Variance-Weighted Estimators to Improve Sensitivity in Online Experiments

KEVIN LIOU, Facebook
SEAN J. TAYLOR, Lyft

As companies increasingly rely on experiments to make product decisions, precisely measuring changes in key metrics is important. Various methods to increase sensitivity in experiments have been proposed, including methods that use pre-experiment data, machine learning, and more advanced experimental designs. However, prior work has not explored modeling heterogeneity in the variance of individual experimental users. We propose a more sensitive treatment effect estimator that relies on estimating the individual variances of experimental users using pre-experiment data. We show that that weighted estimators using individual-level variance estimates can reduce the variance of treatment effect estimates, and prove that the coefficient of variation of the sample population variance is a sufficient statistic for determining the scale of possible variance reduction. We provide empirical results from case studies at Facebook demonstrating the effectiveness of this approach, where the average experiment achieved a 17% reduction in variance with minimal impact on bias.

ACM Reference Format:
Kevin Liou and Sean J. Taylor. 2020. Variance-Weighted Estimators to Improve Sensitivity in Online Experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20), July 13–17, 2020, Virtual Event, Hungary.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3391403.3399542

1 BACKGROUND AND MOTIVATION

Companies routinely turn to A/B testing when evaluating effectiveness of their product changes. Also known as a randomized field experimentation, A/B testing has been used extensively over the past decade to measure the causal impact of product changes or variants of services, and has proven to be important success factor for businesses making important decisions [10, 14].

With increased adoption of A/B testing, proper analysis of experimental data is crucial to decision quality [2, 6, 21]. Successful A/B tests must exhibit *sensitivity* – they must be capable of detecting effects that product changes routine generate. From a hypothesis-testing perspective, experimenters aim to have high statistical power – the likelihood that the experiment will detect an non-zero effect when such an effect exists. Two common ways to improve the power of an experiment are increasing sample size and decreasing the sampling variance (often through choosing less variable metrics as outcomes of interest).

Increasing sample size is straightforward to implement – one can simply include more users or conduct longer experiments. However, if the product change is detrimental, a large number of users could be adversely affected. In addition, companies conducting multiple tests may have a limited set of users available for each study [4, 23]. As a result, increasing the sample size is not always feasible or worth the trade-off.

EC '20, July 13–17, 2020, Virtual Event, Hungary

ACM ISBN 978-1-4503-7975-5/20/07…$15.00
https://doi.org/10.1145/3391403.3399542

We focus on improving sensitivity by reducing the sampling variance of the experiment metric. Several methods for variance reduction have been proposed in literature, including post-stratification [13], CUPED [5] and predictive modeling [7, 19]. We will review these methods in section 2.

While standard approaches to variance reduction often require incorporating non-experimental data, there is little prior work on methods that attempt to understand the inherent uncertainty introduced by individual experimental users. To leverage this information, we propose directly estimating the pre-experiment individual variance for each user. For example, if our target metric is "time spent by user on the site per day," we may want to give more weight to users who exhibit lower variance for this metric through their more consistent usage of the product. We can estimate the variance of a user's daily time spent during the month before the experiment and assign weights which are higher for users with less noisy behaviors. We describe this idea more in-depth in section 3.1, where we prove that the amount of variance reduction one can achieve when weighting by variance is a function of the *coefficient of variation of the variance* of experimental users – roughly how variable the users are in their variability.

Applying our approach of using inverse variance-weighted estimators to a corpus of real A/B tests at Facebook, we find that there is opportunity for substantial variance reduction with minimal impact on the bias of treatment effect estimates. Specifically, our results show an average variance reduction of 17% while bias is bounded within 2%. In addition, we show that inverse variance-weighted estimators can achieve improved variance reduction when combined with other standard approaches, such as regression adjustment (also known as CUPED) [5], demonstrating that this method complements existing work.

To summarize, our main contributions in this paper include:

- We propose an efficient average treatment effect (ATE) estimator that weights individual users based on pre-experiment variance estimates, as well as an approach–stratification–to manage bias.
- We derive the theoretical extent of variance reduction one can achieve, a function of the coefficient of variation of the user-level variances, and show how variance reduction depends on the predictive power of the variance model.
- We provide empirical results from a set of A/B tests at Facebook documenting variance reduction from our estimator with only modest increases in bias, and a comparison of how results of existing methodologies (such as CUPED) complement each other.

The remainder of this paper is organized as follows. In section 2, we review A/B testing and existing approaches to variance reduction. Section 3 proposes our variance reduction framework and section 4 introduces inverse variance-weighted estimators. In section 5, we conduct a case study evaluating our proposed estimators. Section 6 concludes with recommendations for future work and improvements.

