# Bayesian Predictive Probabilities for Online Experimentation

Abbas Zaidi, Rina Friedberg, Samir Khan, Yao-Yang Leow, Maulik Soneji, Houssam Nassif, and Richard Mudd

Meta Inc
{abbaszaidi, rinafriedberg, samirk, leowyaoyang, soneji, houssamn, rmudd}@meta.com

## 1 Introduction

Online Randomized Controlled Experiments (A/B Tests) enable product decision-making across the information technology sector [5]. Among a myriad of applications, they are used to support deployment decisions and to evaluate the impact of product changes and feature launches [1]. The design may vary significantly across experiments, and various organizations (even at the same company) may follow different processes - e.g., some may perform replication experiments while others do not. Despite these variations, product decision processes ultimately depend on a company's ability to perform high-quality inference across its collection of A/B tests [13].

However, given the role of experimentation in the product life-cycle, finite resources present an evergreen challenge in this space; compute (e.g., online machine usage) and experimental units (e.g., finite number of users to advertise to) being chief amongst such constraints. Downstream of said limitation, users of experimentation platforms are increasingly motivated to use ad-hoc means of optimizing these resources primarily by way of *peeking* [9]. Their objective is to conclude futile experiments or make decisions from clearly efficacious ones cheaply and quickly given the omnipresence of limited resources. As an example, ongoing monitoring of trends in point-estimates and uncertainty intervals is a common operationalization of this practice. However, the absence of formalized stopping rules [18] are error prone and often strongly misaligned with end-of-experiment outcomes, leading to inflated type-I errors and erroneous conclusions [14, 19].

This work introduces a mechanism for principled interim-analysis using *Bayesian Predictive Probabilities* [17] that have been widely utilized in applications outside of the technology domain. These methods enable statistically efficient, data-driven decision making from experiments that are robust to peeking by way of having a stopping rule that is *ignorable* [7].

Our work contributes to the growing body of research on interim analyses in online experimentation in two ways. First, our approach introduces a novel application of predictive probabilities in online experimentation, which to our knowledge have not been utilized in this domain before. Second, we demonstrate how this system can be deployed at scale as a means of improving the efficiency of online experimentation systems without complex inference schemes. As a result, we expect the proposed approach to be a viable mechanism for practitioners to more optimally utilize these methods for large-scale online experimentation platforms.

## 2 Methods

### 2.1 Notation

Assume that at the conclusion of the experiment, the treatment under consideration is declared a success if $\mathrm{P}(\theta > 0|\hat{\theta}_{1,\ldots,T}) > (1-\alpha)$. Here $\theta$ denotes the true but unknown effect, and $\hat{\theta}_{t=1,\ldots,T}$ the

sufficient statistics of the treatment effects reported on a cadence indexed by $t$ (e.g., daily estimates of the effects). We complete this specification with the likelihood $\mathcal{L}(\hat{\theta}_{t=1,\ldots,T}|\theta)$ and the prior $\pi(\theta)$, which jointly specify the posterior $\pi(\theta|\hat{\theta}_{t=1,\ldots,T})$.

## 2.2 Inference Based on Predictive Probabilities

Our objective is to infer success before the experiment reaches its planned conclusion, i.e., at some analysis point $T' << T$. This is accomplished via the predictive probability of success (PPoS):

$$PPoS = \int_{Y_{t=T',\ldots T}} \mathbf{1}[\mathbf{P}(\theta > 0|\hat{\theta}_{t=T',\ldots,T}) > (1-\alpha)]\pi(\hat{\theta}_{t=T',\ldots,T}|\hat{\theta}_{t=1,\ldots,T'})\mathrm{d}\hat{\theta}_{t=T',\ldots,T}.$$

$\pi(\hat{\theta}_{t=T',\ldots,T}|\hat{\theta}_{t=1,\ldots,T'})$ characterizes the predictive distribution $\int_\theta \mathcal{L}(\hat{\theta}_{t=T',\ldots,T}|\theta)\ \pi(\theta|\hat{\theta}_{t=1,\ldots T'})\mathrm{d}\theta$. One may stop the trial early if $PPoS > \gamma$ for some threshold $\gamma$. The following section demonstrates how this quantity may be inferred via Monte-Carlo simulation.

## 2.3 Estimation Strategy: Posterior Predictive Simulation

We can estimate PPoS via simulation using algorithm 1, which draws from the work of Berry et al. [4]. The objective is to utilize the predictive distribution at the interim point to generate 'data' at the conclusion of the experiment, characterize the posterior at this hypothetical conclusion, and then average over the possible states.

