, 1)$ under H_0.

A common application of significance testing in astronomy has been for the purpose of source detection. A "3σ" detection usually implies that the ratio of the estimated flux to its error is ≥ 3. It is worth noting that this ties "detectability" to flux estimation. Newer methods like *CIAO*/`wavdetect` (Freeman et al. 2002) explicitly separate detection from flux estimation, carrying out the former via a p-value threshold set based on the estimated background alone. Such methods have more power and improve the sensitivity of the detectors, enabling the detection of weaker sources[8]. Using only the background also allows a statistically well-defined definition of an *upper limit* to the source flux, which can be set as the source intensity that would be detectable with a specified power (Kashyap et al. 2010).

One aspect that astronomers must keep in mind when interpreting significance levels of multiple test statistics, however, is how to control the family-wise error rate (FWER). The FWER is defined as the probability of committing at least one type I error (false-positive) among m hypothesis tests, and is smaller than or equal to $1 - (1 - \alpha_{\text{ind}})^m$. The notation α_{ind} denotes the common type I error rate used for each of the m individual hypothesis tests. A good example to illustrate the FWER can be found in Abbott et al. (2016), where the detection of the first gravitational wave is based on the 4.6σ and 5.1σ significance levels of two test statistics, respectively. Clearly, each of the reported significance levels is greater than the 3σ threshold. However, naively comparing each reported significance level with the 3σ threshold is equivalent to maintaining an FWER that is less than or equal to 0.0054 ($= 1 - (1 - 0.0027)^2$). That is, the probability of committing at least one type I error in the two tests for the gravitational wave detection is actually twice as large as the individual type I error rate. To ensure the FWER is less than or equal to 0.0027, as intended, the popular Bonferroni correction (Armstrong 2014) sets the individual type I error rate to be 0.00135 ($= 0.0027/2$), which requires comparing each of the reported significance levels with a 3.2σ threshold, not 3σ.

One possible issue with the Bonferroni method is that it is rarely possible to reject a null hypothesis when the number of hypothesis tests m is large. For example, to ensure that the FWER is less than or equal to 0.0027 among 1,000 hypothesis tests, the individual type I error rate must be 2.7×10^{-6}, which is a threshold that may be difficult for individual p-value to achieve. This can lead to almost no rejection among the 1,000 hypothesis

[7] Early usage in astronomy tended to use the 2σ (5%) thresholds (see discussion in Wall 1979). Higher thresholds have also been adopted; e.g., the processing pipeline for the *Einstein* X-ray observatory used $4 - 5\sigma$ thresholds (Harnden et al. 1984).

[8] Consider a case where the expected background under the source is precisely estimated to be 0.9 counts. The detection threshold is set to the count where the cumulative tail probability of the Poisson distribution under the background only model drops below the 3σ significance, i.e, the detection threshold is 5 counts. The probability of observing 5 counts or more if there were no source is 0.00234, less than the adopted threshold of $p = 0.0027$. Thus if a count of five or more were observed the source would be declared detected. In contrast, the signal-to-noise ratio based on the estimated background-subtracted source strength and Gaussian error propagation approximations (Gehrels 1986) can only exceed 3 when > 15 counts are observed; the source would be declared detected if and only if its flux can be estimated with a sufficiently small uncertainty.

tests. As such, the Bonferroni correction can be too conservative for certain testing scenarios in astronomy, for example, for compiling astronomical catalogs.

An alternative is to instead control the false discovery rate (FDR, Benjamini & Hochberg 1995; Benjamini 2010). Unlike the FWER, the FDR ensures that the expected proportion of false-positives among all of the false- and true-positives is less than or equal to a preset value. That is, controlling the FDR to be less than or equal to 0.0027 means that the proportion of false-positives (false discoveries) among all of the rejected null hypotheses (discoveries) is less than or equal to 0.27%.

In practice, controlling the FDR results in a different procedure for rejecting the null hypothesis. For example, suppose we wish to ensure that the FDR is less than or equal to 0.0027 among 1000 hypothesis tests. We first sort the 1000 p-values in an increasing order, $p_1, p_2, \ldots, p_{1000}$, where p_1 is the smallest p-value and p_{1000} is the largest. Next, we find the largest index such that $p_i < (0.0027 \times i)/1000$. If this largest index were 55, for instance, then we would reject the null hypotheses associated with the first 55 tests. Note that the Bonferroni correction in this case is to reject the null hypothesis in test i if $p_i < 0.0027/1000$. Consequently, the FDR results in more rejections (possibly more false-positives, but also more true-positives) than the FWER. This is a useful feature of FDR when the number of tests m is large. Moreover, controlling FDR is known to be more powerful than controlling FWER, while the former comes with higher type I error rate (more false-positives) (Shaffer 1995). A receiver operating characteristic (ROC) curve can be useful for investigating the balance between the false-positive and true-positive rates obtained using different values of the FDR (or FWER) threshold.