2 REVIEW OF A/B TESTING AND RELATED WORK

2.1 A/B Testing

The primary goal of A/B testing is to estimate whether a change within a sample population leads to a meaningful effect in an observed metric. To measure such an effect, we refer to a two-sample t-test [20]. Denoting the observed metric of users in the test and control groups as Y_t and Y_c, respectively, we can then calculate a t-statistic:

$$\frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\text{Var}(\bar{Y}_t - \bar{Y}_c)}}$$

We denote $\delta = \bar{Y}_t - \bar{Y}_c$ as the Average Treatment Effect (ATE) of the experiment. The null-hypothesis of the test is that $\bar{Y}_t$ and $\bar{Y}_c$ are equal and have no significant difference. The larger the statistic, the less likely we are to have observed the difference in means when there is no effect. If we expand the expression in the denominator and assume independence between the two groups, then

$$\mathrm{Var}(\bar{Y}_t - \bar{Y}_c) = \mathrm{Var}(\bar{Y}_t) + \mathrm{Var}(\bar{Y}_c)$$

meaning that if variance is reduced, then we are more likely to reject the null hypothesis (given that the average treatment effect, $Y_t - Y_c$, is constant.)

2.2 Variance Reduction Methods

Assuming that the effect size of the experiment is constant, there are several standard variance reduction techniques used which have inspired our approach in this paper.

2.2.1 Increasing Sample Size. As mentioned previously, sample size is one of the most direct ways to decrease the variance in an experiment. To see this, note that the variance of a sample mean is

$$\mathrm{Var}(\bar{Y}) = \mathrm{Var}\left(\frac{\sum_i^N Y_i}{N}\right) = \frac{1}{N^2}\sum_i^N \mathrm{Var}(Y_i) = \frac{\sigma_i^2}{N}$$

so an increase in the sample size N leads to a decrease in the variance of the sample mean.

2.2.2 Metric Transformations. Many metrics are heavy-tailed or noisy, and thus simply calculating the raw difference in means may not be the optimal statistic to detect lift. Learning suitable metric transformations is a common method in variance reduction [9].

A simple example of a metric transformation is *winsorization*, which trims extreme values of the data to specified percentiles:

$$f(x) = \begin{cases} x, x < x_0 \\ x_0, x >= x_0 \end{cases}$$

where x_0 is a specified percentile of the data.

Another example of a metric transformation used often for non-parametric data is the *median*, which is referred to as an *Asymptotically Linear Estimator* since it is asymptotically equivalent to taking the mean after transforming each data point independently. Notably, metric transformations can add bias, as we will note later in section 5, so practitioners should manage the bias-variance trade-off with caution.

2.2.3 Regression adjustment (CUPED). Regression adjustment [15], popularly known to practitioners as CUPED (Controlled Experiment Using Pre-Experiment Data) [5], uses pre-experiment data as control variables to explain nuisance variation in outcomes and increase precision of effect estimates. To briefly review, suppose we have a target metric Y, and a closely correlated metric X which is not affected by the treatment. We can define a new variable

$$\tilde{Y} = Y - \theta X \tag{1}$$

where θ is some constant. The expectation of $\tilde{Y}$ is $\mathrm{E}(\tilde{Y}) = \mathrm{E}(Y) - \theta\mathrm{E}(X)$, and the variance of this expression is

$$\mathrm{Var}(\tilde{Y}) = \mathrm{Var}(Y) - 2\theta\mathrm{Cov}(X, Y) + \theta^2\mathrm{Var}(X) \tag{2}$$

Minimizing (2) with respect to θ, we get

$$\theta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Plugging this back into (2), we have

$$\text{Var}(\tilde{Y}) = \text{Var}(Y)(1 - \rho^2) \tag{3}$$

where ρ is the Pearson correlation between X and Y:

$$\rho = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\text{Var}(Y)} \tag{4}$$

If one can identify a variable that is explains variation in the target metric, then sampling variance of the metric can be reduced. The CUPED implementation of this idea is quite simple: the authors advocate choosing X to be the pre-experiment values of the metric. CUPED is implemented in many companies' A/B testing platforms (including Microsoft, Facebook, and Lyft), and often provides substantial variance reduction.

2.2.4 Post-Stratification. Post-stratification [13, 16] is a technique commonly used when a simple random sample of the population is imbalanced or when there may be heterogeneous treatment effects. The sample is first divided into separate strata. The average treatment effect is calculated for each strata, and the estimates are pooled with each strata getting a pre-defined weight. Specifically,

$$\hat{\delta} = \frac{\sum_k w_k \delta_k}{\sum_k w_k} \tag{5}$$

where δ_k is the average treatment effect of strata k.

There are several methods for selecting post-stratification weights. One commonly used weight is the sample proportion of the strata, $w_k = \frac{n_k}{n}$, where n_k is the number of users in strata k, and n is the total number of users in the experiment.

3 VARIANCE-WEIGHTED ESTIMATORS

We now formulate the theory behind variance-weighted estimators and present several empirical case studies showing the effectiveness of using these estimators.

