The Banning of NHSTP in “Basic and Applied Social Psychology”

Abel Rodríguez
Applied Mathematics and Statistics
University of California, Santa Cruz

Background

- 2015 BASP editorial banned null hypothesis significance testing procedure (NHSTP).
- You should look at the 2014 Editorial too:
  - NHSTP is logically invalid.
  - Publishing null effects and results that contradict previous research (publication bias).
  - Mediation effects and causal mechanisms.
- Also need to look at some of the editor's papers:
  - Trafimow, D. (2003). *Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem*. Psychological Review, 110, 526–535

Logically invalid

- 2015 Editorial: NHSTP “has been shown to be logically invalid”

p-value = $\Pr(+ \mid H_0) \neq \Pr(H_0 \mid +)$ = Posterior probability (Bayesian)

- Interpretation of p-balues is tricky:
  - p-values can be used to rule out alternative theories, but not not to support any specific one (we should say ***fail to reject*** rather than *accept*).
  - Type I and type II errors tell you what happens if the experiment was repeated many times, but nothing about what happens with *your data*.
  - Common to use the wrong interpretation …

A basic error of his own, and another reason to criticize p-values

- Trafimow's characterization of the difference is a good starting point, but technically not quite accurate:
  - p-values usually involve the probability of obtaining a value **at least as extreme** ➔ They involve not only the observed data, but also unobserved values, violating the likelihood principle.
  - Posterior probabilities (under subjective priors) involve **only** the observed sample.

$\Pr(|T| > T_{obs} \mid H_0)$ vs. $\Pr(H_0 \mid |T| = T_{obs})$

NHSTP vs. confidence intervals

- 2015 Editorial: “Confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval”.
- Again, an interpretation problem: the confidence of the interval tells you nothing about your specific sample.
- $1 - \alpha$% two-sided confidence intervals are equivalent to NHSTP for point vs. composite with α Type I error (test inversion).
- One consequence is that NHSTP is usually inconsistent.

Posterior probabilities vs. credible intervals

- The same thing is not true for Bayesian procedures: credible intervals behave (asymptotically) like confidence intervals, but posterior probabilities for point vs. composite are usually consistent.
- One way to bridge the (practical, although not the philosophical) gap is to have α depend on the sample size!
  - Control $a\alpha + b\beta$ instead of just α.
  - From a practical perspective, nothing wrong with the basics of the procedure, just with how it is implemented by fixing α.

Point vs. composite hypotheses

- Does it even make sense to test point vs. composite hypotheses?
- Some people propose instead to test

  $H_0: \theta_0 - \Delta \leq \theta \leq \theta_0 + \Delta$ vs. $H_a: \theta < \theta_0 - \Delta$ or $\theta > \theta_0 + \Delta$

  - Solves the statistical vs. practical significance dilemma.
  - How do you choose Δ?
- When hypotheses are driven by specific theories, comparing point vs. composite makes sense:
  - Does the Higgs boson exist?
  - Does Dexamethasone affect survival of cancer patients?

The Laplacian assumption …

- 2015 Editorial: "Bayesian procedures depend on some sort of Laplacian assumption to generate a number when none exists".
- Two components to this comment:
  - Eliciting (proper) subjective priors can be hard! Furthermore, conjugate priors can be poor choices.
  - Default (objective/noninformative) procedures for testing point vs. composite hypotheses based on Laplace's indifference principle can be invalid if they use improper priors.

Default priors

- Standard procedures to generate default priors often lead to improper priors (e.g., Laplace's indifference principle, Jeffreys'), which can lead to testing procedures with very poor performance (Barlett's paradox).
- There is a well established (and still growing) literature on how to deal with this problem.
  - This is implicitly recognized: "However, there have been Bayesian proposals that at least somewhat circumvent the Laplacian assumption, and there might even be cases where there are strong grounds for assuming that the numbers really are there."

Other issues with NHSTP that are not mentioned by the editor

- "Asymmetry" of null and alternative hypotheses.
- How do you deal with multiple hypotheses?
- The likelihood principle and the effect of stopping rules (or what to do when your recruitment does not go according to plan).

I would have liked a stronger defense of Bayesian procedures in the editorial!