

Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing

Alex Deng *
Microsoft
One Microsoft Way
Redmond, WA 98052
alex deng@microsoft.com

Jiannan Lu *
Microsoft
One Microsoft Way
Redmond, WA 98052
jiannanl@microsoft.com

Shouyuan Chen
Microsoft
One Microsoft Way
Redmond, WA 98052
shouch@microsoft.com

ABSTRACT

A/B testing is one of the most successful applications of statistical theory in modern Internet age. One problem of Null Hypothesis Statistical Testing (NHST), the backbone of A/B testing methodology, is that experimenters are not allowed to continuously monitor the result and make decision in real time. Many people see this restriction as a setback against the trend in the technology toward real time data analytics. Recently, Bayesian Hypothesis Testing, which intuitively is more suitable for real time decision making, attracted growing interest as an alternative to NHST. While corrections of NHST for the continuous monitoring setting are well established in the existing literature and known in A/B testing community, the debate over the issue of whether continuous monitoring is a proper practice in Bayesian testing exists among both academic researchers and general practitioners. In this paper, we formally prove the validity of Bayesian testing with continuous monitoring when proper stopping rules are used, and illustrate the theoretical results with concrete simulation illustrations. We point out common bad practices where stopping rules are not proper and also compare our methodology to NHST corrections. General guidelines for researchers and practitioners are also provided.

Category and Subject Descriptors: G.3 [Probability and Statistics]: Statistical Computing

Keywords: A/B testing, controlled experiments, Bayesian statistics, optional stopping, continuous monitoring

1. INTRODUCTION

Many online service companies nowadays have been using online controlled experiments, a.k.a. A/B Testing, as a scientifically grounded way to evaluate changes and comparing different alternatives. A/B testing plays a leading role in establishing the mantra of data driven decision making, and is one of the basic pillars in Data Science.

*First and second authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Most of A/B tests are conducted using the statistical theory of frequentist null hypothesis statistical testing (NHST), namely t-test or z-test.¹ Experimenters using NHST summarize the test result in a p-value and reject the null hypothesis H_0 when the p-value is less than a prescribed confidence level α . The interpretation is that assuming all model assumptions are correct, doing so we can control the Type-I error, i.e. the probability of making a false rejection when H_0 is true, to be no greater than α .

Recently, interests in using Bayesian model comparison for two sample hypothesis testing are growing [8; 22; 14]. The type of statistical interpretations we make from Bayesian tests is fundamentally different from NHST. Under the Bayesian framework, we assume there is a prior probability $P(H_1)$ for H_1 (the alternative) to be true, and similarly $P(H_0)$ for H_0 to be true. The ratio between the two is called the prior odds. After collecting data from an experiment, we update prior odds using the Bayes Rule:

$$\frac{P(H_1|Data)}{P(H_0|Data)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(Data|H_1)}{P(Data|H_0)}, \quad (1)$$

which is commonly referred as

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Bayes Factor}.$$

Note that the Bayes Factor (BF) is the likelihood ratio of observing the data between H_1 and H_0 . From the posterior odds, it is straightforward to calculate the posterior probabilities $P(H_1|Data)$ and $P(H_0|Data)$. The test result can be interpreted as follows. Conditioning on a posterior odds K , rejecting H_1 will expose us to a risk of a false rejection/discovery with probability $P(H_0|Data) = 1/(K + 1)$.

In this paper we are interested in a common practice called continuous monitoring or optional stopping. This practice is best described as the following Example.

EXAMPLE 1 (OPTIONAL STOPPING).

We observe data sequentially through time and at any time we can conduct statistical analysis on the data already observed. Let $t = 1, \dots, N$ be all the interim check points that we can take a peek at our A/B test results². For any given metric M , let R_t be its test result at check-point t . We define

¹For this paper, we assume readers are already familiar with the concepts of null hypothesis statistical testing (NHST) in controlled experiments. Readers new to these concepts should refer to references such as Kohavi et al. [16].

²A test result could be something like test statistics, p-value or Bayes Factor/posterior. We use this vague notion when the detail is not important.

an event S_t that is observed at time t and stop the experiment at the first t such that the event happens ($R_t \in S_t$) and return result R_t , e.g. when we deem the result is “significant” or “conclusive”. Typically event S_t is defined as $p\text{-value} < \alpha$ or $P(H_0|Data) < r$. If this event didn’t happen at $t = N$ we return test result R_N . In fact we don’t even need to collect data after that. A general version could have infinite horizon.

Pitfalls of continuous monitoring under NHST framework have been documented in various publications. We say the interpretation of the result is unbiased with continuous monitoring if the validity of the interpretation holds regardless of whether continuous monitoring is used. NHST is valid for fixed horizon test. But it is known to underestimate Type-I error when continuous monitoring is used. To quickly see why, if experimenters are allowed to stop the first time p -value is less than 5%, we will only reject more often, but no less comparing to a fixed horizon design, because the event of rejection in a fixed horizon design, i.e. only reject at time N , is strictly a subset of the event of rejection in the continuously monitoring design. As a result, if the Type-I error in the fixed horizon design is 5%, the Type-I error with continuous monitoring will in general exceed 5%. An application of the law of iterated logarithm shows when incoming data are i.i.d. continuous monitoring will inflate Type-I error to 100% when the horizon N goes to infinity, see Siegmund [24] and Section 6. Johari et al. [12] provided simulation results showing that the inflation of Type-I error is significant and could be typically above 50% or more.

The main result of this paper is the following theorem.

THEOREM 1. *Let \mathbf{X}_t be all the observed data up to time t and BF_t be the Bayes Factor defined as $\frac{P(\mathbf{X}_t|H_1)}{P(\mathbf{X}_t|H_0)}$ and posterior odds $PostOdds_t$ defined as in (1) with known prior odds $P(H_1)/P(H_0)$. Let τ be any stopping time defined by a proper stopping rule, i.e. a mechanism for deciding whether to continue or stop on the basis of only the present and past events and τ is finite almost surely. Then the interpretation of the Bayes Factor and posterior odds remains unbiased with optional stopping at τ . Specifically, we have*

$$\frac{P(H_1|PostOdds_\tau)}{P(H_0|PostOdds_\tau)} = PostOdds_\tau. \quad (2)$$

Theorem 1 says that even though posterior odds is calculated at a random stopping time τ , conditioning on observing a posterior odds of K , rejecting H_1 will expose us to a risk of a false rejection/discovery probability of $1/(K + 1)$. This is a nontrivial result, because BF_τ and $PostOdds_\tau$ are calculated as if τ is a fixed time t . Theorem 1 guarantees that the Bayesian test result remains the same interpretation even with continuous monitoring, provided that the Bayes Factor (and hence the posterior odds) are calculated using all available observations up to the stopping time τ , and the stopping rule is properly defined to be based on only the present and past events. In particular, the theorem does not hold if Bayes Factor is calculated on a selected subset of the observations available at time t , or if the stopping rule peeked ahead into the future. These requirements are met in all common practices of continuous monitoring as in Example 1 where the stopping time is called a “hitting time”. In conclusion, Theorem 1 formally endorsed the practice of continuous monitoring in the framework of Bayesian Hypothesis Testing. This is in stark contrast to NHST, where special

adjustment has to be done. Still, practices like “re-analyze the same data using continuous monitoring after failed to reject using all data” is not supported by Theorem 1. More bad practices are discussed later in Section 5.