The variance-weighted estimator we propose uses the same general expression from post-stratification. First, note that Equation 5 is unbiased when the treatment effects δ_k are independent, since given true effect δ,

$$\text{E}(\hat{\delta}) = \text{E}\left(\frac{\sum_k w_k \delta_k}{\sum_k w_k}\right) = \frac{1}{\sum_k w_k} \sum_k w_k \text{E}(\delta_k) = \delta \tag{6}$$

Our approach focuses on using pre-experiment estimated variance to create weights for the estimator. Importantly, the effectiveness of using a weighted estimator to reduce variance in an experiment is dependent on how well we can estimate variance.

To begin, first note that the outcome for each user can be written as

$$Y_i = \alpha + \delta_i Z_i + \epsilon_i$$

where α is the overall mean of the outcome, δ is the treatment effect, $Z_i \in \{0, 1\}$ is a binary treatment indicator, and ϵ_i is i.i.d. random error. The average treatment effect is given by

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \text{E}(Y_i | Z_i = 1) - \text{E}(Y_i | Z_i = 0) \tag{7}$$

Assume that the error term ϵ_i is drawn from a normal distribution: $\epsilon_i \sim N(0, \sigma_i^2)$. Using this formulation, one can show that

$$\text{Var}(Y_i|Z_i = 0) = \text{Var}(Y_i|Z_i = 1) = \text{Var}(\epsilon_i) = \sigma_i^2$$

We typically assume homogeneity in the variance of the error terms: $\sigma_i^2 = \sigma^2$, but practically this assumption may not be true. Since the variance of each user typically has to be estimated, the degree to which one can estimate these error terms will then determine how much variance can be reduced.

3.1 Case 1: Perfect Knowledge of Variance

Suppose that one were able to measure the variance of each individual user or strata with no error. Then, the most efficient weights to use would be the inverse variances of each measurement. The proof is straightforward: note that we have a weighted sum $\hat{\delta} = \sum_i w_i \delta_i$, where the weights w_i are normalized, $\sum_i w_i = 1$. Assuming that the δ_i are independent, the variance of $\hat{\delta}$ is given by

$$\text{Var}(\hat{\delta}) = \sum_i w_i^2 \text{Var}(\delta_i) = \sum_i w_i^2 \sigma_i^2 \tag{8}$$

To minimize (8), we use a Lagrange Multiplier:

$$\text{Var}(\hat{\delta}) = \sum_i w_i^2 \sigma_i^2 - w_0 (\sum_i w_i - 1)$$

Setting the gradient to 0,

$$0 = \frac{d\text{Var}(\hat{\delta})}{dw_k} = 2w_k \sigma_k^2 - w_0$$

so that $w_k = \frac{w_0}{2\sigma_k^2}$. Thus,

$$\frac{2}{w_0} = \sum_i \frac{1}{\sigma_i^2} := \frac{1}{\sigma_0^2}$$

where σ_0^2 is the sum of the variances σ_i^2, and

$$\text{Var}(\hat{\delta}) = \sum_i \frac{\sigma_0^4}{\sigma_i^4} \sigma_i^2 = \sigma_0^4 \sum_i \frac{1}{\sigma_i^2} = \sigma_0^4 \frac{1}{\sigma_0^2} = \sigma_0^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}} \tag{9}$$

which is the inverse of variance. This proves that weighting by the inverse of variance, $w_i = \frac{1}{\sigma_i^2}$, will result in the optimal estimator if the variances, σ_i^2, are known [22].

Let the unweighted and weighted estimators be $\hat{\delta}_u$ and $\hat{\delta}_w$, respectively. To formalize how much variance is reduced when using a weighted estimator, we define the **variance ratio (VR)** as the expectation of the ratio of the variance of the unweighted estimator to the variance of the weighted estimator, or

$$VR = \text{E}\left(\frac{\text{Var}(\hat{\delta}_u)}{\text{Var}(\hat{\delta}_w)}\right) \tag{10}$$

The following results can then be derived.

Proposition 3.1. *The variance ratio is a function of the square of the* ***coefficient of variation*** *of the user-level variances. Or,*

$$\text{E}\left(\frac{\text{Var}(\hat{\delta}_u)}{\text{Var}(\hat{\delta}_w)}\right) \cong \frac{E(\text{Var}(\hat{\delta}_u))}{E(\text{Var}(\hat{\delta}_w))}\left(1 + \frac{\text{Var}(\text{Var}(\hat{\delta}_w))}{(\text{E}(\text{Var}(\hat{\delta}_w)))^2}\right) \tag{11}$$

Proof. For simplicity in notation let $U = \text{Var}(\hat{\delta}_u)$, $W = \text{Var}(\hat{\delta}_w)$, and $f(U,W) = \frac{U}{W}$. Let $\theta = (\bar{U}, \bar{W})$ and Taylor expand around θ:

$$\begin{aligned} f(U,W) \cong & f(\theta) + \frac{df(\theta)}{dU}(U-\bar{U}) + \frac{df(\theta)}{dW}(W-\bar{W}) + \frac{1}{2}\left(\frac{d^2f(\theta)}{dU^2}(U-\bar{U})^2\right. \\ & \left. + 2\frac{d^2f(\theta)}{dUdW}(U-\bar{U})(W-\bar{W}) + \frac{d^2f(\theta)}{dW^2}(W-\bar{W})^2\right) \end{aligned} \tag{12}$$

