Souvenir O: Interpreting Probable Flukes

There are three ways to construe a claim of the form: A small *P*-value indicates it's improbable that the results are statistical flukes.

(1) The person is using an informal notion of probability, common in English. They mean a small *P*-value gives grounds (or is evidence) of a genuine discrepancy from the null. Under this reading there is no fallacy. Having inferred *H**: Higgs particle, one may say informally, "so probably we have experimentally demonstrated the Higgs," or "probably, the Higgs exists."

"So probably" H_1 is merely qualifying the grounds upon which we assert evidence for H_1.

(2) An ordinary error probability is meant. When particle physicists associate a 5-sigma result with claims like "it's highly improbable our results are a statistical fluke," the reference for "our results" includes: the overall display of bumps, with significance growing with more and better data, along with satisfactory crosschecks. Under this reading, again, there is no fallacy.

To turn the tables on the Bayesians a bit, maybe they're illicitly sliding from what may be inferred from an entirely legitimate high probability. The reasoning is this: With probability 0.9999997, our methods would show that the bumps disappear, under the assumption the data are due to background H_0. The bumps don't disappear but grow. Thus, infer H^*: real particle with thus and so properties. Granted, unless you're careful about forming probabilistic complements, it's safer to adhere to the claims along the lines of U-1 through U-3. But why not be careful in negating D claims? An interesting phrase ATLAS sometimes uses is in terms of "the background fluctuation probability": "This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with … the Standard Model Higgs boson" (2012b, p.1).

(3) The person is interpreting the *P*-value as a posterior probability of null hypothesis H_0 based on a prior probability distribution: $p = \Pr(H_0|\boldsymbol{x})$. Under this reading there is a fallacy. Unless the *P*-value tester has explicitly introduced a prior, it would be "ungenerous" to twist probabilistic assertions into posterior probabilities. It would be a kind of "confirmation bias" whereby one insists on finding a sentence among many that could be misinterpreted Bayesianly.

ASA 2016 Guide: Principle 2 reminds practitioners that *P*-values aren't Bayesian posterior probabilities, but it slides into questioning an interpretation sometimes used by practitioners – including Higgs researchers:

> *P*-values do not measure (a) the probability that the studied hypothesis is true, or (b) the probability that the data were produced by random chance alone. (Wasserstein and Lazar 2016, p. 131)[4]

[4] The ASA 2016 Guide's Six Principles:

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

I insert the (a), (b), absent from the original principle 2, because, while (a) is true, phrases along the lines of (b) should not be equated to (a).

Some might allege that I'm encouraging a construal of *P*-values that physicists have bent over backwards to avoid! I admitted at the outset that "the problem is a bit delicate, and my solution is likely to be provocative." My question is whether it is legitimate to criticize frequentist measures from a perspective that assumes a very different role for probability. Let's continue with the ASA statement under principle 2:

> Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (Wasserstein and Lazar 2016, p. 131)

Start from the very last point: what does it mean, that it's not "about the explanation"? I think they mean it's not a posterior probability on a hypothesis, and that's correct. The *P*-value is a methodological probability that can be used to quantify "how well probed" rather than "how probable." Significance tests can be the basis for, among other things, falsifying a proposed explanation of results, such as that they're "merely a statistical fluctuation." So the statistical inference that emerges is surely a statement about the explanation. Even proclamations issued by high priests – especially where there are different axes to grind – should be taken with severe grains of salt.

As for my provocative interpretation of "probable fluctuations," physicists might aver, as does Cousins, that it's the science writers who take liberties with the physicists' careful U-type statements, turning them into D-type statements. There's evidence for that, but I think physicists may be reacting to criticisms based on how things look from Bayesian probabilists' eyes. For a Bayesian, once the data are known, they are fixed; what's

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

These principles are of minimal help when it comes to understanding and using *P*-values. The first thing that jumps out is the absence of any mention of *P*-values as error probabilities. (Fisher-N-P Incompatibilist tribes might say "they're not!" In tension with this is the true claim (under #4) that cherry picking results in spurious *P*-values; p. 132.) The ASA effort has merit, and should be extended and deepened.

random is an agent's beliefs or uncertainties on what's unknown – namely the hypothesis. For the severe tester, considering the probability of {d($\boldsymbol{X}$) ≥ d($\boldsymbol{x}_0$)} is scarcely irrelevant once d($\boldsymbol{x}_0$) is known. It's the way to determine, following the severe testing principles, whether the null hypothesis can be falsified. ATLAS reports, on the basis of the *P*-value display, that "these results provide conclusive evidence for the discovery of a new particle with mass [approximately 125 GeV]" (ATLAS collaboration 2012b, p. 15).

Rather than seek a high probability that a suggested new particle is real; the scientist wants to find out if it disappears in a few months. As with GTR (Section 3.1), at no point does it seem we want to give a high formal posterior probability to a model or theory. We'd rather vouchsafe some portion, say the SM model with the Higgs particle, and let new data reveal, perhaps entirely unexpected, ways to extend the model further. The open-endedness of science must be captured in an adequate statistical account. Most importantly, the 5-sigma report, or corresponding *P*-value, strictly speaking, *is not the statistical inference*. Severe testing premises – or something like them – are needed to move from statistical data plus background (theoretical and empirical) to detach inferences with lift-off.