AEA Papers and Proceedings 2019, 109: 43–47
https://doi.org/10.1257/pandp.20191003

Empirical Bayes Estimation of Treatment Effects with Many A/B Tests: An Overview†

By Eduardo M. Azevedo, Alex Deng, José L. Montiel Olea, and E. Glen Weyl*

The use of large-scale experimentation to screen innovations is increasingly common. Large internet companies run thousands of experiments, called A/B tests, to test product improvements. In academia and policy, randomized controlled trials are widely used, from finding better anti-poverty interventions to designing nudges.

This is a practical guide on how to use treatment effect estimates from a large number of experiments to improve estimates of the effects of each experiment. This is a common practical issue in the technology industry. When thousands of new features are A/B tested, the winners tend to be a combination of good features and features that got lucky experimental draws.[1] Empirical Bayes methods (Robbins 1964) are a commonly used tool in statistics to separate good features from lucky draws (Efron 2012). We do not report any novel results. Instead we give a user-friendly overview of both classic and recent approaches to this problem.[2]

*Azevedo: Wharton, 3620 Locust Walk, Philadelphia, PA 19104 (email: eazevedo@wharton.upenn.edu); Deng: Microsoft Corporation, 555 110th Avenue NE, Bellevue, WA 98004 (email: shaojie.deng@microsoft.com); Montiel Olea: Columbia University, 1018 International Affairs Building, New York, NY 10027 (email: montiel.olea@gmail.com); Weyl: Microsoft Research, One Memorial Drive, Cambridge, MA 02142, and Yale University (email: glenweyl@microsoft.com).

†Go to https://doi.org/10.1257/pandp.20191003 to visit the article page for additional materials and author disclosure statement(s).

[1] Efron (2011) refers to this phenomenon as "selection bias" or "winner's curse."

[2] See Athey and Imbens (2017) for a survey of the econometrics of experiments and Berman et al. (2018) and Feit and Berman (2018) for other recent work on the practice of A/B testing.

I. Empirical Bayes Estimation and the Importance of Fat Tails

We consider a simplified version of the *A/B testing problem* proposed by Azevedo et al. (2019). A firm has ideas $i = 1, 2, \ldots, I$. In internet applications, ideas are features that can be implemented in a product. Idea i has true quality Δ_i. In applications, the true quality is how much the idea improves some performance metric, such as click-through rate or revenue. True quality is the population causal treatment effect of implementing an idea.

The firm does not know the true quality, but has a prior distribution G over it, and quality is independent across ideas. The distribution G captures the firm's uncertainty about the true quality, but it also models the heterogeneity in idea quality in the population.

For each idea, the firm performs an experiment, or A/B test, with n users. The experiment yields an estimated quality $\hat{\Delta}_i$. The estimated quality is an estimated treatment effect. We assume that estimated quality is normally distributed with mean Δ_i and variance σ^2/n. This is reasonable because of randomization and because of the large samples used by internet companies. The firm chooses which ideas to implement to maximize the expected sum of the quality of the implemented ideas, minus an implementation cost of c per implemented idea. Realizations of the random variables Δ_i and $\hat{\Delta}_i$ are denoted δ_i and $\hat{\delta}_i$.

Azevedo et al. (2019) show that the optimal choice of which ideas to implement is simple. The firm should use Bayes' rule to calculate $E\left[\Delta_i | \hat{\Delta}_i = x\right]$, which we denote as the posterior mean quality function $P(x)$. The optimal choice is for the firm to implement the ideas for which $P\left(\hat{\delta}_i\right)$ is greater than the implementation cost c. Therefore, the posterior mean quality function

$P(\cdot)$ is the key object for making optimal implementation decisions. We term $P(\hat{\delta}_i)$ the unfeasible Bayes estimator of quality, because it depends on knowing the prior G. In practice, the firm has to estimate the prior G from data on past experiments, a problem that we discuss in the next section.

For now, we will show that the shape of the posterior mean depends crucially on the tails of the distribution of ideas. There is evidence that the effects of innovations are fat-tailed in data from Microsoft's Bing search engine (Azevedo et al. 2019), Facebook (Coey and Cunningham forthcoming; Peysakhovich and Eckles 2018; Peysakhovich and Lada 2016), and eBay (Goldberg and Johndrow 2017). Moreover, Azevedo et al. (2019) show that outlier ideas account for a large share of the gains, even in a mature product like Bing. This suggests that it is important to take into account the potential existence of fat tails.