Next, since we are interested in $\text{E}(f(U,W))$, we observe that

$$\text{E}(f(U,W)) \cong f(\theta) + \frac{1}{2}(\frac{d^2f(\theta)}{dU^2}(\text{Var}(U) + 2\frac{d^2f(\theta)}{dUdW}\text{Cov}(U,W) + \frac{d^2f(\theta)}{dW^2}\text{Var}(W)) \tag{13}$$

Since $f(\theta) = \frac{\bar{U}}{\bar{W}}$, $\frac{d^2f(\theta)}{dU^2} = 0$, $\frac{d^2f(\theta)}{dW^2} = \frac{2U}{W^3}$, and $\frac{d^2f(\theta)}{dUdW} = -\frac{1}{W^2}$, the equation above simplifies to

$$\begin{aligned} \text{E}((f(U,W)) = \text{E}\left(\frac{U}{W}\right) &\cong \frac{\bar{U}}{\bar{W}} + \frac{1}{2}\left(\frac{2U}{W^3}\text{Var}(W) - \frac{2}{W^2}\text{Cov}(U,W)\right) \\ &= \frac{\text{E}(U)}{\text{E}(W)} + \frac{\text{E}(U)}{(\text{E}(W))^3}\text{Var}(W) - \frac{1}{(\text{E}(W))^2}\text{Cov}(U,W) \\ &= \frac{\text{E}(U)}{\text{E}(W)} + \frac{1}{(\text{E}(W))^2}\left(\frac{\text{E}(U)}{\text{E}(W)}\text{Var}(W) - \text{Cov}(U,W)\right) \end{aligned}$$

Assuming $\text{Cov}(U,W) = 0$, then

$$\text{E}\left(\frac{U}{W}\right) \cong \frac{\text{E}(U)}{\text{E}(W)}\left(1 + \frac{\text{Var}(W)}{(\text{E}(W))^2}\right) \tag{14}$$

Plugging back $U = \text{Var}(\hat{\delta}_u)$ and $W = \text{Var}(\hat{\delta}_w)$ concludes the proof. □

The implication in Proposition 3.1 is that the total possible variance reduction is a function of the square of the **coefficient of variation (CV)** of the distribution of the user-level variances. Moreover, if one were to estimate the true variances of each user as accurately as possible, then the coefficient of variation would be a sufficient statistic to determines the total variance reduction using a variance-weighted estimator. This insight is significant as it quantifies the amount of variance reduction one can achieve simply by inspecting the "variance of the variances" of the experiment sample.

3.2 Case 2: Estimated Value of Variance

In practice, we must estimate the variance of each user, and thus Equation 9 is a lower bound of the weighted variance due to estimation error. There are three types of weights that can be used in any weighted estimator:

- *Equal weights:* $w_i = 1$. This is the standard estimator which we are seeking to improve.
- *Optimal weight:* $w_i = \frac{1}{\sigma_i^2}$. This is the optimal estimator if we know the true variances of each user.
- *Optimal weights with biased error:* $w_i = \frac{1}{\sigma_i^2 e_i}$. This is the estimator we use since we have to estimate the variances for each user.

Efficient estimation of user-level variances is an active area of research. We studied unpooled estimators (using the pre-experiment user-level sampling variance), building a machine learning models to predict out-of-sample variance from features, and using Empirical-Bayes estimators to pool information across users.

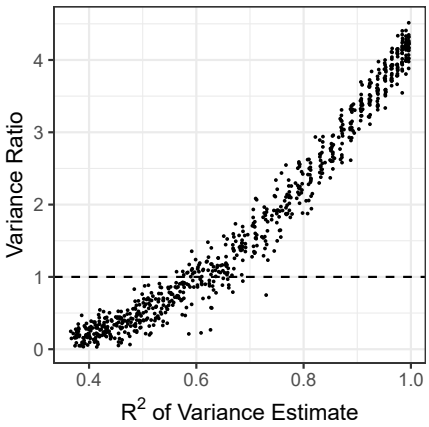

Fig. 1. **Variance Ratio vs. R^2 of model.**
Better estimates of in-experiment user-level variance can provide much larger variance reduction. Poorer models of user-level variance can actually increase variance of the estimator.

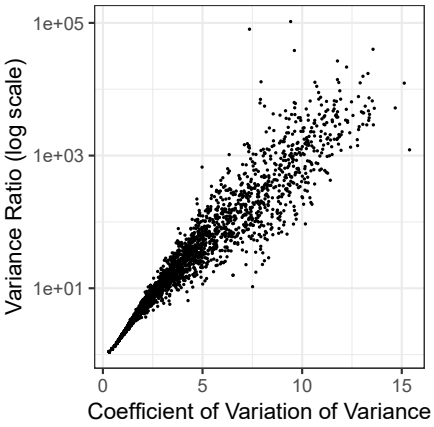

Fig. 2. **Variance Ratio vs. Coefficient of Variation of Variance.**
In agreement with our derivations, the variance reduction that can be achieved in expectation is a function of the "variance of the variances" adjusted for the mean-value of the metric.

