Why optional stopping can be a problem for Bayesians[1]

–

Bayesian Machine Learning

Boëzennec Robin, Hugo Simon

24/03/2022

[1] Rianne de Heide et al. (Nov. 2020). "Why optional stopping can be a problem for Bayesians". In: *Psychonomic Bulletin and Review* 28.3, pp. 795–812.

Table of content

- Misusing data analysis to find patterns in data that can be presented as statistically significant.
- Because of incentives to publish positive results, conflict of interest or simply statistical misunderstanding.

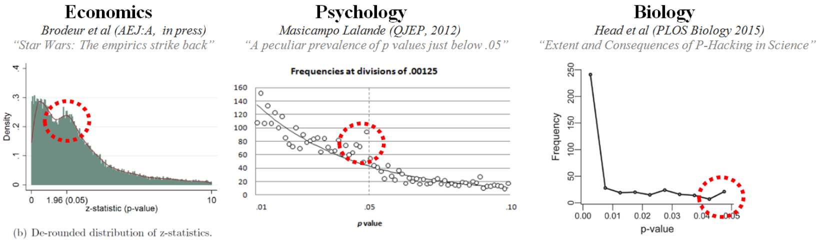


Figure 1: Dotted circles highlight excess of .05, though most p-values are way smaller.

Continuous monitoring

- Experimenter can choose to arbitrarily stop aggregating data depending on already observed ones.
- One practice to be careful with using p-value in NHST.
- E.g. stopping as soon as statistical significance is reached leads to rejecting null hypothesis at a higher alpha risk than the p-value estimates.

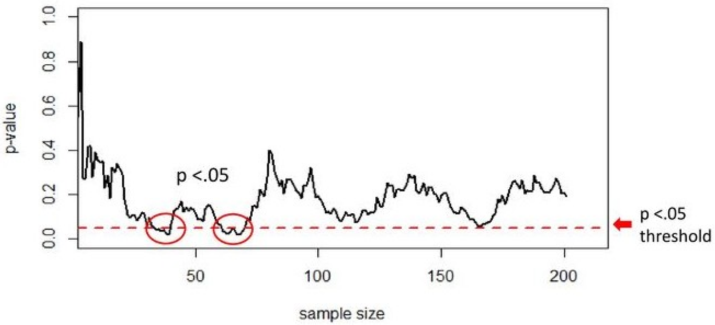


Figure 2: p-hacking with continuous monitoring on non-effect data.

Definition (Post-odds and FDR)

Let be tested pair of hypotheses (H_0, H_1) and measurable event $A \in \mathcal{X}$. Then we define the odds and False Discovery rate (FDR) of H given A as

$$\mathbb{O}\,(H \mid A) = \frac{\mathbb{P}(H_1 \mid A)}{\mathbb{P}(H_0 \mid A)} = \frac{1}{\mathbb{P}(H_0 \mid A)} - 1 = \frac{1}{\mathrm{FDR}(H \mid A)} - 1$$

$\mathbb{O} = 9 \iff \mathrm{FDR} = 0.1$ means "discovering" (rejecting H_0, **accepting** H_1) would be false with proba. 10%.

Definition (Bayes Factor)

Applying Bayes theorem to the previously define post-odds leads to

$$\mathbb{O}\,(H \mid A) = \frac{\mathbb{P}(H_1 \mid A)}{\mathbb{P}(H_0 \mid A)} = \frac{\mathbb{P}(A \mid H_1)}{\mathbb{P}(A \mid H_0)} \cdot \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} = \beta_{\mathrm{H}}(A) \cdot \mathbb{O}\,(H)$$

where β_{H} is called the Bayes Factor (BF) or likelihood ratio.

- Quantifies how prior belief is updated from observations.
- No need to compute data evidence $\mathbb{P}(A)$.

Definition (Post-odds calibration)

A functional $\gamma : \mathcal{X} \mapsto \mathbb{R}^+$ is said to be post-odds calibrated for the model H if $\forall c$ with non zero probability, $\mathbb{O}\left(H \mid \gamma(X) = c\right) = c$

We know that for fixed n

- $\mathbb{O}\left(H \mid \mathbb{O}\left(H \mid X^n\right) = c\right) = c$
- $\mathbb{P}\left(\text{p-value}(X^n) \leq c \mid H_0\right) = c$

But does calibration hold with optional stopping ?

$$\underbrace{\mathbb{O}\left(H \mid \mathbb{O}\left(H \mid X^\tau\right)\right)}_{\substack{\textit{observed post-odds} \\ \text{(under "}\textit{nominal post-odds}\text{ claim")}}} \stackrel{?}{=} \underbrace{\mathbb{O}\left(H \mid X^\tau\right)}_{\textit{nominal post-odds}}$$

Yes[2,3] ! (As long as $H \perp\!\!\!\perp \tau$)

[2] Alex Deng et al. (Oct. 2016). *Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing*.

[3] Allard Hendriksen et al. (2021). "Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations". In: *Bayesian Analysis* 16.3.

