Discussion: Classical p-values and the Bayesian posterior probability that the hypothesis is approximately true

Author: Brendan Kline
Discussant: Ahnaf Rafi

Northwestern University

March 31st 2023

Summary I: high level stuff

- P-values are ubiquitous as a standard for what constitutes (sufficient) empirical evidence for scientific discovery.
- Usually framed in the context of rejecting **an exact "neutral" null hypothesis** in favor of **("more interesting") alternative(s)**.
- **This paper:** are p-values are actually informative about the (Bayesian posterior) probability that a null is **approximately correct**?
- **Why?** We construct tests for **exact nulls**, but draw conclusions about **approximate nulls**.

Summary II: framing

- Throughout, for user-specified c, ε
  - **Exact null:** $H_0 : \theta = c$.
  - **Approximate null:** $H_0^\varepsilon : \theta \in [c - \varepsilon, c + \varepsilon]$
- Low p-value is associated (*conflated*) by researchers with "low probability" that θ is "close to zero".
- **Existing literature:** If the prior has an atom at $\theta = c$, there is an increasing relationship between $\Pr(\theta = c|\text{Data})$ and p-value.
- **Paper's framing:** in social sciences,
  1. Even if we test H_0, we really mean H_0^ε and draw conclusions about the latter.
  2. In addition, no reason to have prior with $0 < \Pr(\theta = \tau) < 1$ for any $\tau \in [c - \varepsilon, c + \varepsilon]$.

Summary III: results

1. If we assume a continuous prior with positive density in $[c-\varepsilon, c+\varepsilon]$, then p-value and $\Pr\left(H_0^\varepsilon|\text{Data}\right)$ no longer have an increasing relationship.
2. In particular, $\Pr\left(H_0^\varepsilon|\text{Data}\right)$ can be higher for lower p-values (and vice versa) - suggests caution against using low p-values as a standard for judging empirical findings.
3. Even though main results are asymptotic, the phenomenon is true generally and in finite samples - not an "asymptotic curiosity".

Some comments

I like the motivation of the particular Bayesian framework from what is done/said in practice:

- Bayesian approach is appropriate since we want to draw (probabilistic) conclusions about true values of the parameters.
- Approximate null and continuous priors are motivated by how researchers think about null hypotheses.
- The use of Bernstein-von Mises approximations is justified by the continuous prior.

My take on broader implications of the paper

- Existing results on the increasing relationship between p-values and $\Pr(H_0|\text{Data})$ with an atomic prior are internally consistent, but interpreting them outside their context is incorrect. The paper does a good job of driving home that point.
- Highlights the potential cost to the overall community of using p-values as a scientific standard: can miss out on treatment effects that are probably not close to zero.
- The large sample approximations provide one alternative standard for evaluating empirical findings.
- Provides a useful tool to retroactively assess evidence about whether published (non-zero) treatment effects are in fact likely close to zero.

Some criticisms

- The finite sample analysis is nice, but seems “tied” to the large sample results since sampling and posterior distributions are t and F distributions in finite samples.
- Some Monte-Carlo simulations with alternative continuous posteriors perhaps? (In lieu of closed form results.)
- Along those lines, asymptotic approximations are nice, but how good are they really? Is there something like a (uniform) Berry-Esseen bound for BvM?

Controversial stuff (my own thoughts, time permitting) I

- To me, drawing conclusions about parameter values on the basis of p-values has always seemed awkward from a frequentist perspective:
  - Researchers can compute $\Pr(\text{Data}|H_0)$, but want to make statements about $\Pr(H_0|\text{Data})$.
  - By Bayes' theorem,

$$\begin{aligned}
\Pr(H_0|\text{Data}) &= \frac{\Pr(\text{Data}|H_0)\Pr(H_0)}{\Pr(\text{Data})} \\
&= \frac{\Pr(\text{Data}|H_0)\Pr(H_0)}{\Pr(\text{Data}|H_0)\Pr(H_0) + \Pr(\text{Data}|\neg H_0)\Pr(\neg H_0)}.
\end{aligned}$$

  Red $=$ not available in frequentist world. Blue $=$ basically what the p-value corresponds to.

- Additionally, there is a logical leap in the act of drawing conclusions about parameter values on the basis of p-values - they are computed "conditional" on both the **estimator and the hypothesized value of the parameter**.

Controversial stuff (my own thoughts, time permitting) II

- As a standard for judging empirical findings, "small p-values" (Fisherian paradigm, null hypothesis significance testing [NHST]) thus seems rather strange.
- The Neyman-Pearson null+alternative, Type I + II error control paradigm is also not always helpful for evaluating scientific findings.
- Should not surprise anyone - the p-value and confidence intervals are all about characterizing sampling error **assuming the null is true**.
- Not new, see e.g. Gigerenzer, Krauss, and Vitouch (2004), and Szucs and Ioannidis (2017) and references therein.
- Begs the question: from a frequentist perspective, how do we evaluate empirical findings without NHST?