In Abbott et al. (2016) for the detection of the first gravitational wave, the two p-values corresponding to the reported 4.6σ and 5.1σ significance levels are 4.2×10^{-6} and 3.4×10^{-7}, respectively. (We assume that test statistics are distributed as $N(0,1)$ under the null for two-sided tests.) Therefore, the largest index satisfying $p_i < (0.0027 \times i)/2$ is 2, leading to the rejection of the null hypothesis in both tests. Even though both FDR and FWER end up with the same rejection results in this example, it is worth noting that the former is ensuring that the FDR is less than or equal to 0.0027 while the latter is controlling the FWER.

7. ALL MODEL CHECKS CONSIDER VARIATIONS OF THE DATA, BUT SOME VARIANTS ARE MORE RELEVANT THAN OTHERS.

The notion of replicate data or repeated experiments is fundamental to frequentist statistical methods. In hypothesis testing, for example, evidence is quantified by (mathematically or numerically) computing the distribution of the test statistic that would result if multiple replicate data sets were generated under the null. In Bayesian data analysis, on the other hand, the posterior predictive distribution of additional data given the observed data is used to generate replicate data sets. In both cases, the replicate data represent the statistical variability and possible range of a test/summary statistic, and are used to quantify the expected deviation between the observed data and the null/posited model, thus enabling researchers to quantify uncertainty.

At first blush generating replicates may seem to be a well-stipulated proposition. In practice, however, researchers must consider how the replicates should be generated to be most comparable with their real data. For example, a researcher may only wish to consider replicate data with the same experimental conditions, instrumental effects, exposure time, and sample size as their real data. These are known quantities; varying them among the replicate data sets can make our uncertainty quantification less relevant for the actual uncertainty we care about. Conditioning on these factors reduces the variability of the replicate data sets and makes them more comparable with the real data. This in turn reduces uncertainty, error bars on fitted parameters, and the lengths of confidence intervals; similarly it increases the statistical power to distinguish between the null and the alternative in a hypothesis test.

Such considerations have led to the broad emphasis in statistics on conditioning as much as feasible; see Section 5.2 of Craiu et al. (2023) for a succinct overview. It has also led statistical theorists to consider if there is flexibility to condition on further attributes of the data in order to further increase statistical power. In a goodness-of-fit test, for example, the aim is to quantify the deviation between the observed data and the fitted model and to assess if it is greater than would be expect under the null. It seems entirely appropriate in this setting to only consider replicate data that have the same fitted model as the real data, e.g., by conditioning on the fitted model parameters[9]. Such a procedure is expected to reduce variability among the replicates (as

[9] Monte Carlo simulations in astronomy, for instance, obtain uncertainties of unknown parameters by generating replicated data sets given the maximum likelihood estimate computed on the observed data, fitting the model on each replicate set again, and quantifying the variations of the estimated parameters (e.g., Tewes et al. 2013).

they all have the same fitted parameters), make them more comparable with the real data, and increase statistical power. In this case, p-values are typically obtained via the parametric bootstrap (Efron 1985), where the estimated parameters are used as the "ground truth parameter" when generating replicate data sets.

Roe & Woodroofe (1999) consider the specific example of background contaminated Poisson counts. Letting the observed count $Y^{\rm obs}$ equal the sum of the unobserved source, Y_S, and background Y_B counts, Roe & Woodroofe (1999) make that astute observation that while Y_B is unknown, it is bounded by $Y^{\rm obs}$, i.e., we know $Y_B \leq Y^{\rm obs}$. By considering only replicates data with $Y_B^{\rm rep} \leq Y^{\rm obs}$, Roe & Woodroofe (1999) devise more coherent confidence intervals for the source intensity.