Table of content

$$\mathbb{O}\,(H \mid X) = \frac{\exp\left(\frac{n^2\bar{x}^2}{2(n+1)}\right)}{\sqrt{n+1}} \cdot \frac{1}{1}$$

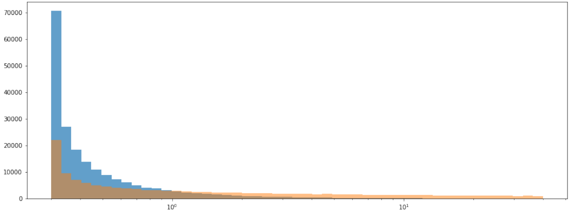
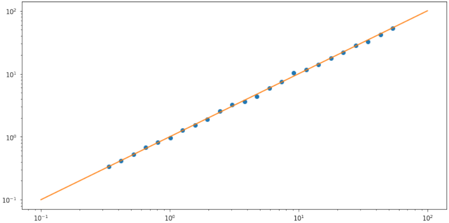


Figure 3: Histogram for the two sttings



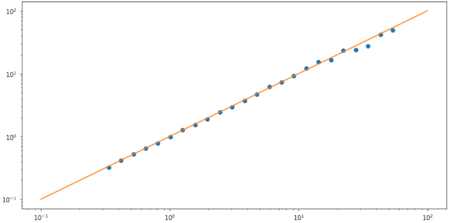


Figure 4: Experiment 0 calibration with n = 10 on the top and with optional stopping on the bottom

But what happens if the prior is not fully believed ?

$$\mathbb{O}\left(H \mid X\right) = \frac{1}{\sqrt{n+1}} \left(1 - \frac{\left(\frac{1}{n+1} \sum_{i=1}^{n} x_i\right)^2}{\frac{1}{n+1} \sum_{i=1}^{n} x_i^2} \right)$$

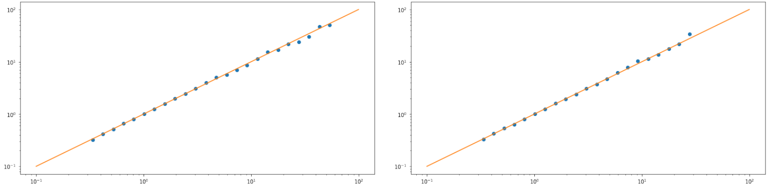


Figure 5: Experiment 1 calibration with n = 10 on the left and with optional stopping on the right

Experiment 2: *Type I prior*

$$\mathbb{O}\left(H \mid X\right) = \frac{\exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right)}{\int_{-\infty}^{\infty} \frac{1}{\pi(1+\mu^2)} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \mathrm{d}\mu}$$

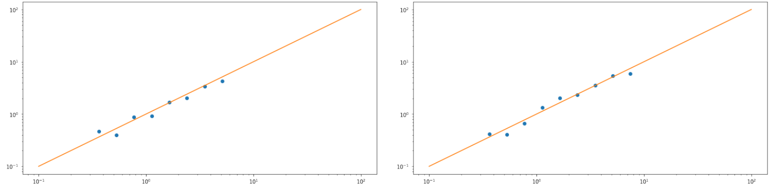


Figure 6: Experiment 2 with Cauchy prior and calibration with n = 10 on the left and with optional stopping on the right (our results)

Experiment 2: *Type I prior*

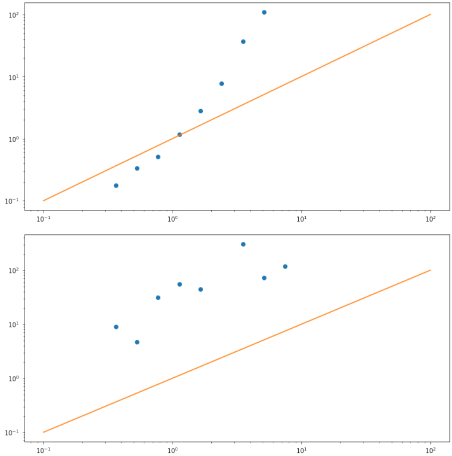


Figure 7: Experiment 2 with constant prior and calibration with n = 10 on the left and with optional stopping on the right

Type II prior

- Priors that depend on the sample size and sometimes data itself.
- Common in some Bayesian literature though they do not define a generative model.
- Furthermore, the introduction of a stopping time makes it unclear which prior we should use
  - prior with fixed n or with τ?
  - Jeffreys prior considering optional stopping?

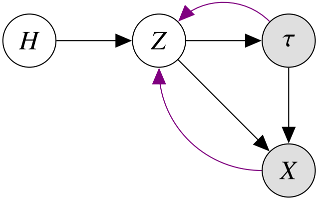


Table of content

What conclusions for subjective™ and objective™ Bayesians ?

Objective Bayesian

- *Type 0 prior* $\boxed{\checkmark}$
  - works fine
- *Type I prior* $\boxed{\text{X}}$
  - robustness to priors is crucial
  - τ just amplifies calibration loss
- *Type II prior* $\boxed{\text{X}}$
  - no satisfying calibration definition

Subjective Bayesian

- *Type 0 prior* $\boxed{\checkmark}$
  - works fine
- *Type I prior* $\boxed{\approx}$
  - well if this is what you believe...
  - ...but might look for consensus
- *Type II prior* $\boxed{?}$
  - how can this formalize any belief ?

Deng, Alex et al. (Oct. 2016). *Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing*.

Heide, Rianne de et al. (Nov. 2020). "Why optional stopping can be a problem for Bayesians". In: *Psychonomic Bulletin and Review* 28.3, pp. 795–812.

Hendriksen, Allard et al. (2021). "Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations". In: *Bayesian Analysis* 16.3.

Thank you!

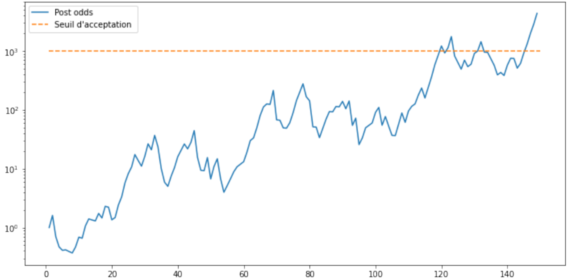


Figure 8: Post odds for our experiment on real data