Comments and discussion points on today's talks

Main theme: "machines" can free us from strictures of parametric methods, but how many of the old issues remain?

Bob Cousins

Univ. of California, Los Angeles

PhyStat-Anomlies

24 May 2022

Constructing alternative H from the data

For me it is interesting to consider Andrea's talk as another step in the history of GOF methods based on the likelihood ratio of:

H_0: specified model to be tested, vs

H_1: an alternative model *constructed from the data.*

Oldest example I know is J. Neyman and E. Pearson (1928):

Histogram with:
bins labeled by i,
observed integer bin contents $\{d_i\}$,
mean contents predicted by model $\{y_i\}$.

NP considered case of total contents fixed ($\Rightarrow$ multinomial), whereas we typically have indep. Poisson model for each bin.

Both cases reviewed by Baker and Cousins, NIM 221 (1984) 437.

Poisson case: For alternative model, take any parametric model with as many indep parameters as there are histogram bins (saturated model). Then can get perfect fit to data, so that under H_1, estimate $\{y_i\} = \{d_i\}$. Then,

$$\mathcal{L}(H_0) = \prod_i \frac{y_i^{d_i} \exp(-y_i)}{d_i!} \qquad \mathcal{L}(H_1) = \prod_i \frac{d_i^{d_i} \exp(-d_i)}{d_i!}$$

$\lambda = \mathcal{L}(H_0)/\mathcal{L}(H_1)$ **, and**

$$-2 \ln \lambda = 2 \sum_i y_i - d_i + d_i \ln(d_i/y_i)$$

is approximately distributed as χ^2 .

(See Eqn. 40.13, https://pdg.lbl.gov/2021/reviews/rpp2021-rev-statistics.pdf)

Like all GOF tests, power depends on what the (unknown) true distribution is in nature.

See Baker & Cousins for advantages over Pearson χ^2 .

The usual χ^2 GOF for Gaussian uncertainties in intro lab courses can also be viewed as a LR with respect to saturated model:

We have "data points" $\{d_i\}$ with uncertainty $\{\sigma_i\}$ measured as a function of some control variable (say current measured as function of applied voltage). H_0 again specifies "true values" $\{y_i\}$.

$$\mathcal{L}(H_0) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-(d_i - y_i)^2/2\sigma_i^2\right)$$

Again, a saturated model for H_1 can be constructed to give a perfect fit to the data with estimates $\{y_i\} = \{d_i\}$, so

$$\mathcal{L}(H_1) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}}.$$

$$\lambda = \mathcal{L}(H_0)/\mathcal{L}(H_1) = \prod_i \exp\left(-(d_i - y_i)^2/2\sigma_i^2\right),$$

$$\boxed{\chi^2 = -2\ln\lambda = \sum_i \frac{(d_i - y_i)^2}{\sigma_i^2}.} \quad (\neq -2\ln\mathcal{L}(H_0))$$

Note: d.o.f. for Wilks = # params in saturated – # params in H_0

$\Rightarrow$ familiar rule: d.o.f. for χ^2 = # data points – # params in H_0

Note the crucial role of the denominator in the likelihood *ratio*.

It provides "best possible likelihood" as reference for $\mathcal{L}(H_0)$.

More fundamentally, for continuous data, it makes the result independent of the metric used for the data: Jacobians in numerator and denominator cancel. Without denominator, you can get any answer you want by transforming data metric.

(As Fred James pointed out years ago, there is a metric where data is uniform on (0,1), so likelihood is unity for all data sets.)

There was a period in HEP in which some people used likelihoods (numerator only) without ratio as GOF statistic; this was quite properly abandoned. See my Durham 2002 summary, http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/cousins.pdf, Sec. 3.7); also my lectures https://arxiv.org/abs/1807.05996, Sec. 4.1 ; also Ben's mention and pointer that "rare is not invariant".

The point of my saying all this is to remind us that we have a *lot* of experience with GOF tests based on likelihood ratios of a specified H_0 vs alternative H_1 that is *constructed from the data*.

They are ubiquitous because they are *generally useful* (for the effort put into them). But (!) we know that they can be *terrible* at finding deviations from H_0. NP Lemma is not applicable.
The classic intro lab course χ^2 :

Throws away info on ordering of data points: misses trends.