At the time of writing, there is still a lack of general agreement on whether continuous monitoring is a proper practice when Bayes test is used, more details in Section 2. The purpose of this paper is to provide arguments assessable by practitioners and engineers, while at the same time provide rigorous proofs for researchers in A/B testing community as well as related fields. With this main goal, the contributions of this paper are

1. We formally prove Theorem 1 in Section 4.
2. We also adopt a simulation based approach to help understand Bayes Factor and posterior odds in a more tangible way, which we believe is more effective for both researchers and practitioners to understand the result and gain intuitions.
3. We discuss practical implications of using Bayesian Hypothesis Testing by comparing it to NHST. We also review and compare to NHST adjustment of continuous monitoring as in Johari et al. [12].
4. For practitioners, we make recommendations on when and when not to use continuous monitoring. We emphasize cases Theorem 1 does not apply.

All model assumptions required in our models are taken as granted. Although both NHST and Bayesian tests make extra model assumptions, the latter requires more such as prior and distribution under H_1 . In practice many people use subjective priors or so called non-informative priors. These practices have been criticized a lot since there are no agreement among researchers and practitioners on which prior is appropriate. However, with the existence of rich historical A/B tests data, Deng [8] showed that we can learn prior objectively from the empirical data. The only assumption we are making when using historical tests data is that we assume the prior behind those historical A/B tests are the same as the current A/B test. See Johnstone and Silverman [13] for a similar technique applied in signal processing. We will provide the algorithm in Appendix B. But the general discussion of objective prior learning is a purely orthogonal topic to this paper and therefore out of our scope.

The rest of the paper is structured as follows. We review related work in the next Section. Section 3 illustrates Theorem 1 using simulation studies. Proof of the main theorem is in Section 4, with intuitive explanations. We emphasize bad practices where Theorem 1 does not apply in Section 5. Section 6 discuss practical implications of using Bayesian Test vs. NHST. We also review and compare to “always valid inference” in Johari et al. [12]. Section 7 concludes the paper with practical recommendations.

2. RELATED WORK

The need of a different theory to allow continuous monitoring in NHST framework has long been known as the subject of sequential hypothesis testing, which dates back to 1945 [25]. Sequential hypothesis testing and later on group sequential testing have been widely used in Clinical Trials, see [1] for a recent survey of the area. The idea of sequential test is only been recently popularized by Johari et al. [12] in A/B testing community, by including it as part of the offering of the commercial A/B testing platform Optimizely. Despite it being newly introduced to A/B testing community,

the theories behind sequential tests under NHST frameworks are well known by statisticians and practitioners in related areas such as clinical trials, psychology, econometrics and other social sciences. There is little dispute about the validity of the methodology.

Bayesian hypothesis testing, on the other hand, is much less accepted and established than its frequentist counterpart. This was largely due to the need of prior knowledge that commonly requires a subjective choice or so called “non-informative” priors which also lack of justification. However, as argued in Deng [8], for A/B testing in the big data era, much of the issue in choosing prior can be mitigated by using historical A/B tests data to empirically learn prior, assuming we have no evidence that the current experiment will have difference chance of success than the past. Putting the issue of choosing a prior aside, many Bayesians have argued that Bayesian reasoning should be immune to stopping rule. For example, Dawid [6] brought up the notion of conditional independence and argued that posterior based on stopping time shouldn’t alter likelihood ratios. This issue is also discussed in Berger and Berry [4] and later Berger and Bayarri [3] referred to the idea as the “stopping rule principle” and said “once the data have been obtained, the reasons for stopping experimentation should have no bearing on the evidence reported about unknown model parameters.” Although the idea is well received by leading Bayesians, there are still a lot of debates going on among researchers and practitioners (who cares more about practical applications instead of entrenched debate between Bayesian and frequentist) on whether Bayesian testing and more generally Bayesian analysis is adjustment-free when optional stopping is applied. John K. Kruschke, a professor of psychological science, author of the well received book *Doing Bayesian Analysis* [17], made the point that Bayesian testing can be biased under optional stopping in 2013 [18]. Andrew Gelman, a professor of statistics and political science, is on the other side and claimed that optional stopping *is* Kosher in Bayesian analysis in 2014 [11]. This debate is also still heated in *Psychon. Bull. Rev.*, a journal where Bayesian hypothesis testing is relatively well received. In a 2014 paper Rouder [21] used simulation to support the case of Bayesian test with optional stopping, and to counter criticisms from Erica et al. [10] and Sanborn and Hills [23], both published also in 2014.

We believe the lack of a general agreement on this issue even in mid 2010s is a clear sign that this is still a big problem for researchers and practitioners in various areas. This is especially the case for A/B testing community because 1) data are always received in near real time in a sequential fashion, 2) the technology enables and even encourages experimenters to frequently check out the test results. We found the simulation argument made in Rouder [21] to be insightful and easier to understand by practitioners and engineers than mathematical formulas. We adopt the same approach and also formally provide rigorous proof of Theorem 1 in Section 4.

3. SIMULATION ILLUSTRATION

Instead of explaining and proving Theorem 1 right away, we present some simulation results first. This section serves two purposes. First, a big part of the debates about whether Bayesian test is biased with continuous monitoring is due to wrong interpretations of the Bayesian test result itself, and

using frequentist measurements such as Type-I error to evaluate a Bayesian test result. This is largely due to most researchers, especially statisticians, are trained with frequentist statistics and methodologies. A correct interpretation of Bayesian posterior odds and Bayes Factor is a prerequisite for readers to understand and appreciate Theorem 1 and rest of the paper. To this end, for most data scientists and engineers, we found replicable simulation results is more tangible and concrete than probability formulas. Secondly, through the simulation results, we hope readers will glimpse some intuitions on why Theorem 1 is indeed expected. Readers who are not interested in rigorous proof can even skip the formal proof in the next section.

Important: Recall

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Bayes Factor}.$$

Prior odds is considered known and is independent of the observations collected for the test. Prior odds is easy to interpret and interpretation of Bayes Factor, hence the posterior odds is the essence of a Bayesian test. Without loss of generality, from now on we will assume a prior odds of 1:1, *i.e.* H_1 and H_0 are equally likely based on our belief without seeing any data. In this case *Posterior Odds and Bayes Factor are the same*. Readers can treat them as interchangeable in this paper.

3.1 Bayes Factor and Fixed Horizon Test

To correctly understand the meaning of Bayes Factor, we first look at the vanilla case — a fixed horizon test first. In a fixed horizon test experimenters prescribe the total sample size, or equivalently, length and traffic of the A/B test first and only consider the test result when we collected all the data as the final result.

We consider a simple problem of testing a normal mean. We observe N *i.i.d.* observations $X_i, i = 1, \dots, N$ from a normal distribution $N(\mu, 1)$ with unknown mean μ . $\mu = 0$ under the null hypothesis H_0 , and $\mu = \delta$ under the alternative hypothesis H_1 . Equivalently, sample mean \bar{X} is the sufficient statistics and it has distribution $N(0, 1/N)$ under H_0 and $N(\delta, 1/N)$ under H_1 . Note that this is even simpler than a A/B test because there is only one group. For two sample A/B test we replace \bar{X} by $\Delta = \bar{X}_T - \bar{X}_C$, *e.g.* difference of two sample means, and the test is essentially the same as one sample test. See Appendix A.

The Bayes Factor is

$$\frac{P(\bar{X}|H_1)}{P(\bar{X}|H_0)} = \frac{\exp(-(\bar{X} - \delta)^2/(2/N))}{\exp(-(\bar{X})^2/(2/N))} = \exp\left(\frac{N}{2}\delta(2\bar{X} - \delta)\right) \quad (3)$$

Conditioning on observing a \bar{X} , if we plug it into Equation 3 and get a number K , what does it mean? To illustrate this, we simulate 100,000 runs and each run we simulate $N = 100$ observations $X_i, i = 1, \dots, N$. Since we assume prior odds 1:1, we simulate 50,000 runs under H_1 , where $X_i \sim N(\delta, 1)$ and the other 50,000 runs under H_0 where $X_i \sim N(0, 1)$. At the end of each run, we calculate Bayes Factor based on Equation 3. The end result of this simulation is 100,000 Bayes Factors, half of them are from H_1 and half of them from H_0 .