**Algorithm 1:** Computing the PPoS via Simulation.

1 Estimate $\hat{\theta}_{t=1,\cdots,T'}$:$\pi(\hat{\theta}_{t=T',\cdots,T}|\hat{\theta}_{t=1,\cdots,T'})$
2 **for** $k \in 1,\cdots,K$ **do**
3  Simulate $\hat{\theta}^{(k)}_{t=T',\cdots,T}$ from $\pi(\hat{\theta}_{t=T'\cdots,T}|\hat{\theta}_{t=1,\cdots,T'})$
4  Estimate $\pi(\theta^{(k)}|\hat{\theta}^{(k)}_{t=1,\cdots,T'})$.
5 Estimate $\widehat{PPoS} = \frac{1}{K}\sum_{k=1}^{K}[\mathbf{P}(\theta^{(k)} > 0|\hat{\theta}_{t=1,\ldots,T'}) > (1-\alpha)]$.

## 2.4 Model Specification

As in Kessler [12], we utilize a conjugate Gaussian parameterization of the Bayesian model which assumes a well behaved estimate of the scale parameter $\sigma^2$:

$$\begin{aligned}\hat{\theta}_{t=1,\ldots,T'}|\theta,\sigma^2 &\sim \mathrm{N}(\theta,\sigma^2),\\ \theta|m_0,\tau &\sim \mathrm{N}(m_0,\tau).\end{aligned}$$

Under this specification, with no meaningful prior information at hand [2], as $\tau \to \infty$, the posterior and predictive distributions become:

$$\begin{aligned}\theta|\hat{\theta}_{t=1\ldots,T'} &\sim \mathrm{N}\left(\frac{\sum_{t=1}^{T'}\hat{\theta}_t}{T'}, \frac{\sigma^2}{T'}\right),\\ \hat{\theta}_{t=T'\ldots,T}|\hat{\theta}_{t=1\ldots,T'} &\sim \mathrm{N}\left(\frac{\sum_{t=1}^{T'}\hat{\theta}_t}{T'}, \frac{\sigma^2}{T}+\frac{\sigma^2}{T'}\right).\end{aligned}$$

### 2.5 Choice of Prior

The choice of prior in this setting depends heavily upon the required properties of the system, like the desired Type I and Type II error rates, or more general operating characteristics [20]. For this paper, in order to mimic classical Frequentist properties, we use a Jeffreys prior [7]. This is particularly useful in the online experimentation domain, where practitioners fuse Bayesian analysis of experiments with Frequentist frameworks for decision making. In section 4 we will discuss future research in this space including the development of *informative* priors.

### 2.6 Ongoing Validation via Predictive Checking

Given the complexities of validation in experimental settings where there is no *ground truth*, we take inspiration from *predictive checking* [8]. This type of assessment uses predictive simulation of new data $\hat{\theta}^{rep}_{t=1,\ldots,T}$ under a model of interest $M$, and a related statistic $f(\cdot)$, to characterize discrepancies against its observed counterpart $g(f(\hat{\theta}^{rep}_{t=1,\cdots,T}), f(\hat{\theta}_{t=1,\cdots,T}))$. This discrepancy can be used to assess any aspect of the model we want to validate (e.g., prior parameter choices or decision boundary), facilitated by the construction of a reference distribution:

$$p[g(f(\hat{\theta}^{rep}_{t=1,\cdots,T}), f(\hat{\theta}_{t=1,\cdots,T}))|M, \hat{\theta}]. \tag{1}$$

This flexible construction enables assessment of quantities like coverage for uncertainty intervals or differences between interim analyses and end of experiment outcomes – the latter of which we do in the next section to gauge performance of the PPoS based decision rule.

## 3 Results

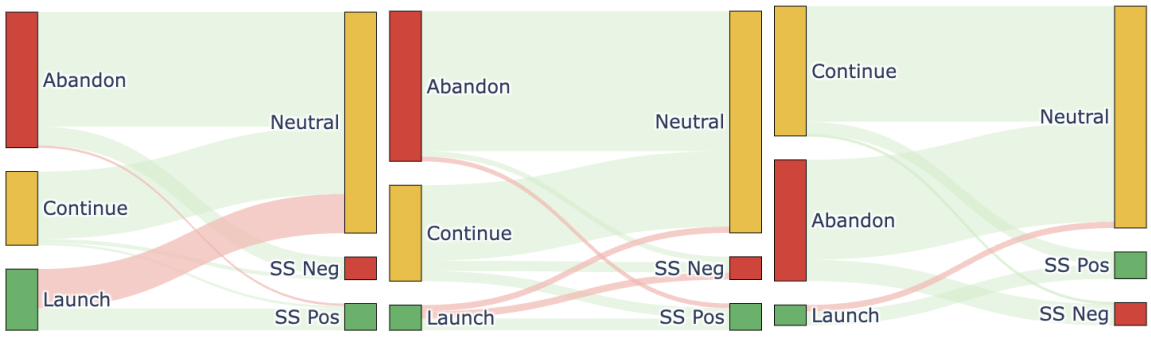
We assess the performance of our proposed technique using 345 experiments conducted over a one month period. We compare three decision-making techniques:

- **Practical Heuristic**: The experiment is abandoned as a failure if the lower endpoint of a 90% uncertainty interval for the effect is below $l$, declared a success if it exceeds $m$, and continues the experiment otherwise.
- **Always Valid**: The experiment is abandoned as a failure if the always-valid p-value is greater than 0.95, declared a success if it is less than 0.05, and continues the experiment otherwise [10].
- **PPoS**: The experiment is abandoned as a failure if the PPoS is less than 0.1, declared a success if it is greater than 0.9, and continues the experiment otherwise.

