Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation

Alex Deng
Microsoft
One Microsoft Way
Redmond, WA 98052
alexdeng@microsoft.com

Tianxi Li
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
tianxili@umich.edu

Yu Guo
Microsoft
One Microsoft Way
Redmond, WA 98052
yguo@microsoft.com

ABSTRACT

Online controlled experiments, also called A/B testing, have been established as the mantra for data-driven decision making in many web-facing companies. A/B Testing support decision making by directly comparing two variants at a time. It can be used for comparison between (1) two candidate treatments and (2) a candidate treatment and an established control. In practice, one typically runs an experiment with multiple treatments together with a control to make decision for both purposes simultaneously. This is known to have two issues. First, having multiple treatments increases false positives due to multiple comparison. Second, the selection process causes an upward bias in estimated effect size of the best observed treatment. To overcome these two issues, a two stage process is recommended, in which we select the best treatment from the first screening stage and then run the same experiment with only the selected best treatment and the control in the validation stage. Traditional application of this two-stage design often focus only on results from the second stage. In this paper, we propose a general methodology for combining the first screening stage data together with validation stage data for more sensitive hypothesis testing and more accurate point estimation of the treatment effect. Our method is widely applicable to existing online controlled experimentation systems.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Experiment Design

General Terms

Measurement, Experimentation, Design, Theory

Keywords

Controlled experiment, A/B testing, Search quality evaluation, Meta analysis, Bias correction, Empirical Bayes

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2568028.

1. INTRODUCTION

Controlled experiments have been used in many scientific fields as the gold standard for causal inference, even though only in the recent decade has it been introduced to online services development (Christian 2012; Kohavi et al. 2009; Manzi 2012) where it gained the name A/B testing. At Microsoft Bing, we test everything via controlled experiments, including UX design, backend ranking, site performance and monetization. Hundreds of experiments are running on a daily basis. At any time, a visitor to Bing participates in fifteen or more concurrent experiments, and can be assigned to one of billions of possible treatment combinations (Kohavi et al. 2013).

Surprisingly, the majority of the current A/B testing literature focuses on single stage A/B testing, where one experiment (with one or more treatments) is conducted and analyzed at a time. A likely cause could be that A/B testing is traditionally conducted at the end of a feature development cycle, to make final feature ship decision and measure feature impact on key metrics. As A/B testing gains more recognition as one of the most effective sources of data-driven decision making, and with scaling of our experimentation platform, A/B testing is now employed earlier in feature development cycles. Thus comes the need for multiple stages of experiment, in which the results of an earlier screening stage can inform the design of a later validation stage. For example, when there are multiple shipping candidates, we design the screening stage experiment to select the most promising one. If such a best candidate exists, we conduct a second validation run to make final ship decision and measure treatment effect. This type of two stage experiments with treatment selection and validation is commonly used in practice. The space of treatment candidates ranges from 2 to 5 or even 10 in the screening stage. When candidate number exceeds 10, we can aggressively sift candidates via offline measurement or "paired test" such as interleaving (Radlinski and Craswell 2013) to boost statistical power in the data analysis.

In this paper, we focus on statistical inference in this two-stage design with treatment selection and validation. The validation stage involves only winner from the screening stage and a control. It is analyzed in the traditional A/B testing framework and is well-understood (Section 2.1). The first screening stage, however, includes simultaneous analysis of multiple treatments. We need to adjust hypothesis testing procedure to control for inflated false positives in multiple comparison (Section 2.2). We improve on traditional adjustments such as Bonferroni and Holm's method

(Holm 1979), as they are typically too conservative. In Section 3 we propose a sharp adjustment method that is exact in the sense that it touches the claimed Type I error. Point estimation is also nontrivial as the treatment selection introduces an upward bias (Lemma 4). One might wonder why this is important since in A/B testing people are generally only interested in finding the best candidate to ship. We found in a data driven organization it is equally crucial to keep accurate records of impacts made by each individual feature. These records help us understand the return on investment, and prioritize development to benefit users/customers. In Section 4 we propose several methods to correct the bias and investigate more efficient estimators by combining data from both stages. In Section 2.3, we show an insightful theoretical result to ensure we can almost always treat the treatment effect estimates from two stages as independent, given treatment procedures which assign independently, despite overlap of experiment subjects in the two stages. With this result, we propose our complete recipe of hypothesis testing in Section 3 and several options for point estimation in Section 4.

To the knowledge of the authors, this is the first paper in the context of online controlled experiments that studies the statistical inference for two-stage experiment with treatment selection and validation. This framework can be widely usable in existing online A/B testing systems. Key contributions of our work include:

- A theoretic proof showing negligible correlation between treatment effect estimates from two stages, given the treatment assignment procedure in the two stages are independent, which is generally useful in theoretical development for all multi-staged experiments.
- A more sensitive hypothesis testing procedure that correctly adjusts for multiple comparison and utilized data from both stages.
- Several novel bias correction methods to correct the upward bias from the treatment selection, and their comparison in terms of bias and variance trade-off.
- Demonstration from empirical evidence that we get more sensitive hypothesis test and more accurate point estimates by combining data from both stages.

2. BACKGROUND AND WEAK DEPENDENCE BETWEEN ESTIMATES IN TWO STAGES

Before diving into two-stage model, we first briefly cover the analysis of one stage test. Here we follow the notation in (Deng et al. 2013).

2.1 Treatment vs. Control in One Stage A/B Test

We focus on the case of the two-sample t-test (Student 1908; Wasserman 2003). Suppose we are interested in a metric X (e.g. Clicks per user). Assuming the observed values of the metric for users in the treatment and control are independent realizations of random variables $X^{(t)}$ for treatment and $X^{(c)}$ for control respectively, we can apply the t-test to determine if the difference between the two groups is statistically significant. Under the random user effect model, for user i in control group,

$$X_i^{(c)} = \mu + \alpha_i + \epsilon_i, \tag{1}$$

where μ is the mean of $X^{(c)}$, α_i represents user random effect and ϵ_i is random noise. Random user effect α has mean 0 and variance $\mathrm{Var}(\alpha)$. Residual $\mathbb{E}[\epsilon|\alpha] = 0$. The random pair (α_i, ϵ_i) are i.i.d. for all user. However, we don't assume independence of ϵ_i and α_i, as the distribution of ϵ might depend on α, e.g. users who click more might also have larger random variation in their clicks. However ϵ_i and α_i are uncorrelated by construction since $\mathbb{E}(\epsilon\alpha) = \mathbb{E}[\mathbb{E}(\epsilon|\alpha)] = 0$.