Figure 1 displays the posterior mean function for different priors G. The fat-tailed t-distribution prior has parameters close to the benchmark empirical estimates from Azevedo et al. (2019), who considered percentage performance improvements in the Bing search engine. The normal prior is a normal distribution with the same mean and scale parameters, but thin tails. Finally, the normal prior with matched moments was chosen to roughly match the mean and variance in the Bing data. The standard error 0.0224 roughly matches a typical Bing experiment with 20 million users.

The examples make three points. First, whether we are in the thin or fat-tailed cases makes a large difference in the shape of the posterior mean function $P(\cdot)$. With the fat-tailed Student-t prior, estimates with small t-statistics should be aggressively shrunk. The intuition is that, because typical innovations have small effects, it is very likely that these experiments were just lucky draws. This is true even for marginally statistically significant experiments. In contrast, outliers should not be shrunk very much. The reason is that, with fat tails, it is much more likely that these outliers are real effects than lucky experimental draws. In contrast, with the normal priors, the posterior mean $P(\cdot)$ is linear. Therefore, incorrectly assuming that the prior is normal can result in large biases in the estimated posterior mean quality.

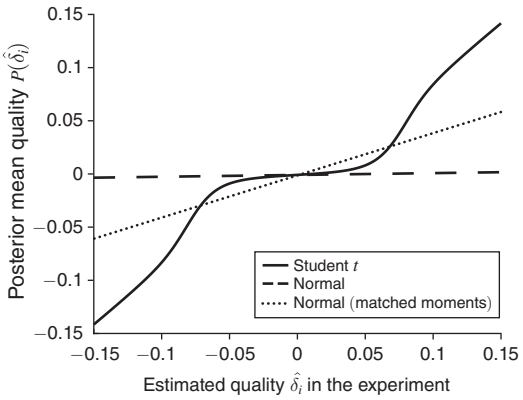

Figure 1. Posterior Mean Quality as a Function of Estimated Quality Under Different Priors

Notes: The Student-t prior has mean −0.0009, scale parameter 0.0030, and degrees of freedom parameter 1.3090. The normal prior has the same mean and scale parameters. The normal (matched moments) has mean −0.0012 and standard deviation 0.0182, which were chosen to be consistent with the mean −0.0012 and variance of experimental estimates in the data on Bing experiments analyzed in Azevedo et al. (2019). They considered percentage improvements in a key metric. The standard error 0.0224 roughly matches a typical Bing experiment with 20 million users.

Second, posterior mean quality $P(\cdot)$ can be calculated with standard Bayesian statistics.[3] Therefore, it is simple to implement this type of Bayes estimate in industry and research settings. Azevedo et al. (2019) show that optimal implementation can substantially improve on the common practice of implementing ideas with estimated quality that is statistically significantly positive at the 5 percent level. Consider the Student-t prior in Figure 1. The optimal implementation strategy can be read off the figure: implement ideas for which the posterior mean P is above c. With zero implementation cost, it is optimal to implement all ideas with positive effect with t-statistic of 0.47, which corresponds to a p-value of 32 percent. With an implementation cost of equivalent to a 0.01 percent gain in quality, it is optimal to implement all ideas with a t-statistic of at least 2.39, which corresponds

[3] Namely, Bayes' rule implies that

$$P(x) = \frac{\int y \cdot g(y) \cdot \varphi\left(\frac{y-x}{\sigma/\sqrt{n}}\right) dy}{\int g(y) \cdot \varphi\left(\frac{y-x}{\sigma/\sqrt{n}}\right) dy}.$$

to a p-value of 0.85 percent. That is, because non-outliers are likely to be lucky experimental draws, even small implementation costs make it optimal to use a small p-value cutoff for implementation. Moreover, because the outliers are not shrunk very much, the threshold t-statistic goes up slowly if implementations costs are even larger. For example, with an implementation cost equivalent to a 0.05 percent gain in quality, it is optimal to implement all ideas with a t-statistic of at least 3.61. That is, multiplying the implementation cost by 5 only increases the optimal threshold t-statistic by 50 percent.