We did this simulation for $\delta = 0.2$ and $N = 100$. Figure 1 shows histograms of those Bayes Factors in log scale, grouped by the ground truth models H_0 and H_1 . Bayes Factors from H_1 are shown on top and those from H_0 are shown

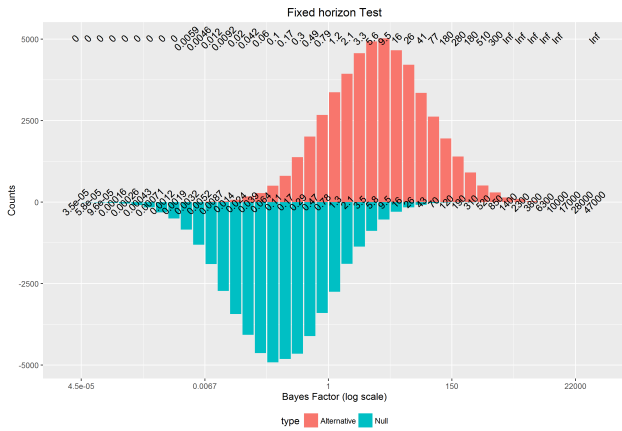


Figure 1: Histograms of Bayes Factor simulated from both models. For each bin, the number on top are the ratio of simulated Bayes Factors from alternative to those from null.

at the bottom. What does it mean if we observe a Bayes Factor of 2.1? On Figure 1, we first group Bayes Factors close to 2.1 together in to the same bin. There are about 4,000 runs from H_1 (height of the red bar) that produced a Bayes Factor close to 2.1 (among 50,000 simulation runs), while around 2,000 (height of the blue bar) are from H_0 (among 50,000 simulation runs). The actual ratio of those from H_1 to H_0 is shown on top of the plot and is 2.1, which is the same as the Bayes Factor 2.1 we started with. In fact, if you go through each bin carefully in Figure 1, you will find the number on the x-axis, which represents Bayes Factor value calculated from Equation 3 and grouped into each bin, are very close to the actual observed count ratio of those from H_1 (top red) to those from H_0 (bottom green), except those at the far tail on both sides. Is this a coincidence? Of course not. When observed a Bayes Factor of K , we know both model H_0 and H_1 could result in such a Bayes Factor. This simulation we did let us replay the data generation process and observe how likely it is for H_1 to generate such a Bayes Factor and how likely for H_0 respectively, which are represented, after binning similar Bayes Factors together, by the height of the top and bottom histograms. Our interest is the odds of this Bayes Factor being from H_1 to H_0 , which is the ratio of heights between the red bar and the green bar. We will expect the observed ratio to be close to the true underlying odds, within some small expected error due to 1) simulation randomness and 2) discretization used in binning similar Bayes Factor together. The error from simulation randomness is smaller for those center bins, *i.e.* bins where more Bayes Factors are observed from 100,000 simulation runs, and are larger for those at the tails³. What we observed so far can be summarized as:

$$\begin{aligned} \text{True underlying odds} &= \text{Observed ratio} \\ &= \text{Bayes Factor calculated from Equation 3.} \end{aligned}$$

This simulation illustrated two things

1. Bayes Factor can be conceptually “materialized” as the ratio of the bar heights from the H_1 histogram and H_0

³Some bins on the two tails are either showing an observed ratio of 0 or Inf, for the obvious reason.

histogram. An observed Bayes Factor of K means it is K times more likely to be generated from H_1 than H_0 .

2. For the fixed horizon case, Equation 3 is the same as the true odds (at least they must be very close).

The fact that Equation 3 is the correct Bayes Factor is no news at all. We just followed the definition of Bayes Factor, and the observation so far confirms that we didn’t make any mistake in Equation 3. In other words, without continuous monitoring, a fixed horizon Bayesian test of H_1 vs. H_0 as in this section using Equation 3 provides the true underlying odds we can use to make decision. This is a special case of Theorem 1 where the stopping rule is to only stop at N , and (2) says if we calculate posterior odds which is just Bayes Factor here using $P(\mathbf{X}_N|H_1)/P(\mathbf{X}_N|H_0)$, it is the same as the true underlying odds conditioning on observing such a posterior odds, the left hand side of (2). We are now ready for some real stopping rules.

3.2 Stopping Rule Based on Bayes Factor

If rejecting H_0 when observing a posterior odds no less than K exposes us to a risk of false discovery at most $1/(1+K)$, a natural stopping rule is to prescribe a false discovery rate (FDR) bound and stop the test immediately if observed posterior odds already can guarantee the FDR control. We can set $K = 9$ to guarantee a FDR bound of $10\% = 1/(9+1)$.

Similarly, we can early stop for futility and accept H_0 if we believe posterior of null is sufficiently large. A symmetric design is to stop if posterior odds is either no less than K or no greater than $1/K$.

Figure 2 illustrated both stopping rules under the same setup in fixed horizon case of Section 3.1, except for we are stopping at the first $t \leq N$ when stopping rule is satisfied. In each of the 100,000 simulation runs, regardless of whether this run is early stopped or stopped in the end, we always calculate Bayes Factor based on Equation 3, replacing N by the observed stopping time τ . Comparing Figure 2 to Figure 1 shows big differences. The biggest one being the spike at the stopping Bayes Factor boundary $K = 9$ (and $1/9$ for futility). However, the interesting observation is, despite the huge histogram shift for both H_1 (red) and H_0 (green), those numbers on the top margin — ratios of observed Bayes Factors in each bin from H_1 to H_0 , remains very close to the theoretical Bayes Factor value calculated from Equation 3, as in a fixed horizon test. This is exactly what Theorem 1 claims, and this simulation study confirms it!

For many who are used to the frequentist thinking of controlling Type-I error, this result seems odd. If we allow early stop, and still using the same rejection criteria of $\text{BF} > 1/K$, we will only reject more so we will be inflating the Type-I error. This is correct, but nonetheless does not conflict with the fact that FDR is still controlled below the designed level.

Table 1: Impact of Early Stopping

| | Type-I | Power | FDR | Early Stop Rate | |
|----------------|--------|-------|-------|-----------------|-------|
| | | | | H_1 | H_0 |
| Fixed Horizon | 0.018 | 0.465 | 0.037 | NA | NA |
| One-sided Stop | 0.060 | 0.599 | 0.09 | 59.5% | NA |
| Two-sided Stop | 0.060 | 0.598 | 0.09 | 64.9% | 65.0% |

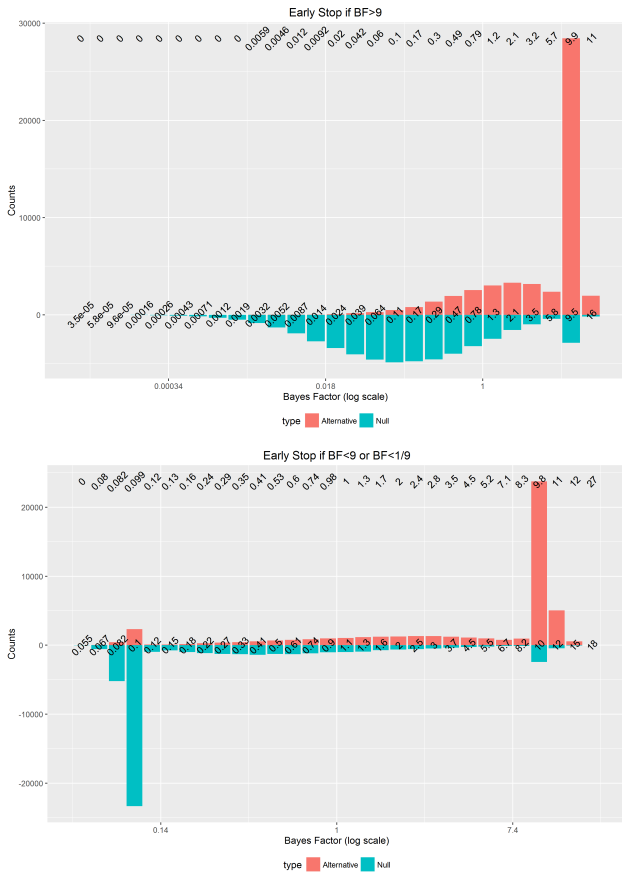


Figure 2: Histograms of simulated Bayes Factor with optional stopping. Top plot: one-sided stopping. Bottom plot: two-sided stopping.