For treatment group,

$$X_i^{(t)} = \mu + \theta_i + \alpha_i + \epsilon_i,$$

where fixed treatment effect θ can vary from user to user but θ is uncorrelated to the noise ϵ. The average treatment effect (ATE) is defined as the expectation of θ. The null hypothesis is that $\delta := \mathbb{E}(\theta) = 0$ and the alternative is that it is not 0 for a two-sided test. For one-sided test, the alternative is $\delta \leq 0$ (looking for positive change) or $\delta \geq 0$ (looking for negative change). The t-test is based on the t-statistic:

$$\frac{\overline{X}^{(t)} - \overline{X}^{(c)}}{\sqrt{\mathrm{Var}\left(\overline{X}^{(t)} - \overline{X}^{(c)}\right)}}, \tag{2}$$

where observed difference between treatment and control $\Delta = \overline{X}^{(t)} - \overline{X}^{(c)}$ is an unbiased estimator for the shift of the mean and the t-statistic is a normalized version of that estimator. In traditional t-test (Student 1908), one needs to assume equal variance and normality of $X^{(t)}$ and $X^{(c)}$. In practice, the equal variance assumption can be relaxed by using Welch's t-test (Welch 1947). For online experiments with the sample sizes for both control and treatment at least in the thousands, even the normality assumption on X is usually unnecessary. To see that, by central limit theorem, $\overline{X}^{(t)}$ and $\overline{X}^{(c)}$ are both asymptotically normal as the sample size m and n for treatment and control increases. Δ is therefore approximately normal with variance

$$\mathrm{Var}(\Delta) = \mathrm{Var}\left(\overline{X}^{(t)} - \overline{X}^{(c)}\right) = \mathrm{Var}\left(\overline{X}^{(t)}\right) + \mathrm{Var}\left(\overline{X}^{(c)}\right).$$

The t-statistics (2) is approximately standard normal so t-test in large sample case is equivalent to z-test. The central limit theorem only assumes finite variance which almost always applies in online experimentation. The speed of convergence to normal can be quantified by using Berry-Essen theorem (DasGupta 2008). We have verified that most metrics we tested in Bing are well approximated by normal distribution in experiments with thousands of samples.

2.2 Multiple Treatments in A/B Test

When there is only one treatment compared to a control, Δ is both the Maximum likelihood estimator (MLE) of the treatment effect, and an unbiased estimator. When there are multiple treatments and we observe $\Delta^{(j)}$ for the jth treatment, it can be shown that $\max(\Delta^{(j)})$ is MLE of $\max(\delta^{(j)})$, where $\delta^{(j)}$ is the true underlying treatment effect of the jth treatment. However, it is obvious that $\max(\Delta^{(j)})$ is no longer an unbiased estimator. To see this, assuming $\delta^{(j)} = 0$ for all j, the distributions of $\Delta^{(j)}$'s are symmetric around mean 0. But the distribution of $\max(\Delta^{(j)})$ will certainly be skewed to the positive side and with a positive mean. Note that $\max(\delta^{(j)}) = 0$. $\max(\Delta^{(j)})$ will have a upward bias. Furthermore, Lemma 4 in Appendix showed that

if pth treatment is selected (p is random as it is picked as the best treatment), $\Delta^{(p)} = \max(\Delta^{(j)})$ is not an unbiased estimator for $\delta^{(p)}$.[1] Section 4 introduces a novel method to correct bias so we can achieve a estimator of $\delta^{(p)}$ with smaller mean squared error (MSE).

Besides point estimation, hypothesis testing in experiments with multiple treatments also suffers from an issue called multiple comparison (Benjamini 2010). Framework of hypothesis testing only guarantees Type I error (False Positive Rate) be controlled under the significant size, which is usually set to 5%, when there is only one test. When there are multiple comparisons, and if we are looking for the "best" treatment among all the treatments, our chance of finding false positive increases. If we have 100 treatments, all have 0 true treatment effect, we still expect to see on average 5 out of 100 of them showing a p-value below 0.05 just by chance. To deal with this, various of p-value adjustment techniques have been proposed, such as Bonferroni correction, Holm's method (Holm 1979) and false discovery rate based methods suitable for even larger number of simultaneous comparisons (Benjamini and Hochberg 1995; Efron 2010). Both Bonferroni and Holm's method are applicable to the general case with unknown covariance structure between test statistics of all comparisons. In the context of online A/B testing, when we have large samples, we live in a simpler multivariate normal world. We have full knowledge of the covariance structure of this multivariate normal distribution and we should be able to exploit it to come up with a better hypothesis testing procedure. Section 3 contains more details.

2.3 Weak Dependence

When we combine results from two stages to form a more sensitive test and estimate treatment effect more accurately, one of the challenges we face is caused by possible dependence of the observed metric values from the two stages. In theory we may force independence between the two stages by running them on separate traffic, so the two stages share no users in in common. This is undesirable in any scaled A/B testing platform (Kohavi et al. 2013) because

1. It means the total traffic in both stages combined cannot exceed 100%, and we suffer from decreased statistical power in both stages.
2. It requires additional infrastructure to ensure no overlap in traffic between the two stages, which can pose technical challenges when we run multiple experiments at the same time.

In this section, we explain why in practice we can safely assume independence between the observed Δ from two stages, as long as the randomization procedure used in the two stages are independent. For online A/B testing, randomization is usually achieved via bucket assignment. Each randomization id, e.g. user id, is transformed into a number between 0 to $K-1$ through a hash function and modulo operation. Independent randomization procedures between any two experiments can be achieved either by using different hashing functions and re-shuffle all K buckets, or preferably, via localized re-randomization described in Kohavi et al. (2012, Section 3.5).

For the same experiment that has been run twice in the two stages, we model user random effect to be the same for the same user in both stages, but the random noises are independent for different stages. In particular, for a pair of measurement from the same user (X_i, Y_i),

$$X_i = \mu^{(1)} + \alpha_i + \epsilon_i,$$
$$Y_i = \mu^{(2)} + \alpha_i + \zeta_i,$$

where ϵ_i and ζ_i are noise for each run. ϵ_i and ζ_i are independent and all random variables with different index i are independent. We allow $\mu^{(1)}$ to be different from $\mu^{(2)}$ to reflect a change in mean due to some small seasonal effect. After exposure to a treatment, there is an additional treatment effect term θ_i in

$$X_i = \mu^{(1)} + \theta_i + \alpha_i + \epsilon_i,$$
$$Y_i = \mu^{(2)} + \theta_i + \alpha_i + \zeta_i,$$

where θ is uncorrelated to both noise ϵ and ζ.

We are now ready for the first result of this paper. Let N be the total number of users that is available for online A/B testing. For the screening run, we picked m users as treatment and n as control. For the second run, we picked m' users for treatment and n' for control. If the random user picking for the two runs are independent of each other, let Δ_1 and Δ_2 be the observed difference between treatment and control in the two runs, then

THEOREM 1 (ALMOST UNCORRELATED DELTAS). *Assuming independent treatment assignment, we have*

$$\mathrm{Cov}(\Delta_1, \Delta_2) = \mathrm{Var}(\theta)/N \tag{3}$$

Furthermore, if $\mathrm{Var}(\theta) \le \rho \mathrm{Var}(X)$ *and* $\mathrm{Var}(\theta) \le \rho \mathrm{Var}(Y)$, *then*

$$\mathrm{Corr}(\Delta_1, \Delta_2) \le \rho.$$

This holds whether m, n, m', n' are random variables or deterministic.

