Statistical Maxims for Sound Astronomical Data Analysis

Hyungsuk (Tak) Tak

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

Aug 2, 2022

Joint work with Vinay Kashyap, Kaisey Mandel, Xiao-Li Meng, Aneta Siemiginowska, and David van Dyk.

Background

The Statistical Editor of the American Astronomical Society (AAS), Eric Feigelson, emailed Jogesh Babu, David van Dyk, and me on Jan 21, 2019.

We were asked to write an article about standard guidelines for more principled data analyses in the AAS publications.

Thus, we have come up with 8 maxims each in the sprint of George Box's famous aphorism,

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

List of Maxims

1. Data Collection: All data have stories behind them, but some stories are mistold.
2. Processing: All data are messy, but some are more easily cleaned.
3. Modeling: All models are a simplification, but some are more justified.
4. Assumptions: All assumptions are fallible, but some are more credible.
5. Methods: All methods have their purpose, but some are more versatile.
6. Model Checking: All models require assumptions, but some assumptions are more easily checked.
7. Computation: All computations are vulnerable to error, but some are more resilient.
8. Interpretation: All results are subject to interpretation, but some interpretations are less contrived.

Maxim 1 in Data Collection

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

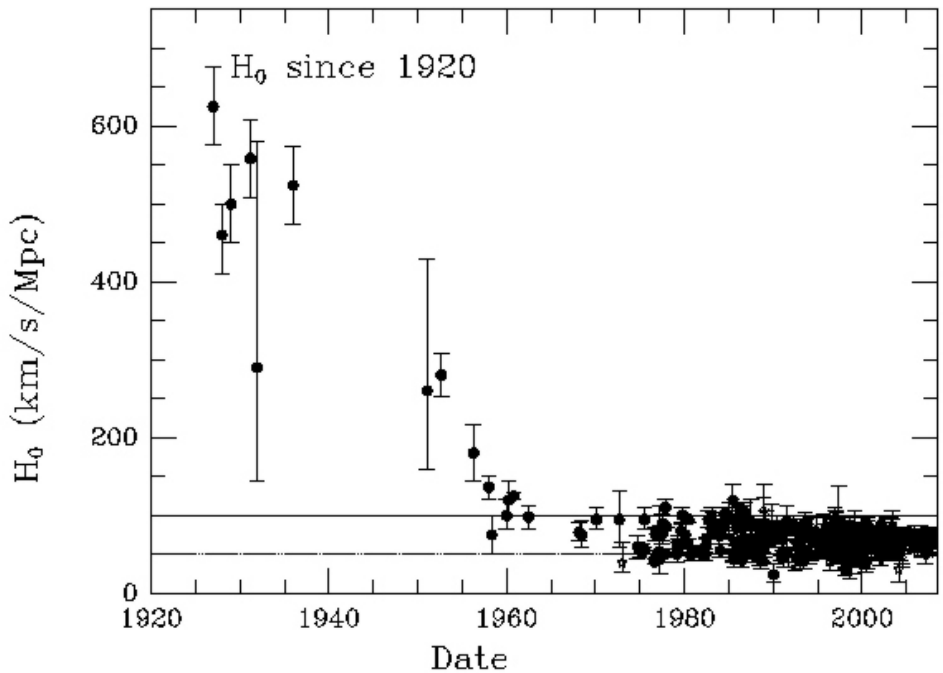
1. Non-uniform coverage: Astro. data are hard to be a random sample.
   - None of the so-called all-sky surveys (SDSS, Rubin Observatory, TESS, eROSITA, etc.) have uniform coverage.
   - Often closer or brighter objects are found in greater abundance in the survey censuses

Using such data without paying attention to the exact nature of how the populations are represented results in incorrect inferences (Kelly 2007).

We caution that the systematics of any survey or measurement must be carefully considered on a case-by-case basis.

Maxim 1 in Data Collection (cont.)

Example: The Hubble tension.



Advances in instrumentation and techniques have reduced systematics of the data measurements over the time.

Maxim 1 in Data Collection (cont.)

2. Selection bias: Data are often obtained purposefully via proposals.
   - Such data become public through archives.
   - Researchers download, use, and sometimes merge them as if the data were randomly and uniformly selected.

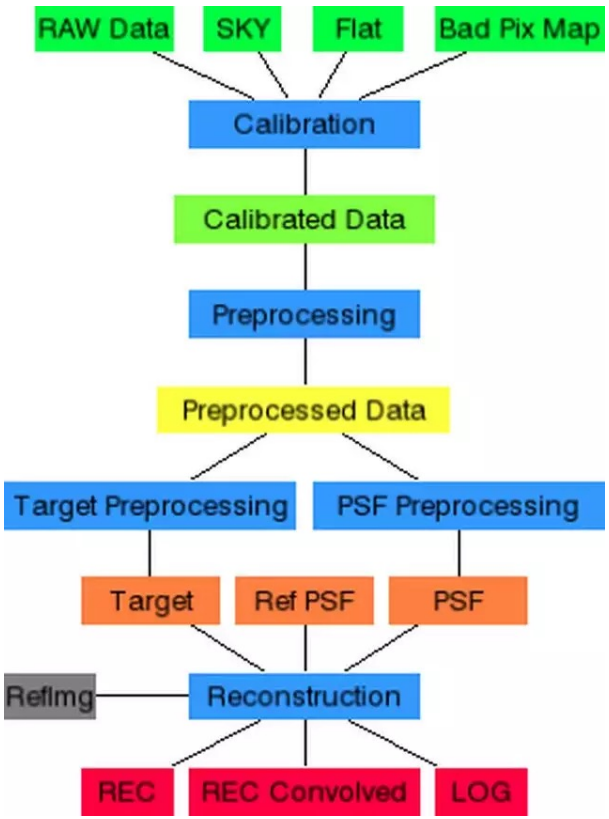
When multiple catalog data sets are merged, individual characteristics can affect overall interpretation in complex ways (Budavari & Szalay 2008).