Table 1 shows the comparison of three simulation studies we did so far in terms of Type-I error, power and FDR. In the fixed horizon design, when we reject for $BF > 9$, the Type-I error (proportion of false rejection among the null cases) is 0.018. This value increased to 0.06 when continuous monitoring/optional stopping is introduced. Because we are rejecting more, the power of the test is also improved from 0.465 to 0.599. (Power in two-sided test is 0.598, slightly smaller than in the one-sided test. We leave it to user to figure out why.) FDR in the finite horizon cases is only 0.037, smaller than the designed bound of 0.1. This is because in the finite horizon cases a lot of rejected cases at the end of the test are actually bearing a BF much larger than the threshold 9, see Figure 1. This suggests that in finite horizon test, using a BF cutoff to calculate FDR might be conservative, also see Efron [9] for the differences of local FDR and FDR. When optional stopping is introduced, FDR become 0.09, very close to the designed level. The small discrepancies here is due to over-shoot, *i.e.* we stop once BF is larger than 9 but not exactly at 9. These over-shoots are reflected in Figure 2 where we found a few bars beyond the spike. In large sample scenario where each individual observation won't make a big change in BF, as in most A/B tests, we can think of the time series of BF_t as continuous. In this case we can stop the test with a BF almost exactly equal

to 9, and the FDR will be also almost exactly 0.1. We saw that FDR control in the fixed horizon setting is conservative because we are wasting sample sizes to collect evidence beyond what we really need, and with early stopping the waste is mitigated. The last two columns in Table 1 shows the percentage of the simulated experiment with early stopping. We saw majority of the simulated runs stopped earlier. We also calculated that the average length of the simulated runs with early stopping is about 55, much smaller than the fixed horizon of $N = 100$. Based on Table 1, one could argue that early stopping is always superior than the fixed horizon test, and should be recommended. More discussions are in Section 6.

3.3 General Stopping Rule

Theorem 1 holds for general stopping rules, not only those based on BF cutoff values. For experimenters who want to “hack” p-values, they could choose to stop once p-value is less than α . Here we did the simulation study with the stopping rule with both criteria: 1) p-value less than 0.1, and 2) the sample sizes is at least 10.

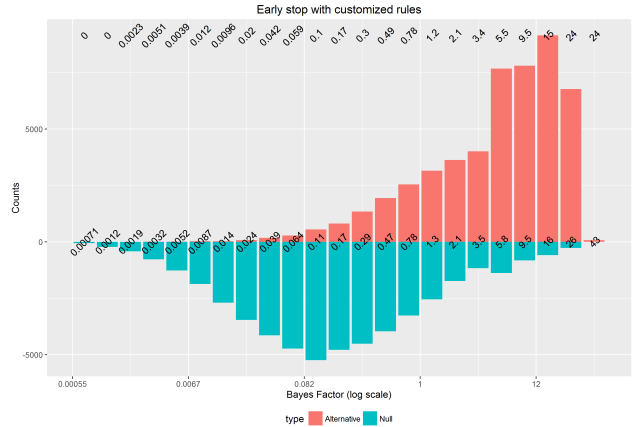


Figure 3: Using a stopping rule based on p-value and minimum required sample size.

Figure 3 shows the simulation results, this time with a rather bizarre histogram for H_1 runs. The important part is the observed actual ratio on the top margin still closely tracks the theoretical Bayes Factor values on the x-axis.

3.4 Composite Alternative

So far in this section we have been using a overly simple alternative model H_1 where the treatment effect is assumed to be known. This is not very realistic since we never know the effect so that alternative is always a composite alternative where δ can be anything nonzero. In Bayesian model comparison we need to put a prior distribution for δ under H_1 , in addition to the prior odds. Following [12] and [8], we put a normal prior $N(0, \sigma_0^2)$. Under this H_1 , $X_i \sim N(0, \sigma_0^2 + 1)$ and the formula for Bayes Factor assuming a fixed sample size N changes to

$$\frac{N(\bar{X}; 0, \sigma_0^2 + 1/N)}{N(\bar{X}; 0, 1/N)} \quad (4)$$

A similar simulation to those above in this section is run by setting $N = 1,000$. We also set $\sigma_0 = 0.1$ to generate

50,000 independent δ first for each of the simulation runs from H_1 . At the end of each runs(or at the stopping time) we compute Bayes Factor based on (4) with N for fixed horizon setting or τ in its place when optional stopping is introduced. Figure 4 shows the results for both fixed horizon setting and optional stopping with BF cutoff at 9. In the fixed horizon setting, the histogram is much more dispersed than the previous precise alternative case. Some BF is as large as several thousands so we only show those no greater than 100. Early stopping effectively eliminated those extremely large BF, creating spikes around 9. We hope readers at this point already noticed that the top margin numbers are very close to the theoretical BF values on the x-axis.

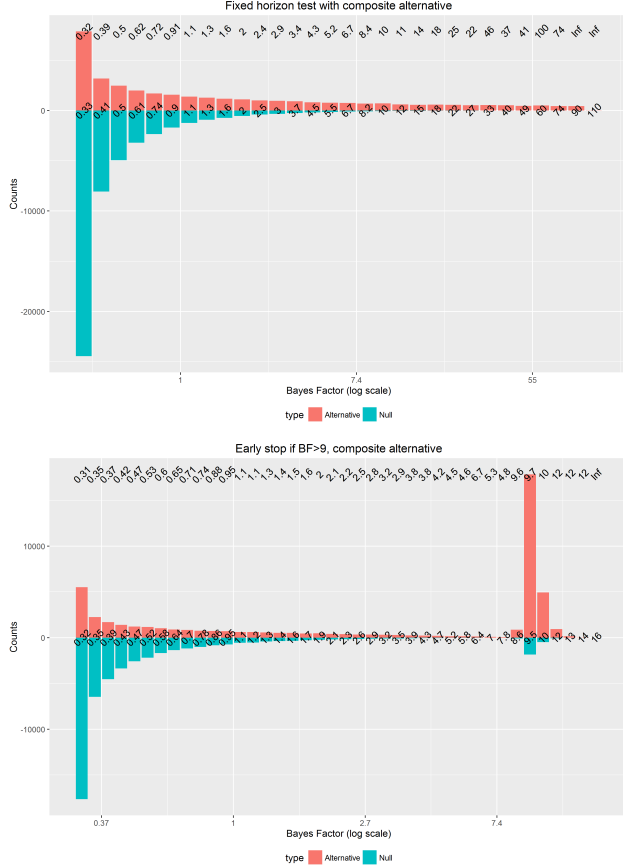


Figure 4: Composite H_1 with Normal prior. Top: fixed horizon(x-axis limit to 100). Bottom: early stopping when $BF > 9$.

