“Why optional stopping can be a problem for Bayesians” reviewed

Boëzennec Robin, Simon Hugo
Master MVA
robin.boezennec@gmail.com, hugo.simon@telecom-paris.fr

Abstract

In the present age, A/B testing is one of the most effective uses of statistical theory, whether you are testing a website, a packaging or a drug. One issue with Null Hypothesis Statistical Testing (NHST), the foundation of A/B testing technique, is that experimenters are not permitted to examine the results in real time and make decisions. This restriction is seen by many as a setback in the technical trend toward real-time data analytics. Bayesian Hypothesis Testing, which seems to be more suited to real-time decision-making, has recently gained popularity as a viable alternative to NHST. While NHST adjustments for continuous monitoring are well established in the literature and well recognized in the community, the argument over whether continuous monitoring is a good idea continues. In this report, we emphasize main points of the optional stopping with Bayes Factor controversy, and we illustrate results and limitations with numerical experiments.

1 Introduction

p-hacking can be described as a range of practices, deliberate or not, of misusing data analysis to find patterns in data that can be presented as statistically significant, thus dramatically increasing and understating the risk of false claims. Whether it is because scientific communities give harsh incentives to statistical significant results publishing, because of conflict of interest, or because of statistical misunderstanding, there have been huge motivations to produce tests that are robust to some of these practices. A family of them are called continuous monitoring or optional stopping. This means an experimenter can choose to arbitrarily stop aggregating data depending on already observed ones. Classical NHST is not robust to optional stopping as p-value can be largely biased for some simple stopping conditions, see Figure 1.

Recently, interests in using Bayesian model comparison for two sample hypothesis testing are growing (Deng et al. 2016). The type of statistical interpretations we make from Bayesian tests is fundamentally different from NHST. As such, a Bayesian alternative to p-value is the Bayes Factor. To understand its role, let’s first introduce conditional odds.

Definition 1.1 (post-odds and FDR)**.** Let be a tested pair of hypotheses $H = (H_0, H_1)$ describing a full generative model, and a measurable event $A \in \mathcal{X}$. Then we define the odds of H conditionally to A (a.k.a. post-odds) as

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

where $\mathrm{FDR}(H \mid A)$ is the False Discovery Rate (FDR) of the tested pair H given A, e.g. $\mathbb{O} = 9 \iff \mathrm{FDR} = \frac{1}{1+\mathbb{O}} = 0.1$ means "discovering" (rejecting H_0, accepting H_1) would be false with proba. 10%.

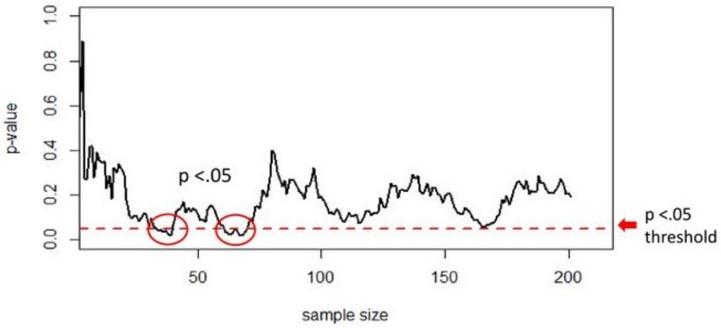

Figure 1: p-hacking with continuous monitoring with non-effect data. Stopping as soon as statistical significance is reached leads to rejecting the null hypothesis at a higher alpha risk than p-value estimates.

Definition 1.2 (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. It quantifies how prior belief is updated from observations i.e. $\beta_{\mathrm{H}} \geq 1$ (resp. $\beta_{\mathrm{H}} \leq 1$) means observations favor H_1 over H_0 (resp. vice versa).

We clearly see with this definition that using odds instead of probability allows us to update prior belief without any need to compute the data evidence $\mathbb{P}(A)$. In the following, we will see an important property that makes Bayes Factor-based model comparison attractive.