We provide an example of dealing with the selection bias of a training set in classifying Type Ia supernovae (Revsbech et al. 2017; Autenrieth et al. 2021).

Maxim 2 in Processing

"All data are messy, but some are more easily cleaned."

Most astronomical data are pre-processed via the multi-stage software pipelines specific to a given telescope. The following is from LINC-NIRVANA Data Reduction Software.



Maxim 2 in Processing (cont.)

Unfortunately, this pre-processing is often ignored even though the pre-processing steps can reveal evidence of potential systematic errors, outliers, censoring, etc.

We suggest incorporating pre-processing procedures into a model as much as possible (Portillo et al. 2017).

Maxim 3 in Modeling

"All models are a simplification, but some are more justified."

A model is a parsimonious summary of a complicated real phenomenon.

Due to this nature, statistical bias is inevitable under any model.

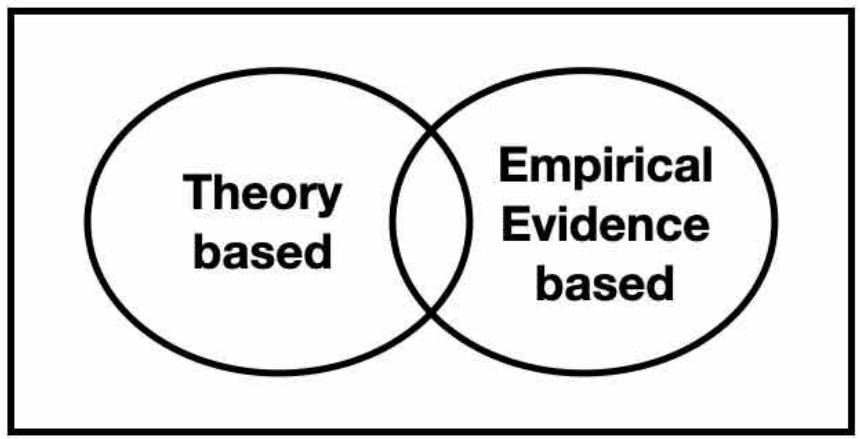
Bias-variance trade-off in machine learning.

Occam's Razor in statistics: Less is more!

Maxim 4 in Assumptions

"All assumptions are fallible, but some are more credible."

Not all assumptions are equally realistic, even though researchers can make their assumptions to suit their own purposes.



Examples include gravitational waves, distance measure via supernovae / the Big Bang theory, ΛCDM / power-law distribution, Poisson distribution, parametric models / subjective assumptions, Uniform priors.

Maxim 5 in Methods

"All methods have their purpose, but some are more versatile."

In many cases, statistical methods are developed for specific purposes or motivated by particular problems. Some of them can be applicable to other cases across various disciplines.

Statistics $\rightarrow$ astronomy: Survival analysis, posterior predictive p-value, measurement error, etc.

Astronomy $\rightarrow$ statistics: RAM algorithm, partially collapsed Gibbs sampler, ASIS, etc.

Thus, it is important for researchers to understand the fundamental ideas behind data analytic tools so that they can choose appropriate tools for their purposes.

Maxim 6 in Model Checking

"All models require assumptions, but some assumptions are more easily checked."

Like the two sides of a coin, model fitting and checking are inseparable, one feeding into the other in the data analytic cycle.

Avoid "letting the tail wag the dog" (choosing the model that best supports one's preferred scientific conclusion) via well-known model checking procedures.

Examples include residual analysis in regression, simulation-based checking procedures such as posterior predictive p-value/check or sensitivity analysis in Bayesian analysis.

Maxim 7 in Computation

"All computations are vulnerable to error, but some are more resilient."

Triple check numerical routines and computer codes.

For example, debugging via SNoTE, using different initial values on iterative procedures (optimization/sampling), multiple chains in MCMC.

In particular, sensitivity to initial values may indicate multiple possibilities in a non-convex parameter space.

Global optimization methods / multimodal samplers may help.

Maxim 8 in Interpretation

"All results are subject to interpretation, but some interpretations are less contrived."

It is essential to understand statistical concepts and methods well enough to ensure that the resulting scientific interpretation is valid.

For example, confidence and credible intervals have different sources of uncertainty.

We also emphasize that potential misinterpretation of p-values or statistical significance must be cautioned against.

Understanding the statistical/mathematical interpretation of analysis results related to physics is highly desirable (e.g., physical interpretations of damped random walk model parameters).