Third, it is important for firms to evaluate whether they are in the thin-tailed or fat-tailed case. Understanding the tails gives useful intuition for how to interpret estimation results and it can also be useful for the design of A/B tests. For example, Azevedo et al. (2019) show that both the optimal implementation strategy and the optimal experimentation strategy in the A/B testing problem depend on the tails of the prior. From a practical perspective, this suggests that one of the first questions firms should ask is whether they are in the fat or thin-tailed case. This seems like a reasonable first step, before more sophisticated data analysis and before building a system to calculate $P(\cdot)$.

II. Estimation Approaches

So far, we considered optimal decisions taking the prior distribution G of quality as given. In practice, the prior has to be estimated from data on quality estimates in previous experiments, $\left(\hat{\delta}_1, \ldots, \hat{\delta}_I\right)$, and on standard errors $\left(\sigma_1/\sqrt{n_1}, \ldots, \sigma_1/\sqrt{n_I}\right)$, which we now allow to vary across observations. There are a number of approaches to this problem that can be used in industry and research settings. We now review some of the techniques and recent applications.

A. *Parametric Empirical Bayes*

Maximum Likelihood Estimation.—Suppose first that g is known up to a finite-dimensional parameter β.[4] Under the independence assumption it is possible to write a parametric likelihood for the data $\left(\hat{\delta}_1, \ldots, \hat{\delta}_I\right)$ which can be maximized with respect to β. This is the estimation strategy used by Azevedo et al. (2019).[5]

The empirical Bayes estimator of the unobserved quality of idea i incorporates the information available on the estimated effects of other A/B tests by computing the posterior mean quality based on the estimated prior distribution $g\left(\hat{\beta}_{MLE}\right)$. The estimator $\hat{\beta}_{MLE}$ is based on independent, non-identically distributed data as in Hoadley (1971). Thus, under some regularity conditions, $\hat{\beta}_{MLE}$ is consistent and asymptotically normal as the number of A/B tests grows large. This means that the posterior mean quality function based on $g\left(\hat{\beta}_{MLE}\right)$ will eventually be close to $P(\cdot)$.

Deng (2015) uses a similar approach. He assumes that the prior is a mixture of a normal distribution and a point mass at zero, and estimates it using an expectation-maximization algorithm.

Bayesian Estimation.—Since there is a parametric likelihood for the data, it is also possible to estimate β by standard Bayesian methods. For example, Goldberg and Johndrow (2017) estimate the parameters of a Student-t prior from A/B tests performed at eBay during 2016. They use the Markov chain Monte Carlo (MCMC) implementation of the hierarchical model given by the distribution of $\hat{\Delta}_i|\Delta_i$ and Δ_i using Stan. The posterior mean function can be approximated as in Azevedo et al. (2019), but replacing the maximum likelihood estimator by the posterior mean of β. One advantage of their estimation procedure is that it can be carried out using off-the-shelf software in MATLAB or the language R.

Lindsey's Method.—Consider the case where the standard error $\sigma/\sqrt{n}$ is constant across experiments. A result known as Tweedie's formula (Robbins 1956; Efron 2011, p. 1602) implies that the posterior mean quality equals

$$(1) \qquad P\left(\hat{\delta}_i\right) = \hat{\delta}_i + \frac{\sigma^2}{n}\left(\frac{d}{d\hat{\delta}_i}\log m\left(\hat{\delta}_i\right)\right),$$

[4] For example, g can be the pdf of the random variable $M + st_\alpha$, where t_α is a t-distribution with $\alpha > 1$ degrees of freedom, and $\beta = (M, s, \alpha)$.

[5] MATLAB programs to implement their estimation strategy can be found at https://github.com/eduardomazevedo/ab-empirical-bayes.

where $m(\hat{\delta}_i)$ denotes the marginal density of $\hat{\delta}_i$.

The usefulness of Tweedie's formula is that the posterior mean can be computed without knowing the prior density g, as only the *marginal* density m enters equation (1). A common approach to estimate m using quality estimates of many A/B tests is *Lindsey's method*; see section 5.2 of Efron (2012). Broadly speaking, the idea consists of modeling $m(\cdot)$ as the exponential of a polynomial with coefficient vector β. Instead of choosing β directly as the Maximum Likelihood estimator, Lindsey's method approximates the maximum likelihood estimator by using a Poisson regression. Lindsey's method can also be viewed as a nonparametric smoothing estimator; see Efron and Tibshirani (1996). The Poisson regression can be implemented using the program locfdr2 in R.