4. PROOF OF MAIN THEOREM

We now prove Theorem 1. Readers who only need intuition are recommended to skip the proof and jump to Section 4.1. Both sides of the (2) are random variables depending on $PostOdds_\tau$. It is equivalent to the following:

$$\begin{aligned} \frac{P(H_1|PostOdds_\tau = K)}{P(H_0|PostOdds_\tau = K)} &= \frac{P(H_1 \text{ and } PostOdds_\tau = K)}{P(H_0 \text{ and } PostOdds_\tau = K)} \\ &= \frac{P(PostOdds_\tau = K|H_1)}{P(PostOdds_\tau = K|H_0)} \times \frac{P(H_1)}{P(H_0)} = K, \end{aligned}$$

for any K where $P(PostOdds_\tau = K) > 0$. Let $K' = K \times P(H_0)/P(H_1)$. The event $\{PostOdds_\tau = K\}$ is equivalent to $\{BF_\tau = K'\}$. The last equality above after rearranging the prior odds $P(H_1)/P(H_0)$ to the right side becomes

$$\frac{P(BF_\tau = K'|H_1)}{P(BF_\tau = K'|H_0)} = K'. \quad (5)$$

Without loss of generality, we only need to prove (5).

We first prove for the fixed horizon case, which is a direct result of *likelihood ratio identity*, or *change of measure*. For any fixed t , let $\mathbb{Q}_t = P(\cdot|H_1)$ and $\mathbb{P}_t = P(\cdot|H_0)$ be the probability measure under H_1 and H_0 respectively for observations up to t , both have a density function w.r.t. to Lebesgue measure on the real line. Let A be any event observable at time t .⁴ The likelihood ratio identity⁵ ensures

$$\mathbb{Q}_t(A) = E_{\mathbb{P}_t} \left(\mathbb{1}_A \times \frac{d\mathbb{Q}_t}{d\mathbb{P}_t} \right),$$

where $d\mathbb{Q}_t/d\mathbb{P}_t$ is the likelihood ratio and $\mathbb{1}_A$ is the binary indicator function for event A . Recall Bayes Factor is defined as the likelihood ratio. Replace $d\mathbb{Q}_t/d\mathbb{P}_t$ by BF_t , set $A = \{BF_t = K'\}$ in the above to get

$$\mathbb{Q}_t(BF_t = K') = E_{\mathbb{P}_t}(\mathbb{1}_A) \times K' = K' \times \mathbb{P}_t(BF_t = K'), \quad (6)$$

which is (5).

We can generalize this argument for random time τ . Theorem 1 requires τ to be a stopping time⁶ so that the event $\{\tau = t\}$ is observable at time t . This is a necessary requirement to ensure that we can apply the likelihood ratio identity for the event $\{BF_t = K' \text{ and } \tau = t\}$ (observable at t) to get

$$\mathbb{Q}_t(BF_t = K', \tau = t) = K' \times \mathbb{P}_t(BF_t = K', \tau = t). \quad (7)$$

If τ can only take value from 1 to a maximum horizon N (experiment stop at N no matter what, which covers all practical cases), summing up (7) over all t entails

$$\begin{aligned} P(BF_\tau = K'|H_1) &= \sum_{t=1}^N P(BF_\tau = K', \tau = t|H_1) \\ &= \sum_{t=1}^N \mathbb{Q}_t(BF_t = K', \tau = t) = \sum_{t=1}^N K' \mathbb{P}_t(BF_t = K', \tau = t) \\ &= \sum_{t=1}^N K' P(BF_\tau = K', \tau = t|H_0) = K' P(BF_\tau = K'|H_0) \end{aligned}$$

which is (5) and the proof is completed. Notice how we changed $BF_\tau = K'$ to $BF_t = K'$ once we restrict ourselves to the set $\tau = t$ in the second and fifth equality. The essence of the proof is to show

$$\frac{P(BF_\tau = K', \tau = t|H_1)}{P(BF_\tau = K', \tau = t|H_0)} = K' \quad (8)$$

⁴ $A \in \mathcal{F}_t$ where \mathcal{F}_t represents the set of measurable events at time t . \mathcal{F}_t is called a filtration because $\mathcal{F}_s \subseteq \mathcal{F}_t$ for any $s \leq t$.

⁵ It is also called change of measure identity because the equation transforms an expectation under a measure \mathbb{P} into an expectation under another measure \mathbb{Q} . This is a special case of the Radon-Nykodym Theorem in measure theory.

⁶ τ is a stopping time w.r.t. to a filtration \mathcal{F}_t if $\{\tau \leq t\} \in \mathcal{F}_t$.

for every $t \leq N$ by applying likelihood ratio identity (7) and then sum up both numerator and denominator of (8) over t to recover (5). For potentially unbounded τ , we just need to sum up to infinity and the result still holds because the sum series of both numerator and denominator of (8) are finite.

Important Remark. We make the remark that Theorem 1 **does not** require observation \mathbf{X} to be sequential *i.i.d.* observations as in earlier simulation examples. All we used in the proof is the likelihood ratio identity for *the whole path of observations \mathbf{X}_t up to t for any t* . In a typical A/B test with user level tracking, users who visit the site multiple times will provide multiple observations sequentially. Since there is a strong between-user correlation, when we look at \mathbf{X}_t , which includes all sequential observations (page-views) from different users, they are not independent. However, at any given time t , we can always first aggregate \mathbf{X}_t to the randomization unit, which is user. Take the metric Revenue per user as example, \mathbf{X}_t is the sequence of revenues for each page-view up to time t . For each user, we can sum up revenues as $Y_{it}, i = 1, \dots, N_t$ where N_t is number of unique users. We can treat Y_{it} as *i.i.d.* when computing likelihoods under both H_1 and H_0 . For ratio metrics such as Click-Through-Rate(CTR), Y_{it} can either be CTR for each user and average of Y_{it} is the average CTR over all users — a double average metric, or Y_{it} can be a pair $(Clicks_{it}, PageViews_{it})$ and the metric is the sum of clicks over all users divided by sum of page-views. Delta Method is required to compute the likelihood for the latter case. The main point is that the original sequential observations might not be *i.i.d.* but after aggregated to randomization unit level, likelihood ratio can be easily calculated using those aggregated values which can be assumed *i.i.d.* because of the randomization design.

4.1 Intuitive Explanation

There is an intuitive explanation of Theorem 1 using the same simulation procedure in Section 3 as a thought experiment. Again we assume prior odds is 1:1 so Posterior Odds equals to the Bayes Factor and (2) becomes (5).

We simulate M paths from H_1 and H_0 . M is a very large number, almost infinite. So every path simulated from H_1 which has nonzero probability under H_0 will have the same path simulated under H_0 , and vice versa. For each path we simulate the whole path \mathbf{X} up to the fixed final horizon N . For any $t \leq N$ and any path \mathbf{X} , we can calculate a Bayes Factor at time t to be $BF(\mathbf{X}_t)$ as likelihood ratio $P(\mathbf{X}_t|H_1)/P(\mathbf{X}_t|H_0)$. No stopping rule has been introduced yet so everything so far belongs to the fixed horizon case. Define $Path(\mathbf{X}_t|H_i), i = 0, 1$ to be the set of all paths simulated from $H_i, i = 0, 1$ having the same subpath \mathbf{X}_t up to time t , and let $|Path(\cdot)|$ denote the total number of paths in a path set, *i.e.* cardinality. Then for any t and any path \mathbf{X} with subpath \mathbf{X}_t , $|Path(\mathbf{X}_t|H_1)|/|Path(\mathbf{X}_t|H_0)| = BF(\mathbf{X}_t)$. Intuitively, this means for every subpath \mathbf{X}_t simulated from H_0 , there are on average $BF(\mathbf{X}_t)$ exact subpaths simulated from H_1 . Because this statement is true for any subpath. If we only look at subpaths such that $BF(\mathbf{X}_t) = K$, we have $|Path(\{\mathbf{X}_t : BF(\mathbf{X}_t) = K\}|H_1)|/|Path(\{\mathbf{X}_t : BF(\mathbf{X}_t) = K\}|H_0)| = K$, for any K .