2 Post-odds calibration

Consider a weather forecaster who, on each day, announces the probability that it will rain the next day at a certain location. It is standard terminology to call such a weather forecaster ***calibrated*** *if, on average on those days for which he predicts "probability of rain is 30%", it rains about 30% of the time, on those days for which he predicts 40%, it rains 40% of the time, and so on. Thus, although his predictions presumably depend on a lot of data such as temperature, air pressure at various locations etc., given* ***only*** *the fact that this data was such that he predicted a, the actual probability is a.*

Rouder 2014 formalize this at-first-look-intuitive property for post-odds

Definition 2.1. A functional $\gamma : \mathcal{X}^n \mapsto \mathbb{R}^+$ is said to be post-odds calibrated for the model H if

$$\mathbb{O}(H \mid \gamma(X^n)) = \gamma(X^n)$$

i.e. $\forall c$ with non zero probability, $\mathbb{O}(H \mid \gamma(X^n) = c) = c$.

One can simply prove that for a fixed $n \in \mathbb{N}$, $\gamma(X^n) = \mathbb{O}(H \mid X^n)$ is calibrated for H. This could be seen as an equivalent of the frequentist p-value calibration, namely

$$\gamma(X^n) = \text{p-value}(\mathbf{X^n}) \text{ is such that } \forall c, \mathbb{P}(\gamma(X^n) \leq c \mid H_0) = c$$

That is to say, under H_0, p-value($\mathbf{X^n}$) $\sim \mathcal{U}(0,1)$. However, one may ask if this calibration remains if instead of fixing the number of samples, we let it be defined by an adapted stopping time τ. Though p-value is not calibrated in general anymore, it as been shown that if $\tau \perp\!\!\!\perp H$, then

$$\mathbb{O}(H \mid \mathbb{O}(H \mid X^\tau)) = \mathbb{O}(H \mid X^\tau)$$

This result has been proven by Deng et al. 2016, and in a more general setting (with improper priors) by Hendriksen et al. 2021. $\mathbb{O}(H \mid X^\tau)$ is called the *nominal post-odds* as it is computed forgetting that data is sampled with a stopping time τ, whereas $\mathbb{O}(H \mid \mathbb{O}(H \mid X^\tau))$ is called *observed post-odds* as it is the odds we actually observe under the "*nominal post-odds* claim", sampling with a stopping time. We illustrate this main result with the following experiment.

2.1 Experiment 0

In the first experiment, we want to test the null hypothesis H_0 that the mean of a normal distribution is equal to 0, against the alternative hypothesis H_1 that the mean is not 0. In the case of H_1 we take the prior on the mean μ to be a standard normal and take 1 for both variance. Our prior odds are set to 1-to-1 so the posterior odds formula is :

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

We then make two different simulations: in the first one we chose n = 10 samples and computed the post odds for these 10 samples. In the second one, we start with one sample, and while the post odds are not 10-1 in favor of either hypothesis (or the number of samples superior to 25), we sample a new case, add it to our list, and compute once again the post odds. We can see in Figure 3 that as expected both methods are calibrated.

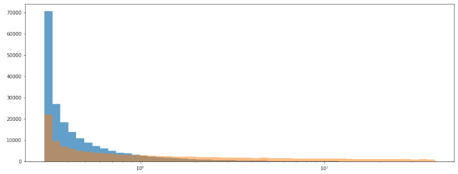

Figure 2: Experiment 0 histogram with n = 10

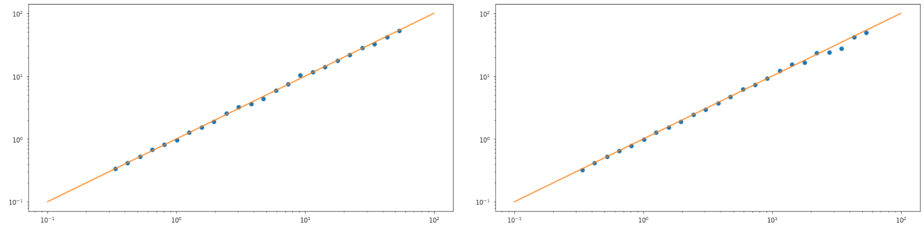

Figure 3: Experiment 0 calibration with n = 10 on the left and with optional stopping on the right