Each interim analysis is conducted at day 7 of the experiment and compared against the end of experiment decision at day 14. Figure 1 plots the results.

The practical heuristic correctly abandons 170 experiments as failures that would have been neutral or statistically significant negative if the experiment had been run to completion. However, 50 of the 78 experiments it suggests launching would have been neutral (i.e., false positives) if run to completion – a common problem stemming from decisions using unadjusted peeking [3]. Similarly, the Always-Valid rule correctly abandons 185 experiments that would have ultimately resulted in failures. It slightly improves the false positive rate relative to the naive rule: only 17 of the 32 experiments it suggests launching are neutral or statistically significant negative at completion.

PPoS based decision correctly curtails 156 experiments It improves false positives even further: only 8 of the 26 experiments being recommended for launch are neutral at completion, with no negative findings. As it correctly stops a large number of experiments while maintaining the lowest false positive rate, PPoS based stopping is preferable in our case to both the practical heuristic, and the always-valid p-value.




(a) Practical Heuristic (b) Always Valid (c) PPoS

Figure 1: Comparison of decision rules.

# 4 Discussion

We introduce Bayesian Predictive Probabilities for the interim analysis of Online Experiments as a statistically efficient and principled means of making decisions, along with a simulation based system for estimation and an ongoing validation strategy. We demonstrate, in accordance with the likelihood principle, that using a decision rule based on the predictive probability of success offers strong performance relative to both heuristic practices common in the online experimentation domain, and also to principled but inefficient Frequentist approaches. In specific, a PPoS based decision mechanism curtails a large collection of experiments while maintaining fidelity with end of experiment outcomes.

Systems based on predictive probabilities offer many avenues for further research. First, given the large slew of experiments from experienced analysts available at any given time, formal informative prior elicitation is possible [11], and can improve expediency in decision making. One can leverage offline evaluation [16] and domain knowledge [15] to construct such informative priors. Second, sign and magnitude errors [6] remain a concern in low power settings – an evergreen concern in the online experimentation domain. These are typically addressed by way of using *shrinkage* priors.

# References

[1] J. Barajas, N. Bhamidipati, and J. G. Shanahan. Online advertising incrementality testing: practical lessons and emerging challenges. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4838–4841, 2021.

[2] J. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(1):1–17, 2004.

[3] R. Berman, L. Pekelis, A. Scott, and C. Van den Bulte. *P-hacking and false discovery in A/B testing*, volume 10. SSRN, 2018.

[4] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller. *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

[5] T. Fiez, H. Nassif, Y.-C. Chen, S. Gamez, and L. Jain. Best of three worlds: Adaptive experimentation for digital marketing in practice. In *The Web Conference (WWW)*, pages 3586–3597, 2024.

[6] A. Gelman and J. Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on psychological science*, 9(6):641–651, 2014.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[8] A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[9] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[10] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

[11] S. R. Johnson, G. A. Tomlinson, G. A. Hawker, J. T. Granton, and B. M. Feldman. Methods to elicit beliefs for bayesian priors: a systematic review. *Journal of clinical epidemiology*, 63(4): 355–369, 2010.

[12] R. Kessler. Overcoming the winner's curse: Leveraging bayesian inference to improve estimates of the impact of features launched via a/b tests. 2024.

[13] R. Kohavi, D. Tang, and Y. Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

[14] A. Maharaj, R. Sinha, D. Arbour, I. Waudby-Smith, S. Z. Liu, M. Sinha, R. Addanki, A. Ramdas, M. Garg, and V. Swaminathan. Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pages 396–400, 2023.

[15] H. Nassif, Y. Wu, D. Page, and E. S. Burnside. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. In *American Medical Informatics Association Symposium (AMIA)*, pages 1330–1339, 2012.

[16] M. A. Radwan, Q. Lanners, J. Zhang, S. Karakulak, H. Nassif, and M. A. Bayir. Counterfactual evaluation of ads ranking models through domain adaptation. In *Workshops of Conference on Recommender Systems (RecSys)*, 2024.

[17] B. R. Saville, J. T. Connor, G. D. Ayers, and J. Alvarez. The utility of bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493, 2014.

[18] N. Stallard, J. Whitehead, S. Todd, and A. Whitehead. Stopping rules for phase ii studies. *British Journal of Clinical Pharmacology*, 51(6):523–529, 2001.

[19] J. Weltz, T. Fiez, A. Volfovsky, E. Laber, B. Mason, H. Nassif, and L. Jain. Experimental designs for heteroskedastic variance. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 65967–66005, 2023.

[20] T. Zhou and Y. Ji. On bayesian sequential clinical trial designs. *The New England Journal of Statistics in Data Science*, 2(1), 2024.