Now we introduce stopping rule. Pick any path \mathbf{X} simulated from H_0 , say the stopping rule will stop at time t and we computed the Bayes Factor to be K . The previous argument shows there will be on average K number of the

the same exact subpath simulated from H_1 . Here comes the important part! *Because the stopping rule does not depend on observations after the stopping time, all subpaths simulated having the exact same subpath \mathbf{X}_t up to t will also have the exact same stopping time at t !* (See the next section for examples of bad stopping rules where this property is not true, hence Theorem 1 does not apply.) After we gathered all paths simulated from H_0 with the same Bayes Factor K at time t which also stopped at time t according to the stopping rule, for each one of them we can find K exact same subpaths which *also stopped at time t* . By one more step of gathering all such set of paths for every possible $t \leq N$, it is then intuitively clear that the number of paths gathered together from H_1 and H_0 have a ratio of exactly K . This is exactly what we tried to demonstrate via various simulations in Section 3.

5. BAD PRACTICES

Theorem 1 is a general result with very mild assumptions which are satisfied in most cases. But failure of satisfying those assumptions can result invalid test results. We list three bad practices so readers can be aware of the limitations of the result in this paper. One critical assumption is that the stopping rule is properly defined that only uses information already observed, without peeking into the future. One example for an improper stopping rule is to reassess all the observations at some time t' , and then decide to only use the data up to an earlier time $t < t'$, *e.g.* stop at t after seeing data at a later time t' . This practice is called *data snooping* and is not supported by Theorem 1. There are two common bad practices related to data snooping:

EXAMPLE 2 (RE-ANALYSIS AFTER FAIL TO REJECT).
Finite horizon test at N failed to reject H_0 . The same data is then reanalyzed using continuous monitoring as in Example 1.

EXAMPLE 3 (OPTIMAL STOPPING).
The basic setup is the same as in Example 1. This time we first collect all the data up to finite horizon N . Then, we look at our data and try to find the best check-point t so the test result R_t is the most favorable. The difference between this example and continuous monitoring is that for the latter the decision of stopping the experiment is made without peeking at the data in future.

In both examples above, if we collected all the data up to horizon N and *did the test*, we should always report this test result instead of re-analysing the data or try to cherry pick the optimal stopping time. This is because Bayesian test is consistent: as we observe more data, posterior $P(H_1|Data)$ converges to 1 if H_1 is true and to 0 otherwise. This means we should always prefer the decision made from more data. However, it is possible that continuous monitoring might have rejected H_1 (if it were used) but the finite horizon test at N does not. Does that mean continuous monitoring makes more error? No! Table 1 shows continuous monitoring does increase the amount of null rejection, without sacrificing the false discovery rate. In the above case the posterior odds realized in the end of the experiment at horizon N shows H_1 is unlikely to be true. It is fair to say *if we had been using continuous monitoring, we would likely be making a false discovery at that time, based on newer observations*. However, let K' be the posterior odds reported by

continuous monitoring, Theorem 1 guarantees that it is K' to 1 odds that we will see posterior odds increasing to ∞ if we keep getting more data, than decreasing to 0. In other words, it is $K' - 1$ more likely our decision still uphold in the end of the experiment, than reversed as in the hypothetical case.

Another critical assumption in Theorem 1 is that the likelihood ratio has to be correctly calculated with all available observations, *i.e.* the whole subpath \mathbf{X}_t . In particular, we cannot cherry pick only those observations that favors one hypothesis. Here is another bad example which happens a lot in practice.

EXAMPLE 4 (CONTINUOUS TESTING UNTIL WIN).

With agile development and continuous A/B testing, a team can iteratively modifying and testing a feature until seeing a successful test result.

In NHST, even if the feature has no effect, there is still α (typically 0.05) chance that the result could be statistically significant. This means for every 20 iterations, we might just declare a success without really having any true effect. This is like continuous monitoring, but the difference is that here each new test only uses its own data. Using Bayesian test, if we need to calculate the likelihood ratio up to the t -th test, we have to aggregate all the evidence together, not just looking at the last one. If all tests are independent replications of the same test, aggregating evidence in Bayes test is trivial, we just need to multiply likelihood ratio for each of the replications all together. This way even if we might have a few large likelihood ratio favoring H_1 , but if H_0 is the ground truth there have to be more smaller likelihood ratios so the product is small. In fact it will converge to 0 if we keep doing replication runs. In practice, since iteration runs are not exactly replications, it is still a challenge how we should properly aggregate evidence from multiple experiment runs together. Ignoring the prior runs can still result in more false discovery than nominally controlled. Technically this is the area of multiple testing and selection bias. See Lu and Deng [20] for some preliminary results.

6. COMPARE TO FREQUENTIST METHODS

6.1 NHST

We compare Bayes testing to NHST in this section to reveal why continuous monitoring is an issue for NHST but not for Bayes testing. For simplicity, we assume $X_i, i = 1, \dots$ are sequential *i.i.d.* observations from $N(\mu, 1)$.

First we test $H_0 : \mu = 0$ against $H_1 : \mu = \delta$. In NHST we reject H_0 when the z-statistics is larger than a constant threshold, while in Bayes testing we reject when the Bayes Factor (3) is larger than a constant threshold. Simplifying both rejection boundary resulted in

$$|\sqrt{n}\bar{X}_n| > C_1 \quad \text{NHST}, \quad (9)$$

$$|\sqrt{n}\bar{X}_n| > C_2\sqrt{n} + C_3/\sqrt{n} \quad \text{Bayesian}, \quad (10)$$

where $C_i, i = 1, 2, 3$ are constants. Asymptotically, NHST reject when test statistics is larger than $O(1)$ and the Bayesian test's rejection boundary is $O(\sqrt{n})$. Why $O(\sqrt{n})$? The reason of having a $O(\sqrt{n})$ in this case is due to the fact that we put a precise alternative $\mu = \delta$ for H_1 . If H_1 is true, by central limit theorem, we would expect most of times we

should observe \bar{X} within $1/\sqrt{n}$ neighborhood of δ . If H_0 is true, then \bar{X} should be within $1/\sqrt{n}$ neighborhood of 0. In NHST, we only assess the likelihood of H_0 , so we reject if \bar{X} is outside of $1/\sqrt{n}$ neighborhood of 0. Bayesian testing is symmetric, *i.e.*, we compare the likelihood of H_1 vs H_0 . In this case since both hypothesis expect observation to be within their $1/\sqrt{n}$ neighborhood, the "classification" line should naturally be for \bar{X} at the midway $\delta/2$, which means $\delta/2 \times \sqrt{n}$ for z-statistics.

Is a rejection boundary of $O(\sqrt{n})$ means Bayesian test will be much less sensitive than NHST? Not necessarily. First of all, if H_1 is indeed true and X_i will have mean δ , then the chance that we observe \bar{X} within \sqrt{n} neighborhood of δ is high, meaning it should not be too much problem for the z-statistics to breach the $O(\sqrt{n})$ boundary. Secondly, if we replace the precise alternative to a composite alternative with a normal prior $N(0, V^2)$ on μ under H_1 , similar calculation reveals an asymptotic rejection boundary of $O(\sqrt{\log(n)})$, very close and almost can be considered $O(1)$ in practice⁷.

