Open Science Framework DOI:10.31219/osf.io/***

Testable and untestable models [draft]

P.G.L. Porta Mana
Kavli Institute, Trondheim <pgl@portamana.org>

***; updated 25 May 2020

***

1 Would you like to know which one is true, eventually?

The word *model* seems to have been taking the place of *hypothesis* since around the 1960s. This replacement does not seem to be connected with the shift from frequentist to Bayesian methods.

Does the replacement of 'hypothesis' by 'model' indicate a shift in concepts in all probability theory and statistics? Or are these two words simply equivalent?

Many hodiernal authors indeed seem to use them interchangeably. An example is Kass & Raftery's famous review[1]. Initially they define a statistical model as something that represents the probability of the data according to a hypothesis (p. 773). Eventually they use the two terms interchangeably, for example saying at times 'the hypothesis H_k', at times 'the model H_k'. ***

'It's just a matter of terminology', some may say. But I want to argue that that there are three deeper important problems, and they need to be solved.

The first problem is pedagogical. Many students of probability and statistics have uncertainties and misconceptions about what a model is. Model of what?

The second problem is an *inadvertent* mismatch between hypotheses declared to be under analysis, and hypotheses that are actually analysed. That is, some authors state that they will compare a particular set of hypotheses, but an analysis of their mathematics reveals that they are unintentionally comparing a different set.

The third problem is the *verifiability* of a hypothesis or model. I mean the following.

[1] Kass & Raftery 1995.

This document is designed for screen reading and two-up printing on A4 or Letter paper

There's a box with three balls, which can be blue or red. You have four hypotheses about their colours: H_0 : no blue balls, H_1 : one blue ball, two red balls, H_2 : two blue balls, one red, H_3 : all blue balls. Each of these hypotheses, granted additional assumptions about the drawing procedure, leads to a probability for the colour in the first draw and in all four draws.

You can assign a probability to each hypothesis. As draws are made, your probability for each hypothesis changes, and so does the probability for each colour in the next draw.

Once all three balls are drawn you see which hypothesis is true (some were proven false along the way). This is reflected in their probabilities conditional on all data: one hypothesis gets unit probability the rest zero.

*** example with composite hypotheses

*** example with models. ... Something peculiar happens: we have collected *all possible* data, but none of the models has reached unit probability. You are still uncertain about the models. This leads to some questions:

- What are your models about? they cannot be about the colour composition, because now you know that perfectly and yet the truth of the models is still unsettled.
- What kind of data would you need to settle their truth?
- Are your models really relevant for this urn problem?

There's a closed box. You have three hypotheses: the box contains one, two, or three balls. By shaking the box a little and listening to the rattle you get the impression that there should be three balls. Maybe two. Unlikely to be just one. You may express your beliefs with a probability distribution.

But how do you *verify* how many balls there are? If possible you could open and look; or X-ray the box; or weigh it (assuming you knew the separate weights of box and balls). The electromagnetic or weigh data would tell you which hypothesis is true. You verify that there are two balls.

The verification of other hypotheses, often entertained in science, is not so straightforward. It would require an infinite amount of data or time.

Someone states that, during the whole history of a specific coin, it will come up heads 51% of the times it will be tossed, and tails 49%

of the times. Another person says 50%/50%. How do you verify these hypotheses? You would need to monitor the coin, probably bequeathing this scientific task to several future generations (the oldest coin existing today is about 2 600 years old[2]), amassing a very long sequence of data. Or imagine ancient Greeks making hypotheses about the Earth's precise distance from the Sun. The data necessary to verify their hypotheses could not be gathered then, owing to lack of technology. But they were some centuries later[3]. In cases like these we say that the hypotheses are verifiable *in principle*. Scientific hypotheses are often of this kind.

There are hypotheses that cannot be verified even with infinite data or futuristic technologies. Their verification hinges on dubious notions, such as multiple copies of this universe with slightly different initial conditions. It is debatable whether we can speak of 'hypotheses' and 'verification' in this case.

The three cases above are not sharply distinct. There is an increasing difficulty with the kind or quantity of data necessary to verify a set of hypotheses.

I believe that whenever we propose a set of hypotheses or models we should always point out what kind of data would be necessary for their verifiability. Such a requirement is in line with today's emphasis on reproducibility and the abandonment of 'significance' in favour of statistical thinking[4]

🧩 'the null must be nested within the alternative' (end of p. 776)

I have never seen a paper in which a probability model is *refuted*. Sure, there was 'hypothesis rejection at some significance level'

***comparing two frequency-priors using exch. data is like using one data point only.

***[5] Raftery: non-testable models ***[6] is the 'hot hand' example really testable?

2 Discussion

The shift from 'hypothesis' to 'model' seems to reflect the gradual abandonment of trying to understanding a phenomenon to simply trying to fit to it some equation pulled out of a hat. This is also reflected in the defacing of the verb *explain*: many authors say 'this distribution

[2] https://rg.ancients.info/lion/article.html [3] Goldstein 1985. [4] ASA 2019 § 1. [5] Raftery et al. 1989. [6] Kass & Raftery 1995.

explains the data’ (or the variance of the data, or similar) when in reality it just *fits* the data – it does not explain them.

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

asa (American Statistical Association) (2019): *Moving to a world beyond "$p < 0.05$"*. Am. Stat. **73**S1, 1–19. Ed. by R. L. Wasserstein, A. L. Schirm, N. A. Lazar.

Goldstein Jr., S. J. (1985): *Christiaan Huygens' measurement of the distance to the Sun*. Observatory **105**, 32–33.

Kass, R. E., Raftery, A. E. (1995): *Bayes factors*. J. Am. Stat. Assoc. **90**430, 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf.

Raftery, A. E., Fairley, D., Joe, H., Weissman, I., Singpurwalla, N. D., Pickands III, J., Smith, R. L. (1989): *[Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone]: Are ozone exceedance rates decreasing? Comments. Rejoinder*. Stat. Sci. **4**4, 378–393. See Smith (1989).

Smith, R. L. (1989): *Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone*. Stat. Sci. **4**4, 367–377. See also comments and rejoinder in Raftery, Fairley, Joe, Weissman, Singpurwalla, Pickands III, Smith (1989).