To measure how well our estimate of variance is, we rely on two main metrics: R^2,

$$R^2 = 1 - \frac{\sum_i \sigma_i^2 - \hat{\sigma}_i^2}{\sum_i (\sigma_i^2 - \bar{\sigma^2})^2} \tag{15}$$

and mean-squared error (MSE),

$$MSE(\hat{\sigma^2}) = E[(\hat{\sigma^2} - \sigma^2)^2] \tag{16}$$

In both metrics above, we denote the theoretical variance, which we can temporarily assume to be the post-treatment variance, as σ_i^2, while the pre-experiment variance is denoted as $\hat{\sigma}^2$.

Based on these theoretical results, we first ran simulations using

$$\log(\sigma_i^2) \sim \text{Normal}(\mu, \nu^2)$$
$$(\log(\hat{\sigma}_i^2) - \log(\sigma_i^2)) \sim \text{Normal}(0, \tau^2)$$

for various values of (μ, ν, τ) to evaluate the trade-offs between variance ratio, variance prediction quality, and the coefficient of variation of the data. Figure 1 demonstrates how better variance estimates improve variance reduction when using weighted estimators.

Using R^2 as the heuristic in determining how well one can predict post-experiment variance using pre-experiment variance, we see that as prediction quality increases, the variance ratio increases. The simulation results estimate that an R^2 of about 0.60 is a necessary threshold to achieve variance reduction. User-level variance estimates that are too large increase the variance of the weights while allocating more weight to higher variance users in many cases.

The threshold from our simulation does not take into account bias that may be introduced when using weighted estimators. Section 5.1 provides a discussion of bias and how to estimate it.

Figure 2 shows in simulation how a higher coefficient of variation of the variances of individual data leads to more variance reduction opportunity, as measured once again by the variance ratio.

This is expected, as Proposition 3.1 proves that the variance reduction is a function of the coefficient of variation.

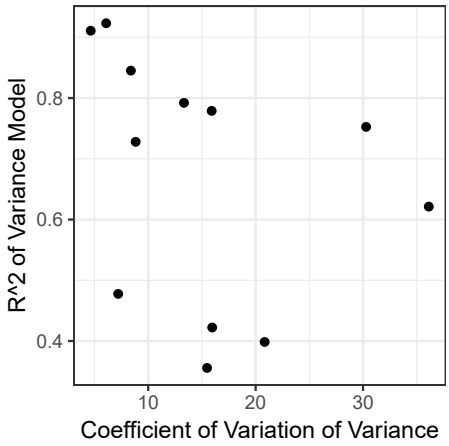

Fig. 3. R^2 **of Variance Model vs. Coefficient of Variation of Variance for 12 Facebook Metrics.** Six metrics have highly accurate user-level variance models (top-left), while two others have very high coefficient of variation (right). The four metrics in the bottom left may not be suitable for our proposed estimator.

To demonstrate that variance-weighted estimators are likely to be useful in practical settings, we collected 12 popular metrics used in A/B tests at Facebook (such as likes, comments, time spent, posts shared) to estimate the predictability of the variance for each metric and its coefficient of variation. The results, in Figure 3, show that the variance of most of the metrics are highly predictable (as measured by R^2). In addition, the coefficient of variation of the variances is large enough that they can be used effectively in a variance-weighted estimator. In section 5 we conduct further case studies showing estimates from applying a variance-weighted estimator to a A/B tests on one of the metrics here, "time spent per user per day".

4 IMPLEMENTATION OF VARIANCE-WEIGHTED ESTIMATORS

To implement variance-weighted estimators, our method utilizes the estimated variance of users prior to the experiment. In addition, we use stratification as a straightforward approach to control the variance in estimating user-level variances, as well bias from unequal treatment effects between users.

4.1 Stratification based on Estimated Variance

One could directly apply the approach we listed in the previous section by estimating the variance of each individual user and weighting inversely by these estimates. However, while overall variance of an experiment can be expected to decrease, we often run into increases in bias from treatment effect heterogeneity. We propose to manage the increase in bias is by creating quantiles bins and stratifying the users based on variance. This approach requires a weaker assumption about the treatment effect heterogeneity in order to be unbiased.