But why not simply $O(1)$? Law of iterative logarithm (LIL) shows in the *i.i.d.* sequential X_i case $\sqrt{n}|\bar{X}|$ even under H_0 will breach the boundary $C\sqrt{\log(\log(n))}$ infinitely often for any C . Any rejection boundary that is not at least $O(\sqrt{\log(\log(n))})$ will be breached with probability 1 under H_0 , and with Type-I error 1 with continuous monitoring, when $n \rightarrow \infty$. If Bayesian test have a rejection boundary not bigger than $O(\sqrt{\log(\log(n))})$, it will reject H_0 every time, and the false discovery rate is solely determined by the prior odds. This is in contradiction to the result of this paper. In other words, because of Theorem 1, any valid Bayes test will have a rejection boundary larger than $O(\sqrt{\log(\log(n))})$.

In practice, it is true that Bayesian tests are in general more conservative than NHST. Partly it is because the rejection boundary need to be at least $O(\sqrt{\log(\log(n))})$ —a price to pay for continuous monitoring. Another more important part is because Prior Odds $P(H_1)/P(H_0)$ usually favors H_0 , and sometimes a lot, *e.g.* knowing certain metric rarely truly moved.

6.2 Sequential Tests

Since fixed horizon NHST does not control Type-I error well, corrections have to be made under frequentist framework to keep controlling Type-I error at desired level. Wald [25] first introduced sequential probability ratio test (SPRT). When assuming both H_0 and H_1 are precise, *e.g.* simple alternative, SPRT make decision based on the likelihood ratio $P(\text{Data}|H_1)/P(\text{Data}|H_0)$ and reject H_0 when $LR > B$ and accept H_0 when $LR < A$. The bounds A and B can be chosen such that Type-I error and Type-II error can be controlled at desired levels. In practice we usually don't know the exact treatment effect under the alternative H_1 . Lai [19] extended the idea of generalized likelihood ratio test into sequential setting. Johari et al. [12] introduced the notion of "always valid inference", where they used a variant of Wald's SPRT with a normal distribution on H_1 . It is called mSPRT (m stands for *mixture*). From a Bayesian test perspective, this is equivalent to putting a normal prior for the average effect size μ under H_1 , same as in Section 3.4, also see

⁷Sample size in practice are typically restricted to 1% of all traffic to 100%, only an order of magnitude of 2.

Appendix B. However, despite the similarity in the form of likelihood ratio there are two main differences:

1. mSPRT does not take prior knowledge into account, while Bayesian test encode these information as prior odds $P(H_1)/P(H_0)$. We know from experience, and from historical that some metrics are easier to move than the others. For mSPRT the rejection boundary for all metrics are the same. In A/B testing, it is reported that most ideas fail to deliver desired movement, or even move the success metric. Prior odds $P(H_1)/P(H_0)$ for most metrics are less than 20%.
2. A more fundamental difference is the interpretation of the results. mSPRT controls Type-I error — the chance of false rejection when H_0 is true, while Bayesian test controls False Discovery rate (FDR) — the chance of false rejection when decided to reject H_0 . There is no simple relation ship between the two. When we reject more aggressively, Type-I error will increase, but FDR does not necessarily increase, as long as more aggressive rejection will also reject more true positives. FDR also extends to multiple testing cases easier than Type-I error⁸. Johari et al. [12] also proposed a way to generalize the classic Benjamini-Hochberg procedure [2] to sequential setting to control FDR of mSPRT tests in multiple testing scenario. This additional step is not necessary in Bayesian test.

For both mSPRT and Bayesian test with normal alternative model under H_1 , the asymptotic rejection boundary for test statistics are both $O(\sqrt{\log(n)})$. Both methods makes the same assumption about the normal alternative model. Bayesian test makes extra assumption that prior odds are also known. To use either method, we strongly suggest learning parameters from empirical data, as explained in appendix, see Deng [8] for more detail.

7. CONCLUSION & RECOMMENDATION

With Theorem 1 rigorously proved, we believe the debate over whether continuous monitoring is a valid practice when Bayesian Hypothesis Testing is used should be settled. The answer is unequivocally yes, in the sense that the Bayesian posterior remains *unbiased* (2) when a *proper* stopping rule is used. We emphasize that the correct understanding of (2) and interpretation of Bayesian test result as controlling FDR is critical. Many researchers, even some versed in Bayesian statistics [17], made the mistake of evaluating Bayesian test results conditioning on either null is true or alternative is true. The correct Bayesian interpretation always requires a prior odds weighing the alternative and the null. Our simulation illustrations in Section 3 serve the very goal of helping readers understand this crucial point in a non-technical way.

Two natural questions are then raised by practitioners. 1) Because the fundamental difference in the statistical conclusions we can make from NHST and Bayesian test, which one shall we use in practice? 2) Is the result of this paper telling us when we use Bayesian test we should always use continuous monitoring?

⁸FDR was originally brought up in frequentist multiple testing scenario. FDR does not make much sense in frequentist framework for single test because H_0 for a test is assumed to be either true or false. FDR is 1 if H_0 is true and 0 otherwise. There is no room in the middle with prior probability.

For the first question, we focus on the difference of controlling Type-I error and FDR. If false rejection of any single test cost us a lot, and the cost of false rejection is considered higher than false negative(fail to reject true H_1), then Type-I error seems to be a better criterion to control. A good example for this case is clinical trial. If our goal is not focus on each individual test, but the overall performance of our decision on a large set of tests, and the cost of false rejection and false negative are in the same order, then we believe FDR is a better criterion. Large scale A/B testing platform is an example of the latter [15]. In an agile environment, where success are built upon a lot of small gains, as long as we are shipping more good features really meet customer needs than useless ones, we are moving in the right direction.

For the second question, continuous monitoring is not always recommended. In many cases, the goal of the experiment is not only to confirm the existence of the treatment effect, but also to measure it. In A/B tests, it is not uncommon for a feature to have time-varying treatment effect such as weekday and weekend differences. To capture the weekly cycle, running tests for a whole week or multiple of weeks are often recommended. It is also possible that the treatment effect only exists in the weekend and we might early stop the experiment during the first few weekdays when we use continuous monitoring with an early acceptance of H_0 . But continuous monitoring should be recommended in many other scenarios. Shutdown a bad experiment is one application. We want to shutdown an experiment once we have enough evidence that the treatment is giving user a very bad experience. Another example is comparing a few closely related alternative candidates, e.g. tuning parameters for a backend algorithm, in which case we might assume the *ordering* of the treatment effects won't be time-varying and hence we can early stop inferior candidates and ramping up outperforming candidates based on Bayesian posterior. The last example is studied in more detail in the literature of multi-armed bandit and Thompson sampling [5] and the result of this paper justifies Thompson sampling for using Bayesian posterior to dynamically change data gathering.

References

- [1] Bartroff, J., Lai, T. L. and Shih, M.-C. [2012], *Sequential experimentation in clinical trials: design and analysis*, Vol. 298, Springer Science & Business Media.
- [2] Benjamini, Y. and Hochberg, Y. [1995], 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J. R. Stat. Soc. Ser. B* pp. 289–300.
- [3] Berger, J. O. and Bayarri, M. J. [2004], 'The Interplay of Bayesian and Frequentist Analysis', *Stat. Sci.* **19**(1), 58–80.
- [4] Berger, J. O. and Berry, D. A. [1988], 'The relevance of stopping rules in statistical inference', *Statistical decision theory and related topics IV* **1**, 29–47.
- [5] Chapelle, O. and Li, L. [2011], An empirical evaluation of thompson sampling, in 'Advances in neural information processing systems', pp. 2249–2257.
- [6] Dawid, A. P. [1979], 'Conditional independence in statistical theory', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–31.
- [7] Dempster, A. P., Laird, N. M. and Rubin, D. B. [1977], 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. Ser. B* **39**(1), 1–38.
- [8] Deng, A. [2015], Objective bayesian two sample hypothesis testing for online controlled experiments, in 'Proceedings of the 24th International Conference on World Wide Web', pp. 923–928.