To give a concrete example of the advantage of conditional goodness-of-fit tests, we consider a Poisson model, where

$$N_i \overset{\rm indep}{\sim} \text{Poisson}(s_i(\boldsymbol{\theta})) \quad (1)$$

is the count in bin i for $i = 1, \ldots, 10$. The test compares a uniform null model, where all of the bins have the same expected count, λ, with the fully saturated alternative model where each bin has its own expected count, s_i:

$$H_0: \; s_i(\boldsymbol{\theta}) = \lambda$$
$$H_a: \; s_i(\boldsymbol{\theta}) = s_i.$$

We use the C-statistic (Cash 1979; Kaastra 2017), defined as minus twice the logarithm of the ratio of the likelihood under the null and that under the alternative, with both likelihoods evaluated at their respective maximum likelihood estimates. Specifically, the maximum likelihood estimate under the null is $\hat{\lambda} = \sum_{i=1}^{10} N_i/10$ and under the alternative is $\hat{s}_i = N_i$. We consider two null distributions for the C-statistic. The *unconditional null* resamples data according to the Poisson$(\hat{\lambda})$ distribution. The *conditional null*, on the other hand, conditions on the maximum likelihood estimate of λ which is equivalent to conditioning on the the total count, $\sum_{i=1}^{n} N_i$, resulting in resampling data from a multinomial distribution.

A simulation study demonstrates that the power of this goodness-of-fit test (i.e., the probability of correctly rejecting the null hypothesis) can increase by more than 30% when the significance level is set to 0.0027 (corresponding to the typical 3σ threshold in astronomy); see Appendix A for more details. Chen et al. (2024) provide a rigorous study of conditional and unconditional goodness-of-fit tests based on the C-statistic, under a general framework designed for realistic high-energy spectral models.

8. CONCLUDING REMARKS

Astronomical data are now being produced at an unprecedented rate and with increasing complexity, and even more large-scale telescopes are expected to come into operation soon. Even though the quantity and complexity of modern astronomical data naturally demand sophisticated statistical tools for various purposes, no single all-purpose statistical tool exists that can be deployed without careful consideration of its limitations and underlying assumptions. Rather state-of-the-art statistical methods require care, both in selecting an appropriate method and applying it properly. In some cases, existing methods do not suffice and new techniques must be developed. All together, this means that astronomers must be cognizant of the limitations and assumptions of the statistical and machine learning tools they employ, and must be cautious when using them.

We have proposed six statistical maxims to promote statistically sound data analytic practices and to improve the quality of scientific findings in astronomy. We hope that researchers are able to easily check these maxims as part of their daily data analytic routines. These maxims, however, are certainly not sufficient to solve all possible problem that might arise from the myriad of data types used in astronomical data analyses. For example, as a reviewer pointed out, one may question "whether our maxims, or any other broad-sweep approaches towards reliability of statistical conclusions, apply to machine learning methods that have either only an algorithmic foundation or can be viewed as models with vast number of parameters."

Whereas several of our maxims are generally applicable to any empirical studies, such as the first and second maxims, we echo this reviewer's call for someone to lead 'Six Maxims for Applying Machine Learning to Astronomy.' More broadly, we hope our work will encourage experts in other components of the data life cycle, such as data management and data visualization (see Wing 2019), to develop their own Six Maxims to benefit the astronomy community, as it continuously improves its ability to extract scientific value from complex astronomical data.

ACKNOWLEDGMENTS

This work was conducted under the auspices of the CHASC International Astrostatistics Center. CHASC is supported by NSF grants DMS-18-11308, DMS-18-11083, DMS-18-11661, DMS-21-13615, DMS-21-13397, and DMS-21-13605; by the UK Engineering and Physical Sciences Research Council [EP/W015080/1]; and by NASA APRA 80-NSSC21-K0285. We thank Eric Feigelson for initiating discussion and for providing insightful comments, which led to this work. We thank CHASC members for many helpful discussions, the graduate students who took the Stat 303 class at Harvard University in 2023 for reading and evaluating an earlier version of the manuscript, and two anonymous referees for their thoughtful comments. HT extends deep appreciation to Jogesh Babu for sharing his invaluable insights into the statistical challenges in astronomy and for offering his helpful guidance during the preparation of this manuscript. DvD's work was supported in part by a Marie Skłodowska-Curie RISE Grants (H2020-MSCA-RISE-2015-691164, H2020-MSCA-RISE-2019-873089) provided by the European Commission. VLK and AS acknowledge support from NASA Contract NAS8-03060 to the Chandra X-ray Center. KSM is supported by the European Union's Horizon 2020 research and innovation programme under ERC Grant Agreement No. 101002652 and Marie Skłodowska-Curie Grant Agreement No. 873089. YC acknowledges support from NASA 22-SWXC22_2-0005, NASA 22-SWXC22_2-0015, and NSF PHY 2027555.