To apply stratification, we create bins based on quantiles of estimated variances of each user, creating strata. For each strata, we then calculate its average treatment effect and estimate its

weight based on within-group estimated variance (this could be any aggregate function, such as mean). From Equation 5,

$$\hat{\delta} = \frac{\sum_k w_k \delta_k}{\sum_k w_k} \tag{17}$$

Note that if the weights are normalized to 1, then the denominator in (17) is not necessary. Denoting Y_i as the metric for user i, s_i as the sample variance, S_k as each strata k, and Z_i as the treatment status, where $i = 1$ is the treatment group and $i = 0$ is the control group, we have the treatment effect in each strata as

$$\delta_k = \frac{\sum_i Y_i \mathbb{1}[s_i \in S_k \cap Z_i = 1]}{\sum_i \mathbb{1}[s_i \in S_k \cap Z_i = 1]} - \frac{\sum_i Y_i \mathbb{1}[s_i \in S_k \cap Z_i = 0]}{\sum_i \mathbb{1}[s_i \in S_k \cap Z_i = 0]} \tag{18}$$

and, if we use the mean of within-group variance as weights,

$$w_k = \frac{\sum_i s_i \mathbb{1}[s_i \in S_k]}{\sum_i \mathbb{1}[s_i \in S_k]} \tag{19}$$

Algorithm 1 runs through the steps in this approach.

Algorithm 1 Stratification based on Pre-Experiment Variance

Input: Experiment users $x_1, x_2, ..., x_n$, and their observed metrics prior to and during the experiment

Output: Weighted average treatment effect, variance

1: Obtain pre-experiment data for each user for a time period prior to the experiment, $x_{11}, x_{12}, ...x_{1t}$, $x_{21}, x_{22}, ...x_{2t}$, $x_{n1}, x_{n2}, ...x_{nt}$
2: For each user i, calculate pre-experiment variance based on the metric using some function $f(x_{i1}, x_{i2}, ..., x_{in})$. One candidate function is the variance of the data.
3: Bin the users into k strata based on the estimated pre-experiment variance.
4: Estimate the weight for each bin based on the mean (or another function) of the pre-experiment variance for users within the bucket
5: Compute the weighted average treatment effect using Equations 17-19.
6: Calculate the variance (or power) z_k of the experiment based on the results from step 5 using either bootstrapping [8] or the Delta Method [3, 18]
7: **Optional** Choose the optimal number of bins by repeating steps 3 to 6 for various values.

4.1.1 Estimating User Variances. A simple method to estimate user variance is to find the variance of the outcome metric for a user over a pre-experiment time period. The noisiness of this result is then a strong indicator of how much signal a user would provide in the experiment.

As a thought experiment, suppose we ran an A/B test on two individuals and wanted to estimate the treatment effect. The month prior to the test, we collect the target metric of each individual on each day, and both have a mean of 20 minutes. However, one individual has incredibly small variance - the metric varies daily between 19.8 and 20.2. The other individual's metric ranges from 12 to 28. If both users were to have the treatment applied to them and both increased their mean to 20.5 over the next five days in the experimental period, then obviously the individual with the lower prior variance would give us a clearer signal on the treatment effect. Conversely, the high-variance individual would provide us with a much noisier signal. Figure 4 illustrates this idea, which is based on a real A/B test at Facebook in which both users had the same mean for the target metric in the month prior to the experiment and also had the same treatment effect on the first week of the experiment. The effect of the user with the lower variance is obviously more apparent.

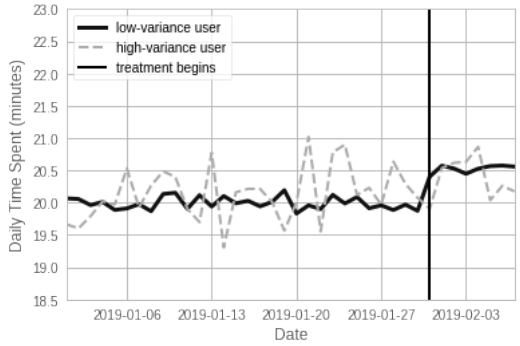

Fig. 4. **Time series of low-variance vs. high-variance users, pre- vs. post-experiment.**
The intuition behind our proposed estimator is that users with more stable behaviors provide greater sensitivity in detectin effects.

4.1.2 Bias in Variance-Weighted Estimators. One caveat here is that weighting may introduce bias – the group of users that have lower variance may exhibit a different effect size than the population. As a result, a recommendation is to compare the unweighted and weighted estimators when implementing this approach [17]. At the very least, even if the sample *is* slightly biased, achieving higher sensitivity on a typically low-powered experiment may still prove to be useful knowledge.

Another approach is to design weights that minimize variance subject to an unbiasedness constraint. For instance, given pre-treatment mean μ_i and variance σ_i^2, one could consider an unbiasedness constraint such that $E(w_i m_i) = E(m_i)$, with a minimum variance objective: $\min_{w_i} Var(w_i m_i)$. This is an area of future research.

4.1.3 Connections to CUPED. Similar to CUPED, which was mentioned in section 2.2.3, the approached described in section 3 leverages pre-experiment data for variance reduction purposes. However, how the pre-experiment data is used is different. In CUPED, one leverages the individual values or mean of a user's data over a period of time prior to the experiment. In contrast, our approach utilizes a user's pre-experiment data to understand whether a user is noisy. Thus, it is the variance of the set of data obtained from a user's pre-experimental data that is of interest, not the mean or individual values. However, an estimator that utilizes both pre-experiment mean and variance would be an intriguing addition, and in section 5 we explore the additive gains empirically.