- [9] Efron, B. [2012], ‘Large-scale simultaneous hypothesis testing’, *Journal of the American Statistical Association*.
- [10] Erica, C. Y., Sprenger, A. M., Thomas, R. P. and Dougherty, M. R. [2014], ‘When decision heuristics and science collide’, *Psychonomic bulletin & review* **21**(2), 268–282.
- [11] Gelman, A. [2014], ‘Stopping rules and bayesian analysis’.
- [12] Johari, R., Pekelis, L. and Walsh, D. J. [2015], ‘Always valid inference: Bringing sequential analysis to a/b testing’, *arXiv preprint arXiv:1512.04922*.
- [13] Johnstone, I. M. and Silverman, B. W. [2004], ‘Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences’, *Annals of Statistics* pp. 1594–1649.
- [14] Kass, R. and Raftery, A. [1995], ‘Bayes factors’, *J. Am. Stat. Assoc.* **90**(430), 773–795.
- [15] Kohavi, R., Deng, A., Frasca, B., Xu, Y., Walker, T. and Pohlmann, N. [2013], ‘Online controlled experiments at large scale’, *Proc. 19th Conf. Knowl. Discov. Data Min.*.
- [16] Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R. M. [2009], ‘Controlled Experiments on the Web: survey and practical guide’, *Data Min. Knowl. Discov.* **18**, 140–181.
- [17] Kruschke, J. [2010], *Doing Bayesian data analysis: A tutorial introduction with R*, Academic Press.
- [18] Kruschke, J. [2013], ‘Optional stopping in data collection: p values, bayes factors, credible intervals, precision’.
- [19] Lai, T. L. [1988], ‘Nearly optimal sequential tests of composite hypotheses’, *The Annals of Statistics* pp. 856–886.
- [20] Lu, J. and Deng, A. [2016], ‘Demystifying the bias from selective inference: a revisit to dawid’s treatment selection problem’.
- [21] Rouder, J. N. [2014], ‘Optional stopping: no problem for Bayesians.’, *Psychon. Bull. Rev.* **21**(March), 301–8.
- [22] Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. and Iverson, G. [2009], ‘Bayesian t tests for accepting and rejecting the null hypothesis.’, *Psychon. Bull. Rev.* **16**(2), 225–37.
- [23] Sanborn, A. N. and Hills, T. T. [2014], ‘The frequentist implications of optional stopping on bayesian hypothesis tests’, *Psychonomic bulletin & review* **21**(2), 283–300.
- [24] Siegmund, D. [2013], *Sequential analysis: tests and confidence intervals*, Springer Science & Business Media.
- [25] Wald, A. [1945], ‘Sequential tests of statistical hypotheses’, *The Annals of Mathematical Statistics* **16**(2), 117–186.

APPENDIX

A. TWO SAMPLE TEST AS ONE SAMPLE TEST

We review two sample t-test and its large sample z-test version. We also show how to “normalize” the scale and turn the two sample test into an equivalent one sample test. This transformation also introduce the notions of **effect size** and **effective sample size** which is important in Bayesian test.

Suppose observations for treatment and control groups are i.i.d. from two distributions with unknown mean τ_T and τ_C respectively. Denote our observations by $Y_i, i = 1, \dots, N_T$ and $X_i, i = 1, \dots, N_C$. We test the null hypothesis $H_0 : \tau_T - \tau_C = 0$ against the alternative $H_1 : \tau_T \neq \tau_C$. Without assuming distributions of X and Y , we use the central limit theorem and hence use Wald test which is large sample version of the well-known t-test. The test statistic is

$$Z := \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_T^2/N_T + \sigma_C^2/N_C}} = \frac{\Delta}{\sqrt{\sigma_T^2/N_T + \sigma_C^2/N_C}},$$

where σ_C and σ_T are variances of X and Y . The variances are also unknown but in large sample scenario we assume they are known and use their estimates. Note that metrics are often in different scales. We first define $N_E = 1/(1/N_T + 1/N_C)$ to be the **effective sample size**. And then let σ^2 be

the **pooled variance** such that $\sigma^2/N_E = \sigma_T^2/N_T + \sigma_C^2/N_C$. With $\delta = \Delta/\sigma$, Z-statistics can be rewritten as

$$Z = \frac{\delta}{\sqrt{\sigma^2/N_E}}. \quad (11)$$

δ is Δ scaled by pooled standard deviation and is called the **effect size**. Finally, define

$$\mu := E(\delta) = E(\Delta)/\sigma = (\tau_T - \tau_C)/\sigma \quad (12)$$

is the average treatment effect scaled by σ . When σ is known, inference on $\tau_T - \tau_C$ and μ are equivalent. In Bayesian analysis it is common to define prior for μ as it is scaleless.

B. OBJECTIVE PRIOR LEARNING

Recall μ is the average effect size. Under H_0 , $\mu = 0$. Under H_1 , we assume a prior π for μ . For both cases we observe $\delta \sim N(\mu, 1/N_E)$. In addition, we assume a prior probability p for H_1 being true, and also under H_1 , $\pi \sim N(0, V^2)$ for some V . Our challenge is to learn both p and V without the need of subjectively assigning one.

Here we take advantages of historical experiment results and use them to learn the prior. Suppose for a given metric, we have N previously conducted tests with observed effect size and effective sample size $(\delta_i, N_{Ei}), i = 1, \dots, N$. We have no idea which of those are from H_0 or H_1 . Fitting the model to find MLE isn’t straightforward, due to the fact that we don’t know each δ_i belongs to H_0 or H_1 . Fortunately, a solution for this type of hidden latent variable problem, called Expectation-Maximization, is well-known[7]. EM algorithm in our case reduces to a fairly intuitive form as the following.

Step I. If p and V are known, the posterior odds for each δ_i belonging to H_1 against H_0 have the simple form

$$\frac{\phi(\delta_i; 0, 1/N_{Ei} + V^2)}{\phi(\delta_i; 0, 1/N_{Ei})} \times \frac{p}{1-p} \quad (13)$$

where $\phi(x; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 . Convert posterior odds to $P_i := P(H_1|\delta_i; p, V)$. Step II. Set p to be $\overline{P(H_1|\delta; p, V)}$ by taking average of all P_i calculated in Step I. Step III. To update V , note that under H_1 , $\text{Var}(\delta_i) = E(\delta_i^2) = V^2 + E(1/N_{Ei})$. Although we don’t know absolutely whether a δ_i belongs to H_1 , we can use posterior P_i in Step II as weights:

$$V^2 = \text{WAvg}(\delta_i^2; P_i) - \text{WAvg}(1/N_{Ei}; P_i) \quad (14)$$

where $\text{WAvg}(x_i; w_i) = \sum w_i x_i / \sum w_i$. To avoid numerical issue that V^2 in (14) can take negative value, we bound V^2 away from 0 by a lower bound.

The EM algorithm starts with an initial value of p and V , iterates through the 3 steps above until they converge. Step I is the E-step. Step II and Step III are the M-step updating p and V . (Technically Step III is generalized M-step. We update V using method of moment estimator knowing with high probability it increase the expected log-likelihood.)

The lower bound in Step III is not purely a numerical trick. It is needed for model identification. When $V = 0$, $\mu \equiv 0$ under both H_1 and H_0 . We cannot distinguish H_1 and H_0 , leading to an unidentifiable model. We recommend setting the lower bound $V^2 = k^2 \times \text{Avg}(1/N_{Ei})$ and set k to 2, see [8] for explanation.