To understand why Theorem 1 holds, remember N is the total user size available for experimentation, θ is the treatment effect. Also, $\mathrm{Var}(\Delta_1) = \mathrm{Var}(X^{(t)})/m + \mathrm{Var}(X^{(c)})/n$ and similarly $\mathrm{Var}(\Delta_2) = \mathrm{Var}(Y^{(t)})/m' + \mathrm{Var}(Y^{(c)})/n'$. We are interested in the correlation defined by

$$\mathrm{Corr}(\Delta_1, \Delta_2) = \frac{\mathrm{Cov}(\Delta_1, \Delta_2)}{\sqrt{\mathrm{Var}(\Delta_1) \times \mathrm{Var}(\Delta_2)}} = \frac{\mathrm{Var}(\theta)/N}{\sqrt{\mathrm{Var}(\Delta_1) \times \mathrm{Var}(\Delta_2)}}.$$

If the percentage average treatment effect is $d\%$, then we argue that $\mathrm{Var}(\theta)/\mathrm{Var}(X)$ is roughly $(d\%)^2$. To see this, if treatment effect is a fixed multiplicative effect, i.e. $\theta/X = d\%$, we have $\mathrm{Var}(\theta) = d\%^2 \mathrm{Var}(X)$. $\mathrm{Var}(\theta)/\mathrm{Var}(X)$ is at the scale of $(d\%)^2$ for any reasonable treatment effect model. Since N is always larger than m, n, m', n', by (3), we know $\mathrm{Corr}(\Delta_1, \Delta_2) \le (d\%)^2$.

In practice, most online A/B testing has a treatment effect less than 10% (In fact a 1% change is quite rare for many key metrics. If the treatment effect is more than 10%, then by the volume of traffic online experiment can

[1] There is a sutble difference between the two scenarios. In the former case we use $\max(\Delta^{(j)})$ to estimate the best treatment effect $\max(\delta^{(j)})$ and in the latter case we only want to estimate the treatment effect of pth treatment without worrying about whether it is truly the best treatment effect.

cheaply gather, detecting such a large effect is usually easy and experimenters don't need to rely on combining results from multiple stages to increase the power of the test.) This means $\text{Corr}(\Delta_1, \Delta_2) \leq 0.01$ for almost all experiments we care about in practice, and $\text{Corr}(\Delta_1, \Delta_2) \leq 10^{-4}$ among most of the cases.

What does this result entail? Remember in our large sample setting, Δ_1 and Δ_2 are approximately normal. For normal distribution, no correlation is equivalent to independence. The result from the above discussion tells us that in almost all interesting cases, we can safely treat Δ_1 and Δ_2 as if they are independent. In two-stage A/B testing with treatment selection and validation, the screening stage involves multiple treatments. A straightforward extension of Theorem 1 shows the vector of observed Δ's $\boldsymbol{\Delta_1}$ and $\boldsymbol{\Delta_2}$ has negligible correlation, hence are almost independent. Furthermore, for hypothesis testing purpose (Section 3), under null hypothesis, we assume there is no treatment effect. Hence we know correlation between Δ from two stages would be less than $(d\%)^2$, thanks to Theorem 1, which is 0.

We make our final remark to close this section. In the model we assumed the user effect α and the treatment effect θ for two runs of the same experiment to be the same for the same user. This can also be relaxed to allow both user effect and treatment effect for the same user in two runs to be a bivariate pair with arbitrary covariance structure. Theorem 1 still holds if we change $\text{Var}(\theta)$ in (3) to covariance of the treatment effect.

3. HYPOTHESIS TESTING

In statistics, combining results from different studies is the subject of a field called meta-analysis. In this section we present a method for hypothesis testing utilizing data from both stages. Using combined data for point estimation is discussed in Section 4.

3.1 Meta-Analysis: Combine Data from Two Stages

Suppose we conduct two independent hypothesis tests and observed two p-values p_1 and p_2. A straightforward attempt of combing two "probability of falsely rejecting null hypothesis" would be to multiply the two p-values together, and claim the product $p = p_1 \times p_2$ to be the p-value of the combined test. This seemingly sound approach actually produce p-values smaller than the true Type I error. To see this, under null hypothesis, p-values p_1 and p_2 follow the uniform(0,1) distribution. Type I error of combined test is $\mathbb{P}(U_1 \times U_2 \leq p)$ where U_1 and U_2 are two independent uniform(0,1) distributed random variables. This probability can be calculated using a double integral:

$$\int_{xy \leq p, 0 \leq x, y \leq 1} dxdy = p(1 - \log p). \tag{4}$$

Since $p < 1$, Type I error $p(1 - \log p) > p$. The underestimation of Type I error could be very significant for common p-values. When $p = p_1 \times p_2 = 0.1$, the true Type I error if we reject the null hypothesis would be 3.3×0.1, meaning the true Type I error is underestimated by more than 3 times if we simply multiply the two p-values.

Equation (4) provides the correct p-value calculation theoretically. To use it in hypothesis testing for usual p-value cutoff at 0.05, the product p required to make $p(1 - \log p) \leq 0.05$ is 0.0087.

The calculation of true Type I error when multiplying more than 2 p-values quickly becomes cumbersome. Fisher (Fisher et al. 1970) noticed natural log function transforms uniform(0,1) distribution into an exponential(1) distribution and exponential(1) is half of a χ^2 with 2 degrees of freedom. In this connection, the product of k p-values under null hypothesis is sum of independent exponential(1) and

$$2\log(\prod p_i) = \sum (2\log(p_i)) \sim \chi^2_{2k}.$$

This result, known as the Fisher's method, can be used to combine tests under the assumption of independent p-values. It is also a model-free method in the sense that it only utilizes p-value without tapping into the distribution of test statistics. It is not surprising that in our cases, by using normality and known covariance structure of our observed Δ's, we should be able to get a more sensitive test. We leave this extension in Section 3.2.

However, we still have the multiple comparison issue to tackle. One standard method is Bonferroni correction. Specifically,

1. First we determine the p-value from the screening stage using a Bonferronni correction. If there is K treatment candidates in the screening run, if p_1 the smallest p-value, we just divide this p-value by K.

2. We use this value plus the p-value for the second stage and combine using Fisher's method.

This combined with Fisher's method provides a valid hypothesis testing for two-stage A/B testing with treatment selection and validation. We will just call it BF method and set it as our benchmark.

3.2 Sharp Multiple Comparison Adjustment

In this section we improve BF method in two directions. We will use a sharper multiple comparison adjustment. We also exploit known distribution properties to form a test statistic. We call our method generalized weighed average method since the test statistic is in a form of weighted average.

Although Bonferroni correction is the simplest and most widely used multiple comparison adjustment, it is often too conservative in online controlled experiments. This is because by central limit theorem, we can safely assume all metrics to be approximately normal. More specifically, let $X_1, X_2, \ldots, X_k$ be the observed metric values (e.g. clicks per user) for the k treatments and X_0 be the value for the control, we can estimate the variance of each and take these as known in our model. Moreover, the covariance between $\Delta_i, i = 1, \ldots, k$ can also be estimated. In this scenario of complete distributional information, we can use a generalized step-down procedure (Romano 2005, Section 9.1, p.352).

