Editorial

A Constellation of Statistical Analyses

Nancy D. Morrison
Editor-in-Chief, *Journal of the AAVSO*

Department of Physics and Astronomy and Ritter Observatory, MS 113, The University of Toledo, 2801 W. Bancroft Street, Toledo OH 43606; jaavso.editor@aavso.org

Received June 9, 2022

Statistics is at the heart of variable star research. Whenever we determine a parameter of a variable star—such as its period, amplitude of variation, or time of maximum or minimum light—we use a mathematical model or a statistical analysis. Those operations return an estimate of a value (point estimate) and an estimate of its uncertainty (sometimes expressed as a confidence interval). Such a test could be as simple as the classical eyeball estimate or as complex as a wavelet analysis. Often, we accept the results of the statistical test at face value and move on to the next problem.

But how does a researcher know that the error estimate returned by a statistical test is realistic? Or even that the estimated value, with its error bars or its confidence interval, is reliable? This question has special urgency if the time series has major gaps or a low signal-to-noise ratio.

Recently, researchers in various fields—ranging from psyhchology to finance—have tried to tackle this question by crowdsourcing data analysis projects. They invited other teams of researchers to analyze a given data set, each using a different methodology. Then the project leaders compiled the results, viewing the collection of results as an approximation to the universe of possible estimates of the statistics of the problem. In an open-access comment in the 17 May issue of *Nature*, Wagenmakers *et al.* (2022) discussed the results of about a dozen formal crowdsourced projects, some of them involving over 100 independent teams. In many of those studies, some of the results by individual teams showed a range characterized by error bars that do not overlap.

I chose one of these crowdsourced studies to explore in more detail: Silberzahn *et al.* (2018). Through online advertising, those authors recruited 29 independent teams to analyze a data set on red cards given to soccer players. They aimed to test the hypothesis that dark-skinned players receive red cards more often than light-skinned players, averaging over many game situations and types of infraction.

Red cards are given for egregious bad or aggressive behavior and generally result in the player's ejection from the game. Although objective criteria for their award exist, marginal cases occur often, and the referee's judgement is important. The possible confounding variables (factors to be controlled for) are too numerous to list here. Many different assumptions about independence among variables and about systematic effects could be made. One important decision area was classification of skin tone as "light" or "dark;" Silberzahn *et al.* (2018) discuss this topic in detail. In keeping with all these complexities, the data analysis techniques differed greatly among the research teams.

In Figure 1, the studies are grouped by general methodology. The error bars are 95% confidence intervals, which are larger than the standard deviation usually used in astronomy (roughly a two-thirds confidence interval). The listing of the statistical methods is included here only to illustrate the great range of methodology involved. In statistics, the term "odds ratio" has the obvious meaning: in this case, it indicates the ratio of probability of a dark-skinned player receiving a red card, compared to that for a light-skinned player. For example, an odds ratio of 1.3 would mean that a dark-skinned player would be 1.3 times as likely to receive a red card, overall, as a light-skinned player.

It is clear from Figure 1 that the results range more widely than a single result with error bar would imply. Even though the confidence intervals mostly overlap, qualitatively different conclusions could be reached if individual results were considered in isolation. Although some of the teams' error bars give a fair representation of the overall range in the results, a few of them are exceptionally small.

How might these results apply to variable star astronomy? At first sight, our data sets are simpler than the one analyzed here. Silberzahn *et al.* (2018) admit that discrepant results are less likely in simpler problems with few measured variables, but that, even in such cases, analytical decisions may influence outcomes. Variable outcomes are more likely in case of complex data sets. Interestingly, one of those authors' examples of complex data is a longitudinal data set with missing data—exactly the case in astronomical time series.

What options are available to the individual researcher who is concerned about these problems? Silberzahn *et al.* (2018) make several suggestions. One is to crowdsource your own project: invite other researchers to analyze a well-specified data set with their own favorite methods. Those authors consider this approach to be inefficient; they spent a lot of energy on organizing their project. But they gave the analysis teams the opportunity to interact on analytical issues (without knowing each other's results), and the interaction was highly beneficial.