APPENDIX

A. A SIMULATION STUDY: CONDITIONAL VS UNCONDITIONAL GOODNESS-OF-FIT TESTS

We conduct a simulation study to illustrate the advantage of the conditional test in terms of statistical power (i.e., the probability of correctly rejecting the null hypothesis) against a particular alternative model, namely the model in Eq. (1) with $s_i(\boldsymbol{\theta})$ varying linearly from 2 to 9 across the ten bins. We independently simulate 20,000 data sets under this alternative Poisson model. For each simulation, we compute $\hat{\lambda}^{(j)}$ for $j = 1, 2, \ldots, 20000$, and then simulate a further 5,000 data sets from each of (i) the unconditional null distribution (Poisson with $\hat{\lambda}^{(j)}$) and (ii) the conditional null distribution (multinomial with a total of $10 \times \hat{\lambda}^{(j)}$ counts). This allows us to numerically compute the p-value associated with the C-statistics from each of the 20,000 simulated data sets, and the power of the conditional and unconditional tests as the proportion of these p-values that are less than a given significance level (i.e., the probability of incorrectly rejecting the null).

The results are shown in Figure 7. The upper panel illustrates that the power obtained from the conditional test (denoted by the dashed curve) is uniformly greater than that of the unconditional test (represented by the solid curve). To emphasize this improvement, the percentage increase in power is displayed in the bottom panel. Although the percentage improvement decreases as the significance level increases, it remains at least 10% when the significance level is below 0.1. In particular, when the significance level is set to 0.0027 (corresponding to the typical 3σ threshold in astronomy), denoted by the dot-dashed vertical line, the percentage improvement from using the conditional null distribution exceeds 30%.

REFERENCES

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2016, PhRvL, 116, 061102, doi: 10.1103/PhysRevLett.116.061102

Aggarwal, A., Schanche, N., Reeves, K. K., Kempton, D., & Angryk, R. 2018, The Astrophysical Journal Supplement Series, 236, 15, doi: 10.3847/1538-4365/aab77f

Akaike, H. 1974, IEEE transactions on automatic control, 19, 716, doi: 10.1109/TAC.1974.1100705

Akritas, M. G., & Bershady, M. A. 1996, The Astrophysical Journal, 470, 706, doi: 10.1086/177901

Andrae, R., Kim, D.-W., & Bailer-Jones, C. A. L. 2013, A&A, 554, A137, doi: 10.1051/0004-6361/201321335

Andreon, S., & Hurn, M. 2013, Statistical Analysis and Data Mining: The ASA Data Science Journal, 6, 15, doi: 10.1002/sam.11173

Armstrong, R. A. 2014, Ophthalmic and Physiological Optics, 34, 502, doi: 10.1111/opo.12131

Autenrieth, M., van Dyk, D. A., Trotta, R., & Stenning, D. C. 2024, Statistical Analysis and Data Mining: The ASA Data Science Journal, 17, e11643, doi: https://doi.org/10.1002/sam.11643

Avni, Y. 1976, The Astrophysical Jorunal, 210, 642

Beaton, R. L., Freedman, W. L., Madore, B. F., et al. 2016, ApJ, 832, 210, doi: 10.3847/0004-637X/832/2/210

Benjamini, Y. 2010, Journal of the Royal Statistical Society Series B: Statistical Methodology, 72, 405, doi: 10.1111/j.1467-9868.2010.00746.x

Benjamini, Y., & Hochberg, Y. 1995, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289. http://www.jstor.org/stable/2346101