Generalized step-down procedure

Given observed $\Delta_1, \ldots, \Delta_k$, we first test against the null hypothesis that all treatments are no greater than 0. In the screening stage we assign equal traffic size for all treatments. We use $\max(\Delta_i)$ as the test statistics. $\Delta_1, \ldots, \Delta_k$ follows a multivariate normal distribution with known covariance matrix. We can theoretically compute the distribution of

$\max(\Delta_i)$ under the least favorable null hypothesis, which is when all treatment effects are 0 (Toothaker 1993, Appendix 3, p.374). In practice, we resort to Monte Carlo simulation. We simulate B i.i.d random samples from multivariate normal distribution with mean 0 and the estimated covariance matrix. For each trial we record the $\max(\Delta_i)$. The simulated B data points serve as an empirical null distribution of $\max(\Delta_i)$. A p-value can then be calculated using the empirical distribution. The step-down procedure rejects the null hypothesis for the treatment with the largest Δ. Then it take the remaining Δ's and continue the same test against the null hypothesis for this subset of treatments. This procedure stops when the test fails to reject the subset null hypothesis and it accepts them all. It can be proved that this procedure, like Bonferroni correction, controls the family-wise false positive rate. But it is strictly less conservative than Bonferroni correction. A two-sided test would be looking at the extreme value, i.e. $\max(abs(\Delta_i))$ with otherwise the same procedure.

For our purpose of testing two-stage experiments with treatment selection and validation, we can stop at the first step, since we only care about the selected treatment with the largest Δ at the screening stage. How do we combine this with the validation stage Δ? If there are no multiple treatments at the screening stage, we are just replicating the same experiment twice. Thanks to Theorem 1, we can treat the two Δ's as independent. Then any weighted average of the two would be an unbiased estimator for the underlying treatment effect. We define

$$\Delta_c = w\Delta^{(1)} + (1-w)\Delta^{(2)},$$

where $\Delta^{(1)}$ and $\Delta^{(2)}$ stands for the observed Δ in two stages respectively. To minimize the variance of this unbiased estimator, the optimal weight w would be proportional to $1/\mathrm{Var}(\Delta^{(s)}), s = 1, 2$. This combined test statistics is also normally distributed, and therefore can be standardized into a Z-score. We call it combined Z-score method.

Generalized weighted average test

To adopt combined Z-score test to support treatment selection in the screening stage, we modify the test statistics as:

$$\Delta_c = w\max(\Delta_i) + (1-w)\Delta_*,$$

where Δ_* is the observed Δ at the validation stage for the selected treatment form the screening stage. The optimal weight can also be estimated, by calculating $\mathrm{Var}(\max(\Delta_i))$ theoretically or from simulation. We then form empirical null distribution through simulation with this test statistic, and calculate p-values. This method is in spirit the same as the combined Z-score, just with an adjustment to the null distribution.

This generalized weighted average test is more favorable comparing to more generic methods such as BF method. It exploits the know distributional information otherwise ignored. It is sharp in the sense that it would touch the designated Type I error bound, unlike BF method. The reason is, it does not rely on loose probability inequalities such as Bonferroni inequality, which was required to control for Type I error for all forms of tests. Instead it relies on simulation to get the exact Type I error, no more, no less. The weighted average method combines the data from two stages nicely and optimally. We compare generalized weighted average test to BF method in Section 5.1. The same idea of using weighted average will also be used in Section 4 for point estimation of the treatment effect.

4. POINT ESTIMATION

Another task for A/B testing is to provide good estimation of the true treatment effects, in terms of minimal mean-squared error (MSE) that achieves a balance between bias and variance.

In the screening stage of the experiment, suppose we have k different treatments with metric values $X_1, \ldots, X_k$ respectively. Thanks to the central limit theorem, we can assume $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_1, \ldots, \mu_k)^T$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2)^2$. Moreover, there is a control group with the metric value $X_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$. The estimation of σ_i is easy to achieve with a large sample and of no interest in this paper, thus we assume the variances are known and fixed. Without loss of generality, assume $\sigma = \sigma_0 = \cdots = \sigma_k$. Then we can use $\Delta_i = X_i - X_0$ as the estimation of the effect of the ith treatment. At the end of the screening stage, we choose the treatment with the largest Δ by $maxi = \mathrm{argmax}_i \Delta_i = \mathrm{argmax}_i X_i$, then run the second-stage experiment with only the control group and the $maxi$th treatment.

In the validation stage, $X^*_{maxi} \sim \mathcal{N}(\mu_{maxi}, \sigma'^2)$ for the selected treatment is obtained, as well as the new observation for control group $X_0^* \sim \mathcal{N}(\mu_0, \sigma'^2)$. Let $\Delta^*_{maxi} = X^*_{maxi} - X_0^*$. According to Theorem 1, we can ignore the dependence between the observed Δ's from the two stages.

To estimate the true treatment effect $\delta = \mu_{maxi} - \mu_0$, Δ^*_{maxi} is the MLE and also is unbiased thus optimal for the validation stage. However, according to Lemma 4, Δ_{maxi} is actually upward biased. Thus a traditional choice for the point estimation is only by Δ^*_{maxi}. The MSE for this estimator is σ^2. It is clear that this method is less efficient as we ignore the useful information from the screening stage. Denote $\tilde{\Delta}_{maxi}$ as the estimation in the screening stage. For weighted average $\hat{\delta}_w = w \cdot \hat{\delta}_{maxi} + (1-w) \cdot \Delta^*_{maxi}$:

$$\mathrm{MSE}(\hat{\delta}_w) = w^2 \mathrm{MSE}(\hat{\delta}_{maxi}) + (1-w)^2 \sigma'^2. \tag{5}$$

The best w that minimizes the MSE can be found by solving $\frac{w}{(1-w)} = \frac{\sigma'^2}{\mathrm{MSE}(\tilde{\Delta}_{maxi})}$. In practice, even by a naive choice of $w = 1/2$, we are guaranteed to have a better MSE if $\mathrm{MSE}(\hat{\delta}_{maxi}) < 3\sigma^2$. Thus the task in this section is to find a good estimator $\hat{\delta}_{maxi}$ for the screening stage, or equivalently, a good estimator $\hat{\mu}_{maxi}$ for μ_{maxi}, as we can always let $\hat{\delta}_{maxi} = \hat{\mu}_{maxi} - X_0$. X_0 is independent of $X_i, i = 1, \ldots, k$ and does not suffer from the selection bias, and it is also optimal for μ_0.

Let $\hat{\mu}_{maxi,MLE}$ be the estimator for X_{maxi}. Define the bias of MLE as $\lambda(\mu) = \mathbb{E}_\mu(X_{maxi} - \mu_{maxi})$. $\lambda(\mu)$ is positive as shown in Lemma 4. So we seek a bias correction for $\lambda(\mu)$, and propose the following estimators:

[2]Usually the metric X is in a form of average. We use X here to simplify the notation. Metric can also be a ratio of averages, such as Clicks Per Query. For this kind of metric, vector of numerator and denominator follows a bivariate normal distribution under central limit theorem with known covariance structure. We can then use the delta method to calculate variance of the metric.

