Souvenir Q: Have We Drifted From Testing Country? (Notes From an Intermission)

Before continuing, let's pull back for a moment, and take a coffee break at a place called Spike and Smear. Souvenir Q records our notes. We've been exploring the research program that appears to show, quite convincingly, that significance levels exaggerate the evidence against a null hypothesis, based on evidential assessments endorsed by various Bayesian and Likelihoodist accounts. We suspended the impulse to deny it can make sense to use a rival inference school to critique significance tests. We sought to explore if there's something to the cases they bring as ammunition to this conflict. The Bayesians say the disagree-ment between their numbers and P-values is relevant for impugning P-values, so we try to go along with them.

Reflect just on the first argument, pertaining to the case of two-sided Normal testing H_0: $\mu = 0$ vs. H_0: $\mu \neq 0$, which was the most impressive, particularly with $n \geq 50$. It showed that a statistically significant difference from a test hypothesis at familiar levels, 0.05 or 0.025, can correspond to a result that a Bayesian takes as evidence *for* H_0. The prior for this case is the spike and smear, where the smear will be of the sort leading to J. Berger and Sellke's results, or similar. The test procedure is to move from a statistically significant result at the 0.025 level, say, and infer the posterior for H_0.

Now our minimal requirement for data $\boldsymbol{x}$ to provide evidence for a claim H is that

> (S-1) H accords with (agrees with) $\boldsymbol{x}$, and
>
> (S-2) there's a reasonable, preferably a high, probability that the procedure would have produced disagreement with H, if in fact H were false.

So let's apply these severity requirements to the data taken as evidence for H_0 here.

Consider (S-1). Is a result that is 1.96 or 2 standard errors away from 0 in good accord with 0? Well, 0 is excluded from the corresponding 95% confidence interval. That does not seem to be in accord with 0 at all. Still, they have provided measures whereby $\boldsymbol{x}$ does accord with H_0, the likelihood ratio or posterior probability on H_0. So, in keeping with the most useful and most generous way to use severity, let's grant (S-1) holds.

What about (S-2)? Has anything been done to probe the falsity of H_0? Let's allow that H_0 is not a precise point, but some very small set of values around 0. This is their example, and we're trying to give it as much credibility as possible. Did the falsity of H_0 have a good chance of showing itself? The falsity of H_0 here is H_1: $\mu \neq 0$. What's troubling is that we found the probability of failing to pick up on population discrepancies as much as 1 standard error in excess of 0 is rather high (0.84) with $n = 100$. Larger sample sizes yield even less capability. Nor are they merely announcing "no discrepancy from 0" in this case. They're finding evidence for 0!

So how did the Bayesian get the bump in posterior probability on the null? It was based on a spiked prior of 0.5 to H_0. All the other points get minuscule priors having to share the remaining 0.5 probability. What was the warrant for the 0.5 prior to H_0? J. Berger and Sellke are quite upfront about it: if they allowed the prior spike to be low, then a rejection of the null would merely be showing an improbable hypothesis got more improbable. "[W]ho, after all, would be convinced," recall their asking: if "my conclusion is that H_0 has posterior probability 0.05 and should be rejected" since it previously had probability, say 0.1 (1987, p. 115). A slight lowering of probability won't cut it. Moving from a low prior to a slightly higher one also lacks punch.

This explains their high prior (at least 0.5) on H_0, but is it evidence for it? Clearly not, nor does it purport to be. We needn't deny there are cases where a theoretical parameter value has passed severely (we saw this in the case of GTR in Excursion 3). But that's not what's happening here. Here they intend for the 0.5 prior to show, *in general*, that statistically significant results problematically exaggerate evidence.[8]

A tester would be worried when the rationale for a spike is to avoid looking foolish when rejecting with a small drop; she'd be worried too by a report: "I don't take observing a mean temperature of 152 in your 100 water samples as indicating it's hotter than 150, because I give a whopping spike to our coolants being in compliance." That is why Casella and R. Berger describe J. Berger and Sellke's spike and smear as maximally biased toward the null (1987a, p. 111). Don't forget the powerful role played by the choice of how to smear the 0.5 over the alternative! Bayesians might reassure us that the high Bayes factor for a point null doesn't depend on the priors given to H_0 and H_1, when what they mean is that it depends only on the priors given to discrepancies under H_1. It was the diffuse prior to the effect size that gave rise to the Jeffreys–Lindley Paradox. It affords huge latitude in what gets supported.

We thought we were traveling in testing territory; now it seems we've drifted off to a different place. It shouldn't be easy to take data as evidence for a claim when that claim is false; but here it is easy (the claim here being H_0). How can this be one of a handful of main ways to criticize significance tests as exaggerating evidence? Bring in a navigator from a Popperian testing tribe before we all feel ourselves at sea:

> Mere supporting instances are as a rule too cheap to be worth having . . . any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, if it can be refuted. (Popper 1983, p. 130)

The high spike and smear tactic can't be take as a basis from which to launch a critique of significance tests because it fails rather glaringly a minimum requirement for evidence, let alone a test. We met Bayesians who don't approve of these tests either, and I've heard it said that Bayesian testing is still a work in progress (Bernardo). Yet a related strategy is at the heart of some recommended statistical reforms.

[8] In the special case, where there's appreciable evidence for a special parameter, Senn argues that Jeffreys only required H_1's posterior probability to be greater than 0.5. One has, so to speak, used up the prior belief by using the spiked prior (Senn 2015a).