Throws away info on the signs of $(d_i - y_i)$.

The human eye can often easily "see" an alternative hypothesis that has more plausible physics motivation than the saturated model (and can of course be easily fooled)!

We routinely look at the individual contributions to the terms in the sums: bumps, trends, etc. It is standard to plot something like $(d_i - y_i)/\sigma_i$ beneath a histogram, etc.

Generalizations of these tools exist in machine learning.

Basic idea was also applied by statisticians to *unbinned* data.

Historical literature is nearly all for 1D data. Often H_0 transformed to uniform on (0,1) with probability integral transformation.

Cannot "saturate" continuous data set with parameters, so *assumptions are added*.

The *Neyman smooth tests* of goodness of fit, and their generalizations, compare given H_0 to H_1 constructed from the data, approximating LR. Again, the components of the alternative model can be examined for insight into discrepancies.

Key issue: what basis functions to use, and how many components to sum?

I have not seen Neyman smooth tests used in HEP. Has anyone?

With Google, I found a (thinly cited) paper

Gerda Claeskens and Nils Lid Hjort, "Goodness of Fit via Non-Parametric Likelihood Ratios", Scand. J. Stat. 31 (2004) 487.
https://www.jstor.org/stable/4616847.

To explain the basic idea, suppose for a moment that one envisages a specific alternative f to f_0. In this case the Neyman–Pearson lemma tells us that the optimal test procedure consists in rejecting f_0 when the ideal likelihood ratio statistic

$$\Lambda_n(f) = \frac{\prod_{i=1}^{n} f(X_i)}{\prod_{i=1}^{n} f_0(X_i)} \tag{1}$$

is large enough. But non-parametric density estimation strategies are available for producing an estimate $\hat{f}$ of the unknown f. Hence

$$\Lambda_n(\hat{f}) = \frac{\prod_{i=1}^{n} \hat{f}(X_i)}{\prod_{i=1}^{n} f_0(X_i)} \tag{2}$$

is a natural estimate of the underlying optimal $\Lambda_n(f)$, constructed without prior assumptions.

(What they call "nonparametric density estimation" is actually a very flexible parametric form.)

Again, Key issue: what basis functions to use, and how many components to sum?

The next step in history: Machine Learning!

This PhyStat: Talks by Kuusela (April 27) and Wulzer, et al.

Some progress on the longstanding problem of unbinned data with dimension > 1.

Key issues: design of components, how to regularize, basically same issues as in Neyman smooth tests and in Claeskens & Hjort.

"Old" issue: how much to let the data suggest alternative hypotheses, and how to deal with resulting issues.

One way: use part of the data (or one expt) to suggest hypotheses, use disjoint data (or second expt) to test. (ATLAS paper in Inês talk.) One might say that this is essentially how science works (repeatable expts, etc.), but I have voiced some objections in the past…

"Old" issue: test point null against continuous alternative
https://www.stat.cmu.edu/stamps/webinar/robert-cousins-feb-12/

For discussion

1) **Are advances with ML practical or foundational? Larry:**

Aside: why did everyone start calling classification and regression Machine Learning? It's statistics! We've been doing it for 100 years!