4.1.4 Relationship to Meta-Analysis. Meta-analysis is a procedure used to combine the results of multiple experiments [1, 11, 12], and benefits include increasing statistical power and improving estimates of effect size. It is commonly used to synthesize results from multiple studies. The main difference between our proposed estimator and meta-analysis is that instead of observing multiple experiments measuring the same treatment effect and conducting a meta-analysis over many experiments, here we estimate the pre-experiment variance of users *within* one test and weight appropriately.

5 EMPIRICAL RESULTS

Time spent is an important metric of engagement at Facebook. It accounts for the amount of time users actively engage with Facebook apps and features and is measured on all major products.

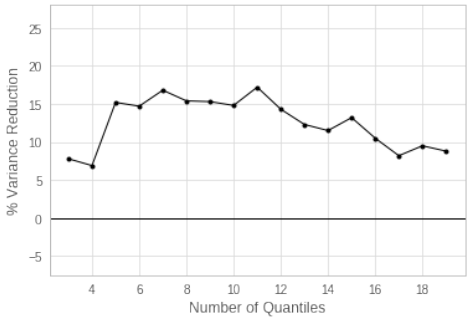

Fig. 5. **Average % Variance Reduction vs. # of Quantiles used in Stratification.**
In our collection of 100 experiments, we observe evidence for a bias-variance trade-off in choosing the number of bins.

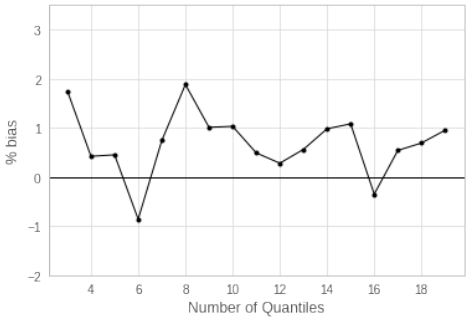

Fig. 6. **Average % Bias of vs. # of Quantiles used in Stratification.**
In our collection of 100 experiments, our estimator introduces a small amount of bias when compared to the unbiased estimator.

Many studies at Facebook measure whether a product change will lead to an increase in time spent. However, the distribution of time spent for users varies widely, and this may impact the signal clarity of their treatment effect in an A/B test depending on the degree of variance.

This presents an opportunity to apply variance-weighted estimators. To do this, we took a sample of 100 Facebook A/B tests that experimented for an increase in time spent, with the average sample size of each test at around 500,000 users. Before analyzing the results of each test, we assembled the daily time spent for each user in the month prior to the experiment and estimated the variance for each user. To see how well the estimated variance of each user was, we compared how well the pre-experiment variance correlated with the post-experiment variance. The results showed an R^2 of 0.696 and a Pearson correlation 0.754, indicating that the pre-exposed variances, when calculated over an extended period of time, do show reasonable estimations of post-exposed variance.

Next, for each experiment all users were ranked based on their estimated variance and applied stratification, as in section 4.1. To do this, we first divided users into quantiles based on pre-experiment estimated variance, and then calculated the sample variance of the experiment based on various numbers of quantiles. Figure 5 shows how an optimal selection of the number of quantiles (based on the coefficient of variation) will result in a decrease in the variance of the experiment, and at 11 quantiles there is a 17% decrease in variance. Note that at some point as the number of quantiles continues to increase, the variance reduction *decreases* as error is introduced with a large number of bins.

5.1 Bias-Variance Trade-off

Importantly, while there are noticeable reductions in variance, an associated increase in bias is likely, so we recommend one to always check for the corresponding bias to ensure that the approach is valid (in fact, note that it is often the case that if a user's variance is correlated with treatment effect, then the estimator will likely incur a noticeable bias.) To do this, we compared the % difference between the average treatment effect for the weighted estimator and the standard unweighted estimator. Denoting the unweighted and weighted estimators as δ_u and δ_w, respectively, % bias is calculated as

$$\% \; bias = 100 * \frac{\delta_u - \delta_w}{\delta_u} \tag{20}$$

A first glance at Figure 6 shows that the bias is minuscule - within 2% in all quantiles. However, that does not mean that the bias should be ignored, and one should always validate using a variance-weighted estimator by monitoring the bias accordingly. In practice, a couple useful diagnostics are to check the signal-to-noise (SNR) ratio and mean-squared error (MSE), and choose a number of quantiles that provides the maximum benefit.

5.2 Comparison to Existing Approaches

We compared the variance reduction achieved with variance-weighted estimators to CUPED [5]. Three separate approaches were considered: 1) CUPED only, 2) Variance-Weighted Estimators only, and 3) CUPED in combination with Variance-Weighted Estimators. The results are shown in Table 1.