Another option is to re-analyze already-published data with a different technique. However, this route is subject to publication bias (Silberzahn et al. 2018 again): researchers doing the re-analysis are most likely to move to publication if

their result disagrees with the original one. Here I can say that *JAAVSO* stands willing to publish re-analyses of published data, even if agreement with the original result is perfect, provided that the analytical methods are truly independent.

There are other options for sampling the universe of analytical methods. In multiverse analysis (Steegen *et al.* 2016), researchers vary the construction of the data set in all the ways they can think of and then perform a similar statistical analysis on all the versions of the analyzed data.

In the realm of statistical analysis, as opposed to data set construction, is specification curve analylsis (Simonsohn *et al.* 2020). Here a given data set is analyzed with all reasonable methods and under all reasonable sets of assumptions. As in Figure 1, the method leads to a plot of results as a function of the assumptions and analysis techniques. Cosme (2022) provides additional explanation.

In astronomy, to my knowledge, researchers employ only one or a few analytical methods to address a given research problem. If any readers know of an example in astronomy of a paper comparing results of more than three analytical methods, I would appreciate learning about it.

Meanwhile, I encourage readers to explore the uncertainties in their analyses by means of one of the techniques outlined here. If crowdsourcing, specification curve analysis, and multiverse analysis are beyond your reach, and you have access to only one or two independent analysis methods, it would be appropriate to approach your results with caution.

References

Cosme, D. 2022.[1]

Silberzahn, R., *et al.* 2018, *Adv. Methods Pr. Psychological Sci.*, **1**, 337 (https://doi.org/10.1177%2F2515245917747646).

Simonsohn, U., Simmons, J. P., and Nelson, L. D. 2020, *Nature Human Behavior*, **4**, 1208 (https://doi.org/10.1038/s41562-020-0912-z).

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel. W. 2016, *Perspectives on Psychological Sci.*, **11**, 702.

Wagenmakers, E.-J., Sarafoglou, A., and Aczel, B. 2022, *Nature*, **605**, 423 (https://doi.org/10.1038/d41586-022-01332-8).

Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93

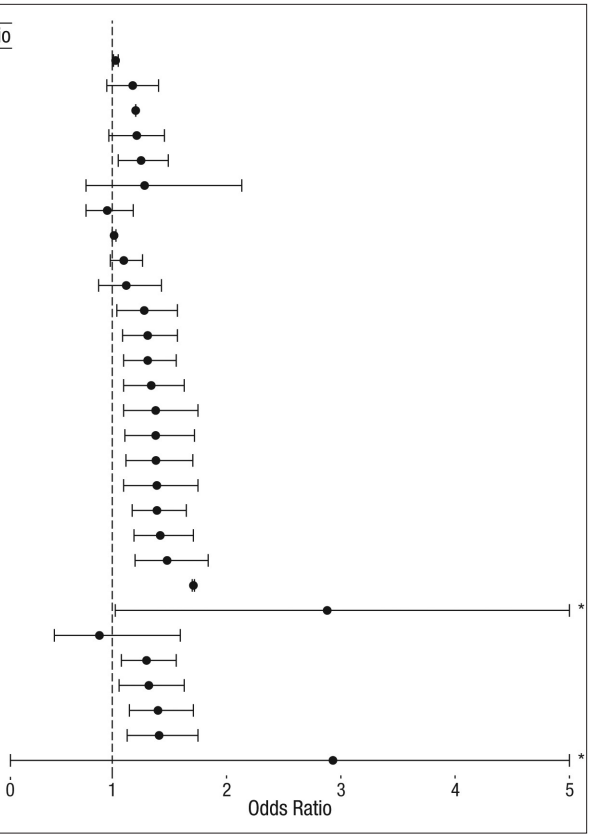

Figure 1. Reproduction of Figure 3 from Silberzahn *et al*. (2018). "Odds ratio" indicates the probability of a dark-skinned player receiving a red card, divided by that for a light-skinned player. A point estimate and 95% confidence interval are shown from each of the 29 analysis teams, with similar analyses grouped together. Asterisked error bars are truncated on the right-hand side for better readability of the graph. OLS = ordinary least squares; WLS = weighted least squares; Misc = miscellaneous.

[1] https://dcosme.github.io/specification-curves/SCA_tutorial_inferential_presentation#1