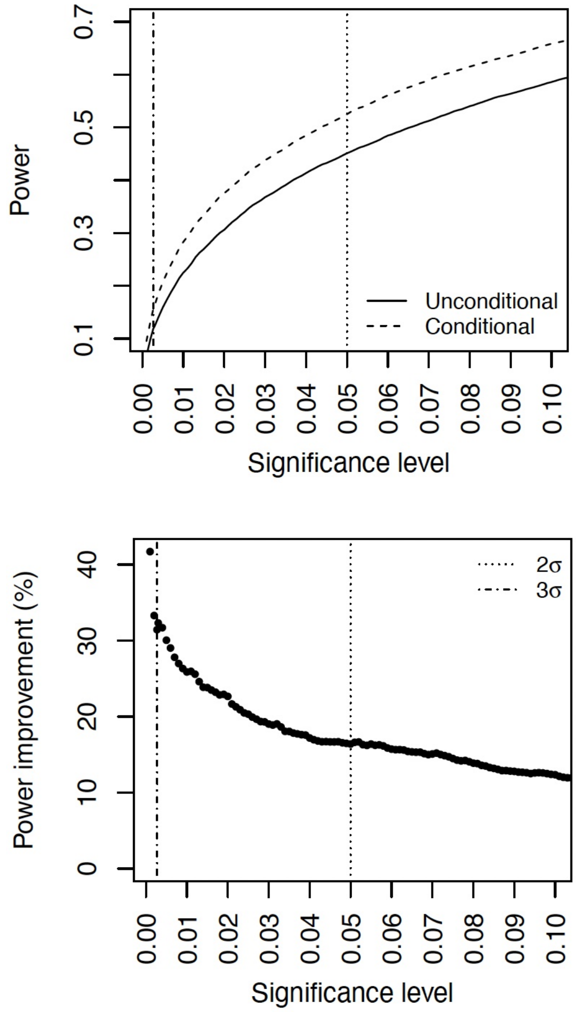

Figure 7. Upper panel: Comparison of the power (i.e., the probability of rejecting the null hypothesis when it is false) of a conditional (dashed curve) and unconditional (solid curve) goodness-of-fit test based on the C-statistic. We compare the uniform null model given in (1) with $s_i(\boldsymbol{\theta}) = \lambda$ to a particular alternative model with $s_i(\boldsymbol{\theta})$ varying linearly between 2 and 9. The power obtained from the conditional test is uniformly higher than that of the unconditional test. The dot-dashed vertical line indicates the significance level of the 3σ threshold (i.e., 0.0027), while the dotted vertical line represents the 2σ threshold (0.05). Bottom Panel: Percentage improvement in power as a function of the statistical significance level. With the 3σ threshold, indicated by the dot-dashed vertical line, the power improves by 31.4%. At the 2σ threshold, denoted by the dotted vertical line, the power improves by 16.4%.

Boller, T., Freyberg, M. J., Trümper, J., et al. 2016, A&A, 588, A103, doi: 10.1051/0004-6361/201525648

Bonamente, M. 2020, Journal of Applied Statistics, 47, 2044, doi: 10.1080/02664763.2019.1704703

Bovy, J., Hennawi, J. F., Hogg, D. W., et al. 2011, ApJ, 729, 141, doi: 10.1088/0004-637X/729/2/141

Box, G. E. P., & Draper, N. R. 1987, Empirical Model-Building and Response Surfaces, 1st edn. (John Wiley & Sons)

Brown, P. J. 1994, Measurement, Regression, and Calibration (Oxford University Press)

Brown, P. J. 2018, Journal of the Royal Statistical Society: Series B (Methodological), 44, 287, doi: 10.1111/j.2517-6161.1982.tb01209.x

Budavári, T., & Szalay, A. S. 2008, ApJ, 679, 301, doi: 10.1086/587156

Bulbul, E., Markevitch, M., Foster, A., et al. 2014, ApJ, 789, 13, doi: 10.1088/0004-637X/789/1/13

Caceres, G. A., Feigelson, E. D., Jogesh Babu, G., et al. 2019, AJ, 158, 57, doi: 10.3847/1538-3881/ab26b8

Cash, W. 1979, The Astrophysical Journal, 228, 939, doi: 10.1086/156922

Chen, Y., Li, X., Meng, X., et al. 2024, in preparation

Chen, Y., Meng, X.-L., Wang, X., et al. 2019a, Journal of the American Statistical Association, 114, 1018, doi: 10.1080/01621459.2018.1528978

—. 2019b, Journal of the American Statistical Association, 114, 1018, doi: 10.1080/01621459.2018.1528978

Craiu, R. V., Gong, R., & Meng, X.-L. 2023, Annual Review of Statistics and Its Application, 10, 699, doi: 10.1146/annurev-statistics-040220-015348

de Vaucouleurs, G., & Bollinger, G. 1979, ApJ, 233, 433, doi: 10.1086/157405

de Vaucouleurs, G., & Peters, W. L. 1986, ApJ, 303, 19, doi: 10.1086/164048

Eadie, G. M., Speagle, J. S., Cisewski-Kehe, J., et al. 2023, arXiv e-prints, arXiv:2302.04703, doi: 10.48550/arXiv.2302.04703

Efron, B. 1985, Biometrika, 72, 45

Efstathiou, G. 2020, arXiv e-prints, arXiv:2007.10716. https://arxiv.org/abs/2007.10716

Evans, I. N., Primini, F. A., Glotfelty, K. J., et al. 2010, ApJS, 189, 37, doi: 10.1088/0067-0049/189/1/37

Feeney, S. M., Mortlock, D. J., & Dalmasso, N. 2018, MNRAS, 476, 3861, doi: 10.1093/mnras/sty418