3 Post-odds strong calibration

Previous result shows that accounting only for the generative model H with fixed n when the true generative model is H with a stopping time τ does not break calibration. Note that in both cases, the generative model H is fully-believed i.e. reflects what we think about how the observed data is generated. In particular, priors involved in *nominal post-odds* computation such as the standard normal for μ in 2.1 is maintained in true generation. Thus, one is tempted to ask what happens to the calibration if the prior is different from the nominal one. Put differently, does calibration remains for accounting for the generative model H with fixed n when the true generative model is some modified H' with stopping time τ? This property is called strong calibration and we'll consequently illustrate conditions under which it holds or not.

3.1 Experiment 1: *Type 0 prior*

For this purpose, we modify the first experiment, supposing this time the variance is unknown. Because of this we have to define a prior on the variance which we will chose to be Jeffreys' prior $P(\sigma) = \frac{1}{\sigma}$ (note : our prior is not sampled according to Jeffreys' prior because we can't sample from it, so in practice we use fixed values for sigma but we do not lose the calibration by changing it). We have the new formula for the post-odds :

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

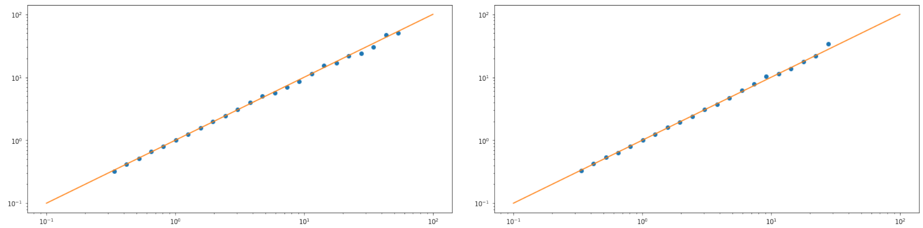

Figure 4: Experiment 1 calibration with n = 10 on the left and with optional stopping on the right

We reproduce the two settings of the first experiment with our new parameters and show that the results are still calibrated 4. This is due to the special property of the prior involving the authors of the article called *Type 0 priors*. These are priors on parameters freely occurring in both hypotheses for which strong calibration holds under optional stopping and the authors conjecture they are only priors that preserve group structure on parameters.

3.2 Experiment 2: *Type I prior*

In the previous experiment, under H_1, we had a normal prior on the mean μ, but this time the prior placed on the effect size will be a standard Cauchy. This gives us this posterior odds :

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

As we don't have an explicit formulation of the posterior odds, we use a numerical integrator to compute the fraction. First, we made the same experiments as before, using a standard Cauchy to sample the mean of the data for the H_1 hypothesis. This gives us results 5 showing calibration holding (points are not as perfectly aligned as before, which is attributed to approximations of numerical integration). In a second times we applied the same algorithm but using a constant prior for H_1 instead of the standard Cauchy. Again, the prior we had is not the one sampled, but in this case, we lost the calibration 6.

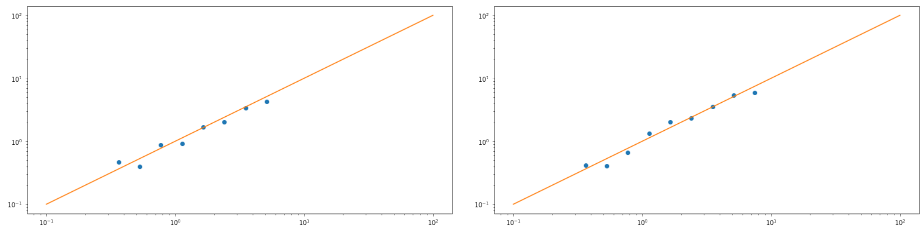

Figure 5: Experiment 2 with Cauchy prior and calibration with n = 10 on the left and with optional stopping on the right (our results)

The Cauchy prior in this experiment is called by the authors a *Type I prior*. Such priors still define a full generative model under optional stopping but strong calibration does not all. This means the model is sensitive to the truly sampled value and thus hypothesis made needs to be carefully founded. Moreover, the effect of optional stopping seems to amplify the calibration loss.