Naive-correction estimator
Based on the screening stage observation $\mathbf{X} = x$, simulate B independent $y_b \sim N(x, \sigma^2 I)$. Calculate the expected bias λ from the simulated samples. Then use $\hat{\mu}_{maxi,naive} = X_{maxi} - \hat{\lambda}$ as the estimator. Note that this is actually a "plug-in" estimation as $\hat{\lambda}(\mu) = \lambda(x)$. Denote this estimator as $\hat{\mu}_{maxi,naive}$.
Compared to $\hat{\mu}_{maxi,MLE}$, the naive-correction has smaller bias but larger variance.

$$\mathrm{Var}(\hat{\mu}_{maxi,naive}) = \mathrm{Var}(\hat{\mu}_{maxi,MLE}) + \mathrm{Var}(\lambda(\mathbf{X})) - 2\mathrm{Cov}(\hat{\mu}_{maxi,MLE}, \lambda(\mathbf{X}))$$

in which $\mathrm{Cov}(\hat{\mu}_{maxi,MLE}, \lambda(\mathbf{X}))$ can be expected to be negative. This makes $\hat{\mu}_{maxi,naive}$ inferior when the true estimation bias of $\hat{\mu}_{maxi,MLE}$ is small, as can be seen from simulation results in Section 5.2. The slight bias in the naive-correction is immaterial for the performance according to our evaluation.

Inspired by the drawback of the naive-correction, we seek some simpler factor that could account for the major factor but with a smaller variance. Apparently the gap between the largest μ_i and the second largest μ_i can be a major factor for the bias. However, μ is unknown so an alternative with similar information is needed. Assume $X_{(1)}, X_{(2)} ... X_{(k)}$ are the order statistics. Define the first order gap as $H(\mathbf{X}) = \frac{X_{(k)} - X_{(k-1)}}{\sqrt{2\sigma^2}}$, then bias $\lambda(\mu)$ is roughly monotonic with the average observed $H(X)$. Thus H can be seen as a major factor for the bias in the sense of expectation. Such relationship can be observed in Figure 1, which is produced by 2000 randomly sampled μ's as described in Section 5.2. Similarly, one can define higher order gap such as the ith order gap as $H_i(\mathbf{X}) = \frac{X_{(k)} - X_{(k-i)}}{\sqrt{2\sigma^2}}$ and try to fit $\lambda(\mu)$ beyond univariate function.

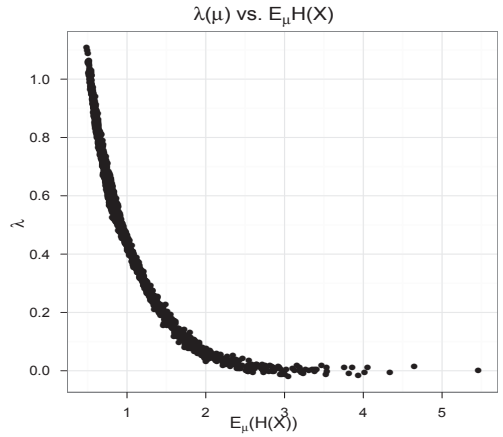

Figure 1: The $\lambda(\mu) = E_\mu(x_{maxi} - \mu_{maxi})$ and $E_\mu(H(x))$ of 2000 randomly sampled μ's.

Simple linear-correction estimator
Based on observation $\mathbf{X} = x$, simulate B independent $y_b \sim N(x, \sigma^2 I), b = 1 \cdots B$. For each y_b, denote the gap between $\max(y_b)$ and the corresponding x covariate as $d(y_b) = \max(y_b) - x_{\mathrm{argmax}(y_b)}$ and $H(y_b)$. Model the relation

$$d(y) \sim f(H(y)) \quad (6)$$

by linear models. We recommend using natural cubic splines (cubic splines with linear constraints outside boundary knots) for the model. Take the expected bias $\hat{\lambda}$ as the model prediction on $H(x)$ and use $\hat{\mu}_{maxi} = x_{maxi} - \hat{\lambda}_+$ as the estimation. Call this `NS` estimation. Simulation shows that its performance is robust against the choice of the exact linear form. For instance, using cubic polynomial regression would achieve similar performance according to our evaluation.
The intuition behind the simulation in `NS` is that suppose we can can simulate from $\mathcal{N}(\mu, \sigma^2 I)$, then we should be able to have a very good recovery of $\lambda \sim f(H(\mu))$. However since we don't know μ, we need a reasonable population is by generating random samples in a similar way as how we get $\mathbf{X}$. In `NS`, we use a "plug-in" population to simulate the data we need.

Bayesian posterior driven linear-correction estimator
Since the naive correction based on MLE suffers from high variance, we try Bayesian approach to avoid over-fitting the data for a low variance estimator. Assume a prior $\mu \sim \mathcal{N}(0, \tau^2 I)$[3]. Here we take an empirical Bayes approach to construct the posterior mean and variance as shown in (Efron and Morris 1973) — the posterior mean turns out to be the famous James-Stein estimator (Stein 1956; James and Stein 1961). Then the formula becomes

$$\mu|\mathbf{X} \sim \mathcal{N}((1 - \frac{(k-2)\sigma^2}{\|\mathbf{X}\|_2^2})X, (1 - \frac{(k-2)\sigma^2}{\|\mathbf{X}\|_2^2})\sigma^2 I).$$

When the shrinkage $(1 - \frac{(k-2)\sigma^2}{\|\mathbf{X}\|_2^2}) \leq 0$, we take it as zero and it becomes a posterior as $P(\mu = 0) = 1$. We call the posterior constructed in this way `JS` posterior. This gives another estimator: for $b = 1, \ldots, B$, we first sample μ_b from the `JS` posterior, and then sample $y_b \sim \mathcal{N}(\mu_b, \sigma^2 I)$. All the remaining modeling steps are the same as `NS`. We call this `JS-NS` estimation. Intuitively, compared to `NS`, `JS-NS` uses hierarchical sampling based on a shrinkage estimation by using an empirical Bayesian prior which is expected to make the model fitting more robust, thus giving a lower variance. Nevertheless, since the posterior shrinks the estimation toward 0, `JS-NS` estimate the bias well when true treatment effects are close to 0 but can over-correct the bias when bias is small. We see `JS-NS` as a further step beyound `NS` to lower variance by allowing for some remaining biases. See Efron (2011) for another application of Empirical Bayes based selection bias correction.

Beyond simple linear-correction
One can further explore more sophisticated (multivariate and non-linear) modeling under the logic of linear-correction. A few examples are included for completeness in Section 5.2. However, we observe no significant gains.

We can set a large number B of simulation trials to fit the functional form (6), as the computation is simple and fast. We observed that beyond $B > 10000$, more runs no longer make much difference. For the natural cubic splines,

[3]It is also possible to assume a general prior mean, which will shrink the estimation to the mean of $\mathbf{X}$ instead. Such estimation will consume one more degree of freedom and result in factor k-3 instead of k-2 in shrinkage. However, in most applications of online experiments, $\mathbf{X}$ has limited dimension thus one might prefer to save that one degree of freedom.

one can simply choose the simple version by using 3 knots. It turns out that the NS estimator with this setting achieved the best compromise between the bias and variance out of the candidates, which will be illustrated in more details in the Section 5.2. In practice, when using weighted average to combine this estimator with the second validation stage MLE, we don't know the MSE of the estimator and therefore can not get the optimal weight. However, it is reasonable to use weights inverse proportional to the sample size in formula (5), i.e. assume MSE of the two estimators in two stages are the same for the same sample size.