Feigelson, E., & Babu, G. 1998, in Symposium-International Astronomical Union, Vol. 179, Cambridge University Press, 363–370, doi: 10.1017/S0074180900129043

Feigelson, E. D., & Babu, G. J. 2012, arXiv preprint. https://arxiv.org/abs/1205.2064

Feigelson, E. D., & Babu, G. J. 2012, Modern Statistical Methods for Astronomy: With R Applications (Cambridge University Press), doi: 10.1017/CBO9781139015653

Feigelson, E. D., De Souza, R. S., Ishida, E. E., & Babu, G. J. 2021, Annual Review of Statistics and Its Application, 8, 493, doi: 10.1146/annurev-statistics-042720-112045

Feigelson, E. D., & Nelson, P. I. 1985, ApJ, 293, 192, doi: 10.1086/163225

Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q. 2002, ApJS, 138, 185, doi: 10.1086/324017

Fuller, W. A. 1987, Measurement Error Models (New York, NY: John Wiley & Sons, Inc.), doi: 10.1002/9780470316665

Gehrels, N. 1986, ApJ, 303, 336, doi: 10.1086/164079

Gelman, A. 1996, Statistica Sinica, 6, 215. http://www.jstor.org/stable/24306008

Gelman, A., Carlin, J. B., Stern, H. S., et al. 2013, Bayesian Data Analysis (Boca Raton, FL, USA: CRC Press)

Gelman, A., Simpson, D., & Betancourt, M. 2017, Entropy, 19, doi: 10.3390/e19100555

Graham, M. J., Djorgovski, S. G., Drake, A. J., et al. 2014, Monthly Notices of the Royal Astronomical Society, 439, 703, doi: 10.1093/mnras/stt2499

Greenwood, P. E., & Nikulin, M. S. 1996, A Guide to Chi-Squared Testing (New York, NY: John Wiley & Sons, Inc.)

Guainazzi, M., David, L., Grant, C. E., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 047001, doi: 10.1117/1.JATIS.1.4.047001

Hainline, L. J., Morgan, C. W., Kochanek, C. S., et al. 2012, in American Astronomical Society Meeting Abstracts, Vol. 219, American Astronomical Society Meeting Abstracts #219, 108.02

Harnden, F. R., J., Fabricant, D. G., Harris, D. E., & Schwarz, J. 1984, SAO Special Report, 393

Hilbe, J. M. 2014, Modeling Count Data (Cambridge University Press), doi: 10.1017/CBO9781139236065

Hu, Z., & Tak, H. 2020, The Astronomical Journal, 160, 265, doi: 10.3847/1538-3881/abc1e2

Hubble, E. 1929, Contributions from the Mount Wilson Observatory, 3, 23

Humphrey, P. J., Liu, W., & Buote, D. A. 2009, ApJ, 693, 822, doi: 10.1088/0004-637X/693/1/822

Isobe, T., Feigelson, E. D., & Nelson, P. I. 1986, The Astrophysical Journal, 306, 490, doi: 10.1086/164359

Izbicki, R., Lee, A. B., & Freeman, P. E. 2017, The Annals of Applied Statistics, 11, 698 , doi: 10.1214/16-AOAS1013

Jurić, M., Axelrod, T., Becker, A., et al. 2023, Data Products Definition Document, Vera C. Rubin Observatory. https://lse-163.lsst.io/

Kaastra, J. 2017, Astronomy & Astrophysics, 605, A51

Kashyap, V. L., van Dyk, D. A., Connors, A., et al. 2010, ApJ, 719, 900, doi: 10.1088/0004-637X/719/1/900

Kasliwal, V. P., Vogeley, M. S., & Richards, G. T. 2015, Monthly Notices of the Royal Astronomical Society, 451, 4328, doi: 10.1093/mnras/stv1230

Kass, R. E., & Raftery, A. E. 1995, Journal of the American Statistical Association, 90, 773, doi: 10.1080/01621459.1995.10476572

Kelly, B. C. 2007, ApJ, 665, 1489, doi: 10.1086/519947

Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, The Astrophysical Journal, 698, 895, doi: 10.1088/0004-637x/698/1/895

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., & Uttley, P. 2014, The Astrophysical Journal, 788, 33, doi: 10.1088/0004-637X/788/1/33

Kelly, B. C., Shetty, R., Stutz, A. M., et al. 2012, ApJ, 752, 55, doi: 10.1088/0004-637X/752/1/55

Kim, D.-W., Protopapas, P., Trichas, M., et al. 2012, The Astrophysical Journal, 747, 107, doi: 10.1088/0004-637x/747/2/107