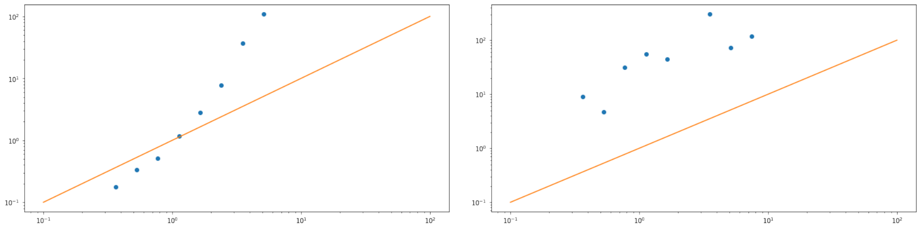

Figure 6: Experiment 2 with constant prior and calibration with n = 10 on the left and with optional stopping on the right

3.3 *Type II prior*

Additionally to *Type 0 prior* and *Type I prior*, authors introduce *Type II prior*. Such priors are quite common in the Bayesian literature but they do not define a full generative model under optional stopping, because they depend on the sample size and sometimes data itself. This means it is impossible to sample from such a model as data depends on sample size which depends on stopping time which depends on sampling from the prior which itself depends on data or sample size. Such examples can be seen in Michael 2010 Lecture 10, Example 5 or Heide et al. 2020, Example 3.

4 Application to real data

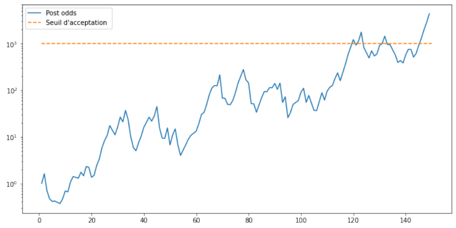

Figure 7: Post odds for our experiment on real data

We use real world data to showcase an example of Bayesian optional stopping, taking a dataset of NBA players, and trying to see if their average height is about 195 cm. As such we conducted the same experiment as in 3.1. Figure 7 shows the evolution of post-odds (in log-scale) with respect to the number of players taken into account. During the experiment, post-odds leans toward and crosses the really safe 1000:1 line, which allows to conclude NBA player's average height is not 195 cm.

5 What conclusions for subjective™ and objective™ Bayesians ?

While *Type 0 priors* would make both happy, there may be some different interpretations on the use of *Type I* and *Type II priors* between subjective and objective Bayesian.

For an objective Bayesian, robustness to prior's choice is crucial so *Type I priors* are problematic and their ability to lose calibration (without and with optional stopping) is to be considered. On the opposite, a radical subjective Bayesian would be ok with such priors because if such are used, this would have to be because they truly represent belief, thus strong calibration is a useless property. But a moderate subjective Bayesian could also think that though priors are fully and subjectively believed, one could pragmatically favor more consensual models which still handle optional stopping under a whole family of priors, as it will make individuals with different subjective priors agree together.

As for *Type II priors*, their inability to define generative models would make them unthinkable to subjective Bayesian, and limits their use by objective Bayesian. The fact that such simple experiments as those conducted in this report (calibration tests) can not be conducted on *Type II priors* question all the more their practice.

References

Deng, Alex et al. (Oct. 2016). “Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing”. In: DOI: 10.1109/DSAA.2016.33.

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. ISSN: 1531-5320. DOI: 10.3758/s13423-020-01803-x. URL: http://dx.doi.org/10.3758/s13423-020-01803-x.

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. ISSN: 1936-0975. DOI: 10.1214/20-ba1234. URL: http://dx.doi.org/10.1214/20-BA1234.

Michael, Jordan (2010). “Lecture Notes on Bayesian Modeling and Inference”. In: DOI: 10.1214/20-ba1234. URL: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/.

Rouder, Jeffrey (Mar. 2014). “Optional stopping: No problem for Bayesians”. In: *Psychonomic bulletin and review* 21. DOI: 10.3758/s13423-014-0595-4.