B. *Nonparametric Empirical Bayes*

General Maximum Likelihood Empirical Bayes Estimation.—Jiang and Zhang (2009) have suggested a nonparametric maximum likelihood approach to use the information of many quality estimates with Gaussian experimental noise. Their approach consists of writing the marginal density in terms of an arbitrary distribution G, and then choosing G (an infinite dimensional parameter) to maximize the marginal likelihood. This suggestion has its roots in the seminal work of Kiefer and Wolfowitz (1956). The General Maximum Likelihood Empirical Estimator is simply the posterior mean of the innovation quality, based on the estimator of G.

Experiment Splitting.—More recently, Coey and Cunningham (forthcoming) suggest a methodology for performing nonparametric empirical Bayes estimation without explicitly estimating the prior distribution of innovation quality. Their idea is to split each A/B test (randomly) into two subexperiments (each with their own treatment and control group). Their approach consists of predicting the estimated treatment effect in the first subexperiment, on the estimated treatment effect of the second subexperiment. The prediction algorithm can be a simple linear regression (possibly incorporating some other variables predictive of treatment), but could also be any other flexible method that allows for nonlinearities and selection of covariates (such as the Lasso).

REFERENCES

Athey, S., and G. W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*, Vol. 1, edited by Abhijit Vinayak Banerjee and Esther Duflo, 73–140. Amsterdam: Elsevier.

Azevedo, Eduardo, Alex Deng, José Luis Montiel Olea, Justin Rao, and Glen E. Weyl. 2019. "A/B Testing with Fat Tails." https://eduardomazevedo.github.io/papers/azevedo-et-al-ab.pdf (accessed February 25, 2019).

Berman, Ron, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte. 2018. "p-Hacking and False Discovery in A/B Testing." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3204791 (accessed February 25, 2019).

Coey, Dominic, and Tom Cunningham. Forthcoming. "Improving Treatment Effect Estimators through Experiment Splitting." *The Web Conference.*

Deng, Alex. 2015. "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments." In *Proceedings of the 24th International Conference on World Wide Web*, 923–28. New York: ACM.

Efron, Bradley. 2011. "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association* 106 (496): 1602–14.

Efron, Bradley. 2012. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

Efron, Bradley, and Robert Tibshirani. 1996. "Using Specially Designed Exponential Families for Density Estimation." *Annals of Statistics* 24 (6): 2431–61.

Feit, Elea McDonnell, and Ron Berman. 2018. "Profit-Maximizing A/B Tests." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3274875 (accessed February 25, 2019).

Goldberg, David, and James E. Johndrow. 2017. "A Decision Theoretic Approach to A/B Testing." https://arxiv.org/abs/1710.03410 (accessed February 25, 2019).

Hoadley, Bruce. 1971. "Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case." *Annals of Mathematical Statistics* 42 (6): 1977–91.

Jiang, Wenhua, and Cun-Hui Zhang. 2009. "General Maximum Likelihood Empirical Bayes Estimation of Normal Means." *Annals of*

Statistics 37 (4): 1647–84.

Kiefer, J., and J. Wolfowitz. 1956. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *Annals of Mathematical Statistics* 27 (4): 887–906.

Peysakhovich, Alexander, and Akos Lada. 2016. "Combining Observational and Experimental Data to Find Heterogeneous Treatment Effects." https://arxiv.org/abs/1611.02385 (accessed February 25, 2019).

Peysakhovich, Alexander, and Dean Eckles. 2018. "Learning Causal Effects from Many Randomized Experiments Using Regularized Instrumental Variables." In *Proceedings of the 2018 Web Conference*, 699–707. New York: ACM.

Robbins, Herbert. 1956. "An Empirical Bayes Approach to Statistics." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, edited by Jerzy Neyman, 157–63. Berkeley: University of California Press.

Robbins, Herbert. 1964. "The Empirical Bayes Approach to Statistical Decision Problems." *Annals of Mathematical Statistics* 35 (1): 1–20.