Kozłowski, S. 2016, Monthly Notices of the Royal Astronomical Society, 459, 2787, doi: 10.1093/mnras/stw819

Kozłowski, S., Kochanek, C. S., Udalski, A., et al. 2010, The Astrophysical Journal, 708, 927

Kubokawa, T., & Robert, C. P. 1994, Journal of Multivariate Analysis, 51, 178

Landy, S. D., & Szalay, A. S. 1992, ApJ, 391, 494, doi: 10.1086/171365

Lee, H., Kashyap, V. L., van Dyk, D. A., et al. 2011, ApJ, 731, 126, doi: 10.1088/0004-637X/731/2/126

Lindley, D. V. 1985, Making Decisions (New York, NY: Wiley)

MacLeod, C., Ivezić, Ž., Kochanek, C., et al. 2010, The Astrophysical Journal, 721, 1014

Mandel, K. S., Scolnic, D. M., Shariff, H., Foley, R. J., & Kirshner, R. P. 2017, ApJ, 842, 93, doi: 10.3847/1538-4357/aa6038

Marquardt, T. 2006, Bernoulli, 12, 1099 , doi: 10.3150/bj/1165269152

Marquardt, T., & James, L. F. 2007, Technical report

Marshall, H. L. 2021, AJ, 162, 134, doi: 10.3847/1538-3881/ac173d

Meyer, A. D., van Dyk, D. A., Tak, H., & Siemiginowska, A. 2023, ApJ, 950, 37, doi: 10.3847/1538-4357/acbea1

Moreno, J., Vogeley, M. S., Richards, G. T., & Yu, W. 2019, PASP, 131, 063001, doi: 10.1088/1538-3873/ab1597

Mushotzky, R. F., Edelson, R., Baumgartner, W., & Gand hi, P. 2011, The Astrophysical Journal Letter, 743, L12, doi: 10.1088/2041-8205/743/1/L12

Oman, S. D., & Srivastava, M. 1996, Scandinavian journal of statistics, 473

Osborne, C. 1991, International Statistical Review, 59, 309, doi: 10.2307/1403690

Partridge, B., López-Caniego, M., Perley, R. A., et al. 2016, The Astrophysical Journal, 821, 61, doi: 10.3847/0004-637x/821/1/61

Payne, E., Talbot, C., Lasky, P. D., Thrane, E., & Kissel, J. S. 2020, Phys. Rev. D, 102, 122004, doi: 10.1103/PhysRevD.102.122004

Pierson, S. 2013, AMSTATNEWS. https://magazine.amstat.org/blog/2013/12/01/science-policy-intel/

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13, doi: 10.1051/0004-6361/201525830

Portillo, S. K. N., Lee, B. C. G., Daylan, T., & Finkbeiner, D. P. 2017, AJ, 154, 132, doi: 10.3847/1538-3881/aa8565

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2002, ApJ, 571, 545, doi: 10.1086/339856

Reeves, J. N., O'Brien, P. T., Braito, V., et al. 2009, ApJ, 701, 493, doi: 10.1088/0004-637X/701/1/493

Revsbech, E. A., Trotta, R., & van Dyk, D. A. 2017, Monthly Notices of the Royal Astronomical Society, 473, 3969, doi: 10.1093/mnras/stx2570

Riess, A. G., Casertano, S., Yuan, W., et al. 2021, The Astrophysical Journal, 908, L6, doi: 10.3847/2041-8213/abdbaf

Riess, A. G., Macri, L. M., Hoffmann, S. L., et al. 2016, ApJ, 826, 56, doi: 10.3847/0004-637X/826/1/56

Rivers, D. L. 2014, Dynamic Bayesian Approaches to the Statistical Calibration Problem (Virginia Commonwealth University)

Roe, B. P., & Woodroofe, M. B. 1999, Phys. Rev. D, 60, 053009, doi: 10.1103/PhysRevD.60.053009

Rots, A. H., Burke, D. J., Civano, F., Hain, R., & Nguyen, D. 2017a, in American Astronomical Society Meeting Abstracts, Vol. 229, American Astronomical Society Meeting Abstracts #229, 156.03

Rots, A. H., Nguyen, D., Budavari, T., et al. 2017b, in AAS/High Energy Astrophysics Division, Vol. 16, AAS/High Energy Astrophysics Division #16, 113.01

Ryan, D. F., Milligan, R. O., Gallagher, P. T., et al. 2012, The Astrophysical Journal Supplement Series, 202, 11, doi: 10.1088/0067-0049/202/2/11

Sandage, A., & Tammann, G. A. 1975, ApJ, 197, 265, doi: 10.1086/153510

Schafer, D. W., & Purdy, K. G. 1996, Biometrika, 83, 813

Schild, R. E. 1990, AJ, 100, 1771, doi: 10.1086/115634

Sereno, M. 2016, Monthly Notices of the Royal Astronomical Society, 455, 2149, doi: 10.1093/mnras/stv2374

Shaffer, J. P. 1995, Annual Review of Psychology, 46, 561, doi: https://doi.org/10.1146/annurev.ps.46.020195.003021

Shalyapin, V. N., Goicoechea, L. J., Koptelova, E., Ullán, A., & Gil-Merino, R. 2008, A&A, 492, 401, doi: 10.1051/0004-6361:200810447

Shy, S., Tak, H., Feigelson, E. D., Timlin, J. D., & Babu, G. J. 2022, AJ, 164, 6, doi: 10.3847/1538-3881/ac6e64

Siemiginowska, A., Eadie, G., Czekala, I., et al. 2019, BAAS, 51, 355. https://arxiv.org/abs/1903.06796

Sutherland, C., Hare, D., Johnson, P. J., et al. 2023, Proceedings of the Royal Society B, 290, 20231261

Tak, H., Ellis, J. A., & Ghosh, S. K. 2019, Journal of Computational and Graphical Statistics, 28, 415, doi: 10.1080/10618600.2018.1537925

Tak, H., Ghosh, S. K., & Ellis, J. A. 2018a, Monthly Notices of the Royal Astronomical Society, 481, 277, doi: 10.1093/mnras/sty2326

Tak, H., Kelly, J., & Morris, C. 2017a, Journal of Statistical Software, 78, 1, doi: 10.18637/jss.v078.i05

Tak, H., Mandel, K., van Dyk, D. A., et al. 2017b, The Annals of Applied Statistics, 11, 1309, doi: 10.1214/17-AOAS1027

Tak, H., Meng, X.-L., & van Dyk, D. A. 2018b, Journal of Computational and Graphical Statistics, 27, 479, doi: 10.1080/10618600.2017.1415911

Tanaka, Y., Nandra, K., Fabian, A. C., et al. 1995, Nature, 375, 659, doi: 10.1038/375659a0

Teerikorpi, P. 2015, A&A, 576, A75, doi: 10.1051/0004-6361/201425489

Tewes, M., Courbin, F., Meylan, G., et al. 2013, Astronomy & Astrophysics, 556, A22, doi: 10.1051/0004-6361/201220352

Vallisneri, M., Meyers, P. M., Chatziioannou, K., & Chua, A. J. K. 2023, PhRvD, 108, 123007, doi: 10.1103/PhysRevD.108.123007

van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, ApJ, 548, 224, doi: 10.1086/318656

Verde, L., Treu, T., & Riess, A. G. 2019, Nature Astronomy, 3, 891, doi: 10.1038/s41550-019-0902-0

Villanueva, G. L., Cordiner, M., Irwin, P. G. J., et al. 2021, Nature Astronomy, 5, 631, doi: 10.1038/s41550-021-01422-z

Voges, W. 1993, Advances in Space Research, 13, 391, doi: 10.1016/0273-1177(93)90147-4

Voges, W., Aschenbach, B., Boller, T., et al. 1999, A&A, 349, 389. https://arxiv.org/abs/astro-ph/9909315

Wall, J. V. 1979, QJRAS, 20, 138

Wang, Q., & Taylor, S. R. 2022, MNRAS, 516, 5874, doi: 10.1093/mnras/stac2679

Wasserstein, R. L., & Lazar, N. A. 2016, The American Statistician, 70, 129, doi: 10.1080/00031305.2016.1154108

Wilks, S. S. 1938, The annals of mathematical statistics, 9, 60, doi: 10.1214/aoms/1177732360

Wing, J. M. 2019, Harvard Data Science Review, 1, doi: 10.1162/99608f92.e26845b4

Xu, J., van Dyk, D. A., Kashyap, V. L., et al. 2014, ApJ, 794, 97, doi: 10.1088/0004-637X/794/2/97

Yu, W., Richards, G. T., Vogeley, M. S., Moreno, J., & Graham, M. J. 2022, ApJ, 936, 132, doi: 10.3847/1538-4357/ac8351

Zu, Y., Kochanek, C. S., Kozłowski, s., & Udalski, A. 2013, The Astrophysical Journal, 765, 106, doi: 10.1088/0004-637X/765/2/106