5. EMPIRICAL RESULTS

5.1 Empirical Results for Hypothesis Testing

In this section we use simulation to compare the BF method to the generalized weighed average method. To simulate the two-stage controlled experiments with treatment selection and validation, we assume a pair of measurement from the same user comes from a bivariate normal distribution with pairwise correlation coefficient $\rho = 0.5$ between stages, and variance 1. The choice of the variance 1 here is irrelevant because one can always scale a distribution to unit variance. The correlation coefficient here is also less important. Also note that the specific distribution of this user level measurement is not very important because of the central limit theorem. For treatment effect, we set a treatment effect on each user using a normal distribution $N(\theta, \sigma_\theta^2)$.

For each run of experiment, we randomly generated $N = 1000$ pairs of such samples from the bi-variate normal distribution, representing 1000 users entering into the two stages of the experiment. In each run we randomly assign equal proportions to be treatments and control respectively. In the screening stage, there are $k = 4$ treatment candidates and we select the best one to run the validation stage. In the context of Theorem 1, this means $m = n = N/5$ and $m' = n' = N/2$. The random sampling for the two runs are independent.

Type I error under the null hypothesis

We study Type I error achieved by BF method and generalized weighted average method under the least favorable null where all treatment effects are 0. We seek positive treatment effects as our alternative hypothesis. For two stage experiments, when validation run shows a Δ that is in different direction comparing to the screening stage run, it generally means the findings in the screening stage cannot be replicated and it then make less sense trying to combine the two runs. Therefore, when either of the two $\hat{\delta}$ were negative, we set pvalues to 1.

We ran $10,000$ simulation trials to show Type I error under null hypothesis where all treatment effects are 0. It confirms our claim that BF method is too conservative, with a 3.2% true Type I error. On the other hand, our generalized weighted average approach (WAvg), at 5.1% closely touches the 5% Type I error as promised.

We then assess the impact of the small correlation in Theorem 1 that was ignored when we performed the test. Although under the least favorable null hypothesis, treatment effects are all 0 and the correlation is exactly 0, we can still relax the null hypothesis and add some variance in the random user treatment effect θ. To do that, we assume even though $\mathbb{E}(\theta) = 0$, but $\text{Var}(\theta) = 0.04$. Since we set the variance of user level measurement to be 1, this setting means the random treatment effect has a standard deviation of 20% of that of the user level measurement. As we discussed, a 10% treatment effect is already already very rare for online A/B testing, needless to mention 20%. To test the robustness of the test against the impact of the correlation between the two stages, we ran another $10,000$ trials when treatment effect has a 20% standard deviation and observe the same phenomenon. BF method has a 3.3% true Type I error while WAvg still achieves 5.0%.We see that both generalized weighted average and BF method are very robust. This empirically justifies Theorem 1 that we can safely ignore the small correlation between the two stages. However, Theorem 1 also shows as the variance of user random treatment effect $\text{Var}(\theta)$ gets larger, the correlation could be significant. As an extreme case, we did the same simulation for $\text{Var}(\theta) = 1$, i.e., treatment random effect has the same standard deviation as the measurement itself. We found the Type I error of generalized weighted average method increased to 12.3% while for BF method it stayed at 4.2%. This suggests BF method might be more robust against the correlation than generalized weighted average method.

Statistical power under the alternative hypothesis

Next we compare the sensitivity of the two tests under alternative. We increase N to $110,000$ and let the treatment effect vary from 0 to 0.03 with step size 0.001. We also increased the number of treatments to 10 and set the same treatment effect on them. Figure 2 shows the power curve estimated from 10000 simulation runs. We observed that BF method could be inferior to validation run t-test for small treatment effects, even though it tried to take advantage of data from the screening run. By contrast, generalized weighted average method had higher power than both of them for the whole range, with a gap of 10% to 15% for the middle range. When $\mu = 0.015$, power for weighted average method was 81% and the power for BF and validation run t-test was only at around 68%. Hence we've seen generalized weighted average method is more sensitive.

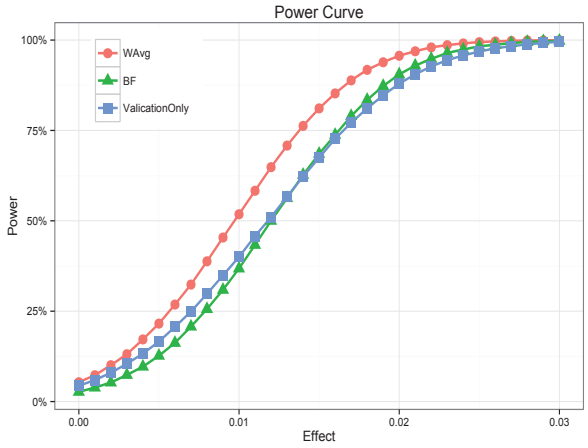

Figure 2: Power curve of generalized weigthed average method, BF method and validation run t-test.

5.2 Empirical Results for Point Estimation

To evaluate the proposed estimators of μ_{maxi}, we simulated 200 experiments with four treatments each, their means μ from $N(\eta, 4I)$, where $\eta = (2, 2, 2, 2)^T$ is irrelevant to the estimation. Without loss of generality, we use $\sigma = 1$ in simulation. Then for each experiment, we generated 1500 independent random vectors x with mean μ, and calculated their MSE Note we can only observe one x in practice. In all the estimations, the B is chosen to be 10000. The knots for splines were chosen as the (0.1,0.5, 0.9) quantiles of the simulated $H(y_b)$.

Three additional methods were included for completeness: the bi-variate natural splines estimation and boosted trees with two variables for bias correction via $H(X)$ and $H_2(X)$. The bi-variate splines model is an example of including more predictors: setting the basis functions with 6 degree of freedom for both of the two variates without any interaction term. We label this estimator `NS-2` and the corresponding univariate version as `NS-1`. The boosted trees can be seen as an example when one seeks to model the bias exhaustively by using the two predictors (but still in a regularized way). In each estimation of equation 6, all base trees were required to be of depth 3 (so at most 2 orders of interaction will be considered in each base learner), and 1200 trees were trained in each estimation, with final prediction model selected according to hold-out squared prediction error. We label this estimator as `Boost`. The last estimator is a mixture between MLE and `JS-NS` according to the rule: if $H(X) > 1$ take MLE as the estimator and otherwise use `JS-NS`. The intuition behind this will be explained in Remark 1. We label it as `MLE-JSNS-Mixture`. Many threshold values other than 1 were evaluated as well but they all suffer from the same problem which will be discussed later.

In addition to MSE, the bias and variance of each estimators were also studied respectively. For the clearness of illustration, only 16 μ's are shown in detail below, with error bars for the MSE and biases estimated by 1500 instantiations. These 16 μ are representative for the general case according to our observations. The error bars of the variance are too small to be shown, thus were ignored. The 16 μ's are ordered according to the first order gap $H(\mu)$. In each figure, they are also split into two groups: the ones with $H(\mu) > 1$ and those with $H(\mu) < 1$.

Figure 3 shows that MLE has smaller bias and larger variance when $H(\mu)$ is large and vice versa. When the true bias is small, the naive correction become inferior to MLE while in other cases, it achieves much smaller MSE. Figure 4 shows that naive-correction results in minimum bias on average, but has large variance. Thus it will be outperformed by MLE when the gain in bias is small. On the other hand, `NS-1` achieved a good compromise between the two. It performs nearly as good as naive-correction when naive-correction works well, while still retains reasonably good MSE when the naive-correction fails. Neither `NS-2` nor `Boost` seems to improve the performance. Thus using the univariate linear estimator is the best choice among the candidates. `JS-NS` performs significantly better than others in the case of $H(\mu) < 1$, and closely to `NS` for moderate $H(\mu)$. Figure 4 reveals the advantage of `JS-NS`. The variance of `JS-NS` is always similar to MLE, much lower than all other variables. This means using hierarchical sampling with JS posterior does help to make the estimation stable. As discussed before, `JS` tends to over-correct the bias when $H(\mu)$ is large though as shown in the bias plot. In summary, `JS-NS` is the best one in most cases while `NS` is a good choice if one conservatively prefers a good performance uniformly.

REMARK 1. *Figure 3 indicates a simple rule of selecting correction method: if $H(\mu)$ is large, use MLE, otherwise use correction estimation (for instance, `JS-NS`). Since μ is unknown in practice, one alternative is to use observed $H(\mathbf{X})$ instead. This is the intuition for `MLE-JSNS-Mixture` and its variants. However, it turns out that these estimators perform poorly in many cases due to the large variances as shown in Figure 4, though the biases are small. Using $H(\mathbf{X}) > 1$ introduces extra variance that degraded performance of the estimators.*

REMARK 2. *When $H(\mu)$ is large, all candidates except MLE have higher MSE than σ^2. Such extra MSE is the price we have to pay for not knowing the oracle of whether we need bias correction beforehand. In the best cases, the MSE of the final estimation is lower than σ^2, even better than the case of single random variable. This is an effect of "learning from the experience of others" discussed in Efron (2010). That is, since μ_i's are close in such cases, knowing all x_i's could help to improve the estimation.*

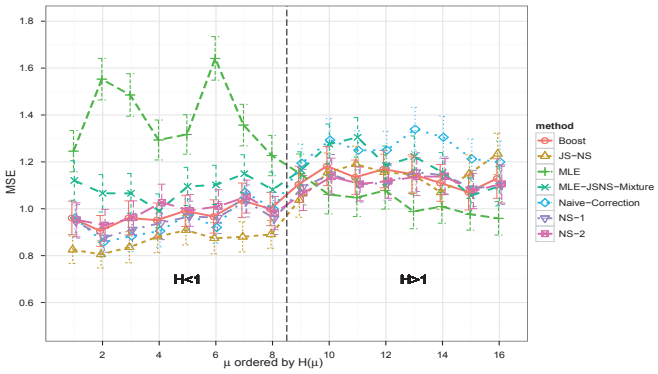

Figure 3: The estimated MSE for candidate estimators on 16 of the 200 randomly generated μ's. The μ's are ordered according the value $H(\mu)$ and are grouped by $H(\mu) < 1$ and $H(\mu) > 1$.

6. CONCLUSION

When data-driven decision making via online controlled experiment becomes a culture and the scale of an online A/B testing platform reaches a point when anyone can and should run their ideas through A/B testing, it is almost certain that A/B testing will eventually be employed through the full cycle of web-facing software developments. This is already happening now at Microsoft Bing. In this paper, we took our first steps to build the theoretical foundation for one of the multi-stage designs that is already a common practice — two-stage A/B testing with treatment selection and validation.

The results and methods we laid out in this paper are more general even though motivated primarily by this specific design. Using generalized step-down test to adjust for multiple comparison can be applied to any A/B testing with relatively few treatments. Bias correction methods are useful when one cares about a point estimate and our empirical

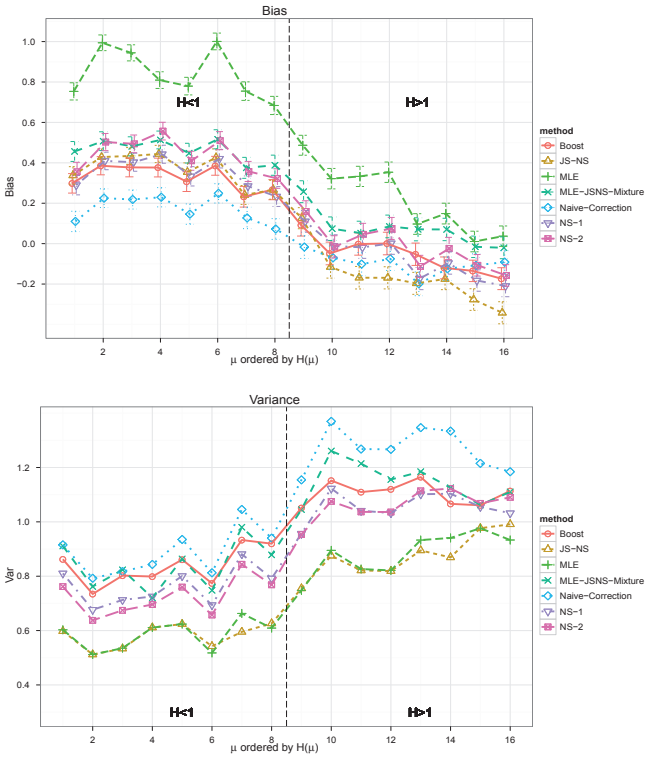

Figure 4: The estimated biases and variances for candidate estimators on the 16 μ's. The μ's are ordered according the value $H(\mu)$ and are grouped by $H(\mu) < 1$ and $H(\mu) > 1$.

results shed lights on how to find further correction methods with even smaller MSE The theoretical proof of weak correlation between estimators from multiple stages is a general result that is applicable beyond two stages.

We wish to point out one concern here. In our model we did not consider any carryover effect. In multi-stage experiments when there is overlap of subjects (traffic) at different stages, treatment effect from first exposure may linger even after the treatment is removed. To eliminate carryover effect, one can use separate traffic for different stages. In practice, however, we didn't find evidence of strong carryover effect in most of the experiments we run or if any, such effects usually fade away after a few weeks' wash-out period. However, we have observed cases where carryover effect can linger for weeks or even months, such as examples we shared in (Kohavi et al. 2012, Section 3.5). One proposed solution is to segment users in the validation stage by their treatment assignment in the screening stage. This is equivalent to including another categorical predictor in the random effect model. If there is no statistically significant benefit from introducing this additional predictor, we assume there is no carryover effect from Occam's razor principle.

7. ACKNOWLEDGMENTS

We wish to thank Brian Frasca, Ron Kohavi, Paul Raff and Toby Walker and many members of the Bing Data Mining team. We also thank the anonymous reviewers for their valuable feedback.

References

Yoav Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Brian Christian. The a/b test: Inside the technology that's changing the rules of business, April 2012. URL `http://www.wired.com/business/2012/04/ff_abtesting/`.

Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, 2008.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 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*, pages 123–132. ACM, 2013.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973. ISSN 01621459.

Sir Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

W. James and James Stein. Estimation with quadratic loss. In Jerzy Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. University of California Press, 1961.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining Knowledge Discovery*, 18:140–181, 2009.

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. *Proceedings of the 18th Conference on Knowledge Discovery and Data Mining*, 2012.

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker nd Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. *Proceedings of the 19th Conference on Knowledge Discovery and Data Mining*, 2013.

Jim Manzi. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books, 2012.

Filip Radlinski and Nick Craswell. Optimized interleaving for online retrieval evaluation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 245–254. ACM, 2013.

Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Calif.*, volume 1, pages 197–206. University of California Press, Berkeley, Calif., 1956.

Student. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

Larry E Toothaker. *Multiple comparison procedures*. Number 89. Sage, 1993.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

Bernard L Welch. The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

APPENDIX

A. DETAILED PROOFS

In this appendix, we give proof of Theorem 1 and a proof of positive selection bias in Lemma 4.

PROOF OF THEOREM 1. We prove the theorem for the case when m, n, m', n' are deterministic. The proof for random case is a simple extension for which we give a similar proof for Corollary 3.

Let I, J and S, T be the indexes of users picked as treatment and control for the two runs. From now on we use the shorthand $\overline{X}_I$ to denote the sample average over index set I.

$$\begin{aligned}\mathrm{Cov}(\Delta_1, \Delta_2) =& \mathrm{Cov}(\overline{X}_I - \overline{X}_J, \overline{Y}_S - \overline{Y}_T)\\ =& \mathrm{Cov}(\overline{X}_I, \overline{Y}_S) + \mathrm{Cov}(\overline{X}_J, \overline{Y}_T) \qquad (7)\\ & - \mathrm{Cov}(\overline{X}_I, \overline{Y}_T) - \mathrm{Cov}(\overline{X}_J, \overline{Y}_S) = \mathrm{Var}(\theta)/N.\end{aligned}$$

To prove the last equation, we first calculate the first term in the expansion. I and S are both treatments so we need to add treatment effect to both.

$$\mathrm{Cov}(\overline{X}_I, \overline{Y}_S) = \mathrm{Cov}(\frac{\sum_I(\alpha_i + \theta_i + \epsilon_i)}{m}, \frac{\sum_S(\alpha_s + \theta_s + \zeta_s)}{m'}),$$

Expand the second term and use the assumptions in the model, we get

$$\mathrm{Cov}(\overline{\alpha}_I, \overline{\alpha}_S) + \mathrm{Cov}(\overline{\alpha}_I, \overline{\theta_S}) + \mathrm{Cov}(\overline{\theta_I}, \overline{\alpha_S}) + \mathrm{Cov}(\overline{\theta_I}, \overline{\theta_S}).$$

By applying Lemma 2 to all 4 terms

$$\mathrm{Cov}(\overline{X}_I, \overline{Y}_S) = \mathrm{Var}(\alpha)/N + 2\mathrm{Cov}(\alpha, \theta)/N + \mathrm{Var}(\theta)/N.$$

Similarly, we can show

$$\mathrm{Cov}(\overline{X}_J, \overline{Y}_T) = \mathrm{Var}(\alpha)/N,$$

$$\mathrm{Cov}(\overline{X}_I, \overline{Y}_T) = \mathrm{Cov}(\overline{X}_J, \overline{Y}_S) = \mathrm{Var}(\alpha)/N + \mathrm{Cov}(\alpha, \theta)/N.$$

Therefore (7) holds because the 4 terms canceled with each other. □

LEMMA 2. *For a random variable* $\mathbf{X}$ *with known variance* σ^2 *and let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be* N *i.i.d. copies of* $\mathbf{X}$. *Let* $I = I_1, \ldots, I_m$ *and* $J = J_1, \ldots, J_n$ *be the indexes of* m *and* n *random selections out of* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *and denote sample average by* $\overline{X}_I$ *and* $\overline{X}_J$ *respectively. Then* $\mathrm{Cov}(\overline{X}_I, \overline{X}_J) = \sigma^2/N$.

Similarly for a pair of random variables $(\mathbf{X}, \mathbf{Y})$ *with covariance* σ_{XY}, *then* $\mathrm{Cov}(\overline{X}_I, \overline{Y}_J) = \sigma_{XY}/N$.

PROOF. WLOG, assume $\mathbb{E}X = 0$. (Otherwise set $X = X - \mathbb{E}X$.) First,

$$\mathrm{Cov}(\overline{X}_I, \overline{X}_J) = \frac{1}{m \times n}\sum_I\sum_J \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_i, X_j),$$

where the last equality holds because $\mathrm{Cov}(X_i, X_j)$ are the same for any $i \in I$ and $j \in J$. To calculate $\mathrm{Cov}(X_i, X_j)$, note that $\mathbb{P}(i = j) = 1/N$.

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathbb{E}[X_iX_j] = \mathbb{E}[X_iX_j|i = j]\mathbb{P}(i = j)\\ &= \mathrm{Var}(X, X)\mathbb{P}(i = j) = \sigma^2/N.\end{aligned}$$

Combine the two we've proved the Lemma. The proof of the second part is essentially the same. □

COROLLARY 3. *In Lemma 2, the result holds if* m *and* n *are two random numbers.*

PROOF. When m and n are random, WLOG assuming 0 mean for X, Lemma 2 essentially proved:

$$\mathrm{Cov}(\overline{X}_I, \overline{X}_J|m, n) = \mathbb{E}(\overline{X}_I\overline{X}_J|m, n) = \sigma^2/N.$$

Using tower property of conditional expectation,

$$\mathrm{Cov}(\overline{X}_I, \overline{X}_J) = \mathbb{E}(\overline{X}_I\overline{X}_J) = \mathbb{E}[\mathbb{E}(\overline{X}_I\overline{X}_J|m, n)] = \sigma^2/N.$$

The proof for bivariate case is similar. □

LEMMA 4 (NON-NEGATIVE SELECTION BIAS). *A sequence of independent random variables* $X_1, X_2, \ldots, X_p$ *have finite expectations of* $\theta_1, \theta_2, \ldots\theta_p$ *respectively. Let* $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ *be the (increasing) order statistics, and denote the corresponding means as* $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(p)}$ *which is one permutation of the model means. Then the selection bias defined as* $\mu = \mathbb{E}(X_{(p)} - \theta_{(p)}) \geq 0$.

PROOF. Without loss of generality, assume $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$. We have

$$\begin{aligned}\mu =& \mathbb{E}(X_{(p)} - \theta_{(p)})\\ =& \int_{x_p = x_{(p)}}(x_p - \theta_p)\mathrm{d}P(X) + \int_{x_p \neq x_{(p)}}(x_{(p)} - \theta_{(p)})\mathrm{d}P(X)\\ =& \int_{\mathbb{R}^p}(x_p - \theta_p)\mathrm{d}P(X) - \int_{x_p \neq x_{(p)}}(x_p - \theta_p)\mathrm{d}P(X)\\ &+ \int_{x_p \neq x_{(p)}}(x_{(p)} - \theta_{(p)})\mathrm{d}P(X)\\ =& \int_{x_p \neq x_{(p)}}(x_{(p)} - x_p + \theta_p - \theta_{(p)})\mathrm{d}P(X).\end{aligned}$$

The conclusion follows by noticing that $x_{(p)} - x_p \geq 0$ when $x_p \neq x_{(p)}$ and $\theta_p - \theta_{(p)} \geq 0$. Note that in the case of Gaussian variables, it will be strictly positive. □