	Variance Reduction (%)
CUPED only	37.24%
VWE only	17.31%
CUPED + VWE	48.38%

Table 1. Comparing the % variance reduction achieved for CUPED vs. VWE

In general, CUPED achieved nearly 40% in variance reduction in our case study. It is also important to recall that the performance of using variance-weighted estimators is dependent on how noisy the users in an experiment are, so our results are specific to the sample of A/B tests we studied. In reality, an experiment with a higher coefficient of variance of user variance or an experiment metric with more accurate variance estimation would improve the variance reduction opportunity. Interestingly, we found the total variance reduction achieved when using both CUPED and VWEs was less than the sum of their individual variance reduction, indicating some of the reduction obtained with VWE can be achieved through CUPED, and vice versa. How these two method work together is interesting question for future research. In particular, there is an opportunity to learn how the pre-experiment mean and variance of a metric can be jointly estimated to best reduce metric variability. Moreover, given the results from Figure 4 that showed the potential of VWEs for different topline metrics at Facebook, a logical next step would be to extend this approach to analyze A/B tests for other metrics.

6 CONCLUSION AND FUTURE WORK

Improving sensitivity of online experiments is increasingly important for businesses as they seek to detect smaller effect sizes to make launch decisions for products. We presented an approach that, like regression adjustment, uses pre-experiment information, but uses the pre-experiment variance rather than the pre-experiment mean in order to achieve a more efficient estimator. The major limitation to our proposed approach is it requires a stronger assumption about the homogeneity of the treatment effect in order to be unbiased. However this assumption is testable and we have not observed large bias in practice.

It is easy to understand when this method is likely to be helpful: we proved the efficiency gains from using variance-weighted estimators are based on two easy-to-interpret quantities: 1) the coefficient of variation of the estimated variance, and 2) the predictive performance of variance models. After demonstrating the applicability of our proposed estimator on 12 topline Facebook

metrics, we showed that the variance of a set of Facebook A/B tests can be reduced by an average of 17%, with a low impact on bias.

There are several opportunities to explore in future work:

- *Better variance estimates.* The approach we proposed in section 3.1 to estimating user-level metric variance can be improved. There may be significant gains to devising conditional variance model that estimate variance more accurately. Figure 1 showed in simulations how increased estimate qualities can improve variance reduction, suggesting very large gains possible for more precise estimation.
- *Interaction with other variance reduction approaches.* We showed the gains of VWEs in combination with CUPED in section 5. We would like to understand how VWEs may improve the variance reduction observed from other approaches (such as machine learning based methods), as analytically understand the interactions when using multiple variance reduction approaches at once.

REFERENCES

[1] Sarah E Brockwell and Ian R Gordon. 2001. A comparison of statistical methods for meta-analysis. *Statistics in medicine* 20, 6 (2001), 825–840.

[2] Thomas Crook, Brian Frasca, Ron Kohavi, and Roger Longbotham. 2009. Seven pitfalls to avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1105–1114.

[3] Alex Deng, Ulf Knoblich, and Jiannan Lu. 2018. Applying the Delta method in metric analytics: A practical guide with novel ideas. *arXiv preprint arXiv:1803.06336* (2018).

[4] Alex Deng, Jiannan Lu, and Jonthan Litz. 2017. Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 641–649.

[5] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 123–132.

[6] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1427–1436.

[7] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 256–266.

[8] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[9] William Fithian and Daniel Ting. 2017. Family learning: nonparametric statistical inference with parametric efficiency. *arXiv preprint arXiv:1711.10028* (2017).

[10] Somit Gupta, Ronny Kohavi, Diane Tang, and Ya Xu. 2019. Top Challenges from the first Practical Online Controlled Experiments Summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[11] Larry V Hedges. 1982. Estimation of effect size from a series of independent experiments. *Psychological bulletin* 92, 2 (1982), 490.

[12] Larry V Hedges and Therese D Pigott. 2001. The power of statistical tests in meta-analysis. *Psychological methods* 6, 3 (2001), 203.

[13] D Holt and TM Fred Smith. 1979. Post stratification. *Journal of the Royal Statistical Society. Series A (General)* (1979), 33–46.

[14] Ronny Kohavi, Thomas Crook, and Roger Longbotham. 2009. Online experimentation at Microsoft. *Data Mining Case Studies* 11 (2009).

[15] Winston Lin et al. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7, 1 (2013), 295–318.

[16] Luke W Miratrix, Jasjeet S Sekhon, and Bin Yu. 2013. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 2 (2013), 369–396.

[17] Kari Lock Morgan, Donald B Rubin, et al. 2012. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40, 2 (2012), 1263–1282.

[18] Gary W Oehlert. 1992. A note on the delta method. *The American Statistician* 46, 1 (1992), 27–29.

[19] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 235–244.

[20] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

[21] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 17–26.

[22] Wikipedia. [n.d.]. Inverse-Variance Weighting. https://en.wikipedia.org/wiki/Inverse-variance_weighting

[23] Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 645–654.