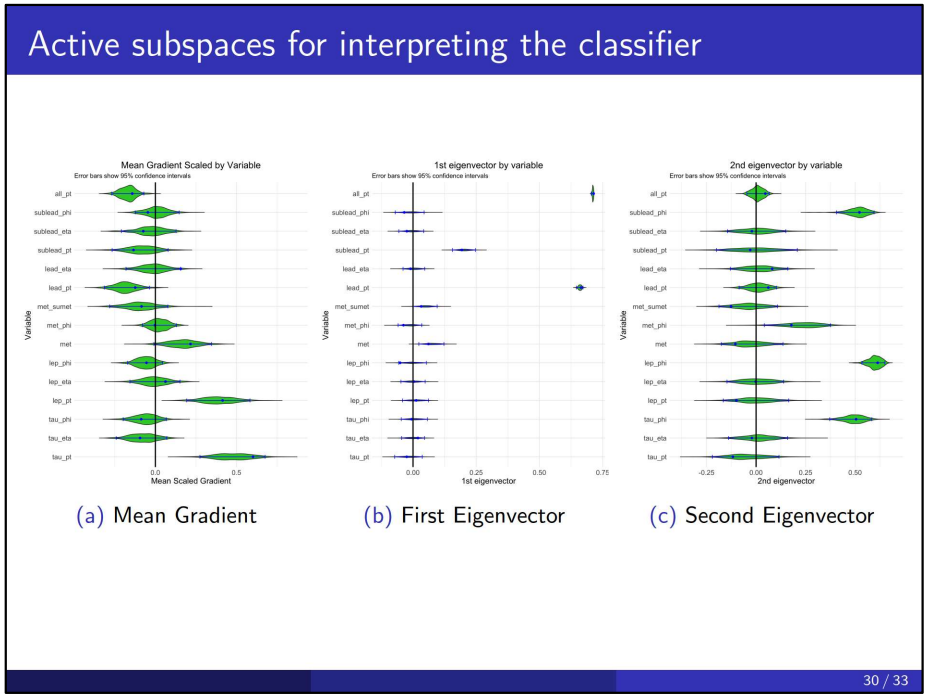
a) **My view: ML brings powerful new tools to multi-D anomaly detection, but helpful to keep in mind that "It's statistics!", with >100-year-old foundational issues.**

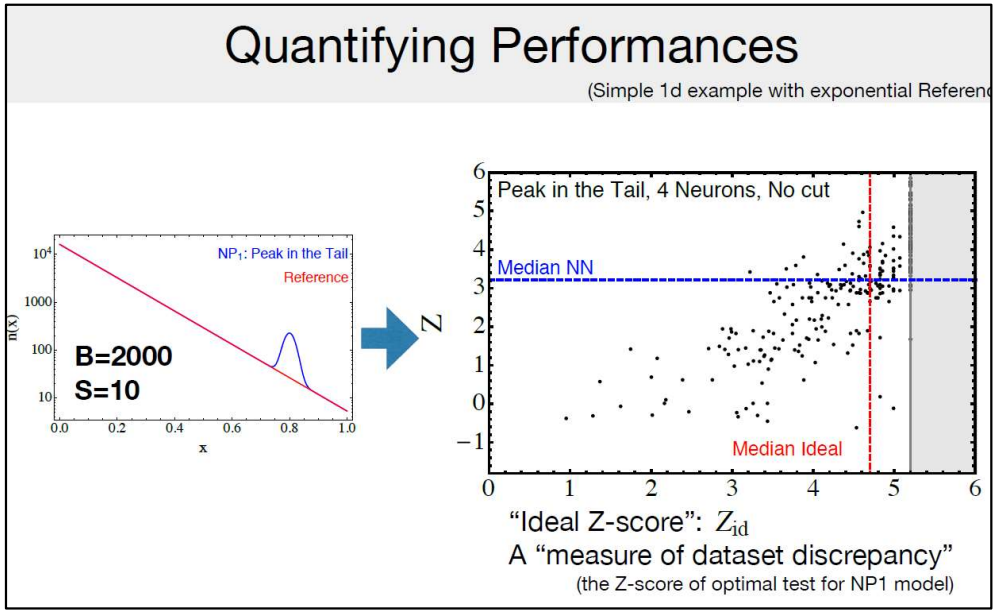
2) **"Optimal" needs to be carefully defined (examples today).**
For simple case, we can use NP: for fixed Type I error prob α, optimal means lowest Type II error prob β. But how to summarize curve of β vs α (as function of unknown params)?
I find "area under ROC curve" to be very blunt criterion: our operating points typically have small Type I error.

For discussion (cont.)

3) I *really* liked Active Subspaces study in Mikael's talk, April 27. This seems to be well-known to experts, but has not propagated to routine use in analysis documentation. IMO crucial for overcoming "black box" reputation.

4) I *really* liked scatter plots shown by Andrea of "ideal Z score" (using NP LR with "true" H_1) vs machine-derived Z-score. Can this become an industry standard for HEP (and beyond)? (Are there precedents?)





Thanks to all, including my “sponsor”, U.S. DOE Office of Science

For some details, references, and acknowledgements, see
Robert D. Cousins, “Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms” (2013),
https://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf