Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments

Minyong R. Lee
Airbnb, Inc.
San Francisco, California
minyong.lee@airbnb.com

Milan Shen
Airbnb, Inc.
San Francisco, California
milan.shen@airbnb.com

ABSTRACT

Online controlled experiments, or A/B testing, has been a standard framework adopted by most online product companies to measure the effect of any new change. Companies use various statistical methods including hypothesis testing and statistical inference to quantify the business impact of the changes and make product decisions. Nowadays, experimentation platforms can run as many as hundreds or even more experiments concurrently. When a group of experiments is conducted, usually the ones with significant successful results are chosen to be launched into the product. We are interested in learning the aggregated impact of the launched features. In this paper, we investigate a statistical selection bias in this process and propose a correction method of getting an unbiased estimator. Moreover, we give an implementation example at Airbnb's ERF platform (Experiment Reporting Framework) and discuss the best practices to account for this bias.

CCS CONCEPTS

• **General and reference → Measurement**; • **Mathematics of computing → Hypothesis testing and confidence interval computation**; **Exploratory data analysis**; • **Information systems → Online analytical processing**; **Business intelligence**;

KEYWORDS

online experiments; A/B testing; bias correction; multiple hypothesis testing

ACM Reference Format:
Minyong R. Lee and Milan Shen. 2018. Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219905

1 INTRODUCTION

Online controlled experimentation, or A/B testing, has been playing an essential role in product analytics of today's data-driven industry. The idea is to deliver incremental changes on a subset of users and conduct statistical tests to understand their effectiveness. For many web companies, their large user volume provides enough sample size for randomized controlled experiments, so that the same population of users is randomly split into two groups with or without the change, and one can draw a causal relationship between the change and any effect observed. Many works including those by the researchers from Microsoft [13], LinkedIn [24], Google [20] and Twitter [4] have discussed the challenges and lessons in building the experimentation platforms.

A/B testing results not only help us to identify whether a change is beneficial but also to quantify the potential impact on business metrics if we launch the change to production. Measuring the impact is critical because we can evaluate ideas quickly and find the promising directions, shaping the product gradually.

The process of identifying and quantifying successful changes using controlled experimentation is called *attribution process*. Although different companies have different infrastructures or frameworks built for A/B testing tailored to their specific challenges, most of them have common characteristics. First, on a daily basis, there are many ongoing experiments, and new experiments are launched onto different segments of users continuously. Second, for a given product team, a test is usually targeted to improve on a set of target metrics, while a set of core metrics are monitored for all experiments company-wise. Third, among a set of tests conducted, some of them are launched to production either to a smaller group or the full traffic according to certain prescribed conditions, which are determined by taking into account both the business goal and statistical properties of the measurement in the experiment. Naturally, if one launches ten experiments out of a hundred experiments conducted in total, aggregating together the measured effect differences in those ten experiments can give us a rough estimate of the overall impact on the target business metric.

Attribution of fully launched features and understanding total aggregated effect are critical to a product's business strategy because it helps the team to quantify as accurately as possible its business impact and thus ROI in any given product direction. As online experiments gained popularity in web analytics, many studies focused on various aspects of it. Still, there has been little discussion on aggregating the results from multiple online tests. In our work, we show that when we aggregate the effects that were statistically significant, we introduce an upward bias in the estimate of the total effect. We call this bias the *winner's curse bias* in online experiments. Winner's curse ([3, 21]) is a phenomenon in common value auctions that the winner tends to overpay than the value of the item. Analogously, the observed effects of selected experiments tend to overestimate the true effects of experiments.

Although it is possible to run well-designed experiments without this kind of bias, direct correction is much more practical and

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219905

easier to be implemented and adopted. Usually, running repeated experiments has a significant cost, and running multi-factor experiment requires enormous sample size. The main trade-off lies between accurate measurement and efficient iteration under time and resource constraints. More of the necessity of correction is in Section 5.3.3.

In this paper, we quantify the winner's curse bias and propose a direct bias-corrected estimate of effects from online experiments. The paper is organized as the following. In Section 2, we give a brief review of selected papers. In Section 3, we quantify the bias and propose a direct bias-corrected estimate of effects from online experiments, followed by a discussion on the bootstrap method to calculate confidence intervals. In Section 4, we show simulation results using bias correction. In Section 5, we discuss how the method is implemented at Airbnb with recommended best practices.

2 LITERATURE REVIEWS

There are many challenges in online experiments from the nature of collected data from users. Kohavi [14] discussed practical rules when running online experiments, and also discussed the characteristics of them in subsequent papers ([13, 15]). Crook [5] summarized several risks to avoid when running online experiments. Deng and Guo [6] proposed a way to correct the bias when there are multiple treatment groups within experiments. However, there is not much discussion on measuring the total effect from multiple experiments.

Winner's curse bias has been investigated in the literature on genome-wide association studies (GWAS). Sun and Bull [19], Garner [12], Zhong and Prentice [25, 26], Zöllner and Pritchard [27], Xiao and Boehnke [22], and Xu, Craiu, and Sun [23] addressed this selection bias where the actual genetic association is typically overestimated due to the selection procedure, and proposed bias-reduced estimators. These studies consider situations where only a few substantial effects exist among many null effects and propose ways that reduce the bias for individual estimates of the selected effects. We compare our proposed bias correction with previous studies on winner's curse in Section 3.3.

More generally, the post-selection inference has been studied extensively in statistics. Bancroft [1] investigated the bias of a variance estimator obtained after performing a preliminary test of significance, and extended the discussion to more examples in [2] and introduced related studies. Efron [7] proposed an Empirical Bayes approach for correcting the selection bias. Fithian et al. [11] proposed inference controlling the selective type I error after model selection. In regression setting, Efron [9] discussed bias in estimation after model selection, and Lee et al. [16] developed an approach for exact inference after model selection. In a different perspective, we can consider the winner's curse bias as quantified false positives in the multiple hypothesis testing framework. For more discussions on multiple hypothesis testing, see Efron [8].

In this paper, we consider the situation where the effects from online experiments are aggregated to estimate the total effect of selected experiments. Our main contribution is that we quantify the bias from the selection process, not the bias in each selected effects which many papers are focusing on. We give a detailed comparison between selected works and our work in Section 3.3.

3 FORMULATION

Suppose that we run n controlled online experiments and observe estimated incremental effects X_i , $i = 1, \cdots, n$. For example, if we are interested in the average revenue per user, then we can define X_i as the difference of the average revenue per user metric in the treatment group compared to that in the control group for experiment i.

3.1 Preliminary Assumptions

We make several assumptions to simplify the formulation.

First, we assume that the effect of estimates is additive. For instance, if we observe 0.5 of increase and 0.1 of decrease from two experiments, then the total effect is 0.4. In the real world, it is common to use the percentage lift to represent the effect for business purpose. When an experiment is running to measure different metrics or for various time length, the absolute effect difference can vary a lot, and percentage lift is easy to report and comparable universally. Under this circumstance, the effects are multiplicative, but in large scale online experiments, the effects are usually minuscule, and the sum of the effects is a good approximate of the total effect. Correcting the bias for multiplicative effects can be easily extended from our work.

Undoubtedly, if the experiments are run simultaneously on different traffic buckets, or using control groups with the same feature sets, the sum of effects measured in individual experiments may not be a reasonable estimate of the total effect unless the experiments are independent without any interaction. In these cases, it is invalid to claim any aggregated effects unless we run a meta-experiment containing all the features being tested, similar to the hold-out setup as discussed in 5.3.3. In spite of that, it is also a common practice to run experiments successively, especially under the scope of one project or product theme. If an experiment was successful, we launch its treatment to all the subjects and run the following experiment. Thus the cumulative effect on the metric should be computed by adding up (or multiplying up) the individual effects. Therefore, the additivity of the effects is a reasonable assumption in practice.

Second, the launch condition for a given experiment is assumed to only depend on the metric being significantly positive. As discussed in Section 1, in reality, the conditions can be much more complicated and compose of multiple different metrics. Most of the time, the metrics at stake are quite dependent on the target metric; it is out of the scope of this work to derive a precise formula for all possible cases, and we think the simplified assumption is sufficient to inspire a better method and reduce bias meaningfully.

Now suppose that $X_1, \ldots, X_n$ are random variables defined on a same probability space, and each X_i follows a distribution μ_i with finite mean a_i and finite variance σ_i^2. (The distributions μ_i are not necessarily identical.) We regard a_i as the unknown *true effect* and usually estimate it by the unbiased estimate X_i.

We select the 'significant' experiments that had positive estimated effects larger than a threshold. Suppose we use significance level α_i for each experiment i. For now, let us assume that σ_i is known. We choose experiments such that $X_i/\sigma_i > b_i$, where b_i is the cutoff from the reference distribution for significance level α_i.

Let us define the set of significant experiments $A = \{i \mid X_i/\sigma_i > b_i\}$. Then, the total true effect of A is

$$T_A = \sum_{i \in A} a_i.$$

If we add up the effects of positive significant experiments, the total estimated effect of A is

$$S_A = \sum_{i \in A} X_i.$$

Note that since A is a random set, therefore $\mathbb{E}[S_A] \neq \mathbb{E}[T_A]$ in general. Also, the total true effect T_A is random. We define the *expected total true effect* as

$$\mathbb{E}[T_A] = \mathbb{E}\Big[\sum_{i \in A} a_i\Big].$$

3.2 Estimation of the Bias

First, we show that $\mathbb{E}[S_A] \geq \mathbb{E}[T_A]$.

$$\begin{aligned}\mathbb{E}[S_A - T_A] &= \mathbb{E}\Big[\sum_{i \in A}(X_i - a_i)\Big] = \mathbb{E}\Big[\sum_{i=1}^{n} I(i \in A)(X_i - a_i)\Big] \\ &= \sum_{i=1}^{n} \mathbb{E}[I(i \in A)(X_i - a_i)] = \sum_{i=1}^{n} \mathbb{E}\Big[I\Big(\frac{X_i}{\sigma_i} > b_i\Big)(X_i - a_i)\Big] \\ &= \sum_{i=1}^{n} \mathbb{E}[I((X_i - a_i) > (b_i\sigma_i - a_i))(X_i - a_i)].\end{aligned}$$

All the summands are all nonnegative because the mean of lower truncated mean-zero distribution is always positive. To be more specific,

1) If $b_i\sigma_i - a_i \geq 0$, then

$$\begin{aligned}&\mathbb{E}[I((X_i - a_i) > (b_i\sigma_i - a_i))(X_i - a_i)] \\ \geq &\mathbb{E}[I((X_i - a_i) > (b_i\sigma_i - a_i))(b_i\sigma_i - a_i)] \geq 0.\end{aligned}$$

2) If $b_i\sigma_i - a_i < 0$, then

$$\begin{aligned}&\mathbb{E}[I((X_i - a_i) > (b_i\sigma_i - a_i))(X_i - a_i)] \\ &= \mathbb{E}[X_i - a_i] - \mathbb{E}[I((X_i - a_i) \leq (b_i\sigma_i - a_i))(X_i - a_i)] \\ &= 0 - \mathbb{E}[I((X_i - a_i) \leq (b_i\sigma_i - a_i))(X_i - a_i)] \\ &\geq -\mathbb{E}[I((X_i - a_i) \leq (b_i\sigma_i - a_i))(b_i\sigma_i - a_i)] \geq 0.\end{aligned}$$

Thus, $\mathbb{E}[S_A] \geq \mathbb{E}[T_A]$. In other words, if we estimate $\mathbb{E}[T_A]$ by S_A, we introduce upward bias.

If we can quantify $\mathbb{E}[S_A - T_A]$, then we can define the unbiased estimate of the expected true effect as

$$\tilde{T}_A = S_A - \mathbb{E}[S_A - T_A].$$

We easily find that

$$\mathbb{E}[\tilde{T}_A] = \mathbb{E}[T_A]. \tag{1}$$

Now let us quantify the bias under certain assumptions. In online experiments, the incremental effect X_i is computed by the difference of the aggregated statistics over many subjects, between control and treatment. The most common aggregated statistics are averages and medians. Thus, in many cases, we can assume that X_i follows Gaussian distribution $N(a_i, \sigma_i^2)$ from the central limit theorem.

Let ϕ, Φ be the density function and the cumulative distribution function of standard normal distribution respectively. Then

$$\begin{aligned}&\mathbb{E}\Big[I\Big(\frac{X_i}{\sigma_i} > b_i\Big)(X_i - a_i)\Big] \\ &= \mathbb{P}\Big[\frac{X_i}{\sigma_i} > b_i\Big]\mathbb{E}\Big[X_i - a_i \Big| \frac{X_i}{\sigma_i} > b_i\Big] \\ &= \mathbb{P}\Big[\frac{X_i - a_i}{\sigma_i} > \frac{\sigma_i b_i - a_i}{\sigma_i}\Big]\mathbb{E}\Big[\sigma_i \frac{X_i - a_i}{\sigma_i} \Big| \frac{X_i - a_i}{\sigma_i} > \frac{\sigma_i b_i - a_i}{\sigma_i}\Big] \\ &= \sigma_i\Big(1 - \Phi\Big(\frac{\sigma_i b_i - a_i}{\sigma_i}\Big)\Big)\frac{\phi(\frac{\sigma_i b_i - a_i}{\sigma_i})}{1 - \Phi(\frac{\sigma_i b_i - a_i}{\sigma_i})} = \sigma_i \phi\Big(\frac{\sigma_i b_i - a_i}{\sigma_i}\Big).\end{aligned}$$

Thus,

$$\beta = \mathbb{E}[S_A - T_A] = \sum_{i=1}^{n} \sigma_i \phi\Big(\frac{\sigma_i b_i - a_i}{\sigma_i}\Big). \tag{2}$$

We find that the bias contributed from a single experiment is always positive, and the bias is a function of standard deviation and the significance of the experiment. It is vital to note that every experiment contributes to the bias, even if they were not selected. We are essentially measuring the bias coming from the selection process.

Figure 1 shows the bias function in terms of true effect and p-values. We observe that the selection bias increases proportionally to the true effect size, which is expected. Roughly, if we see a 2% increase in target metric with 0.3 p-value, if we estimate the true effect to be around 2%, then the experiment contributes 0.5% to the bias.

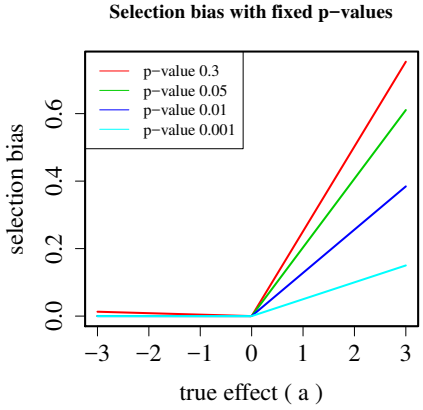


Figure 1: Plot of bias contributed from a single experiment

The bias depends on the unknown parameters a_i and σ_i. We use the estimates X_i and W_i, where W_i is the estimated standard deviation of X_i, to define the bias estimate,

$$\hat{\beta} = \sum_{i=1}^{n} W_i \phi\Big(\frac{W_i b_i - X_i}{W_i}\Big) \tag{3}$$

and the debiased plug-in estimate of $\tilde{T}_A$,

$$\hat{T}_A = S_A - \hat{\beta} = S_A - \sum_{i=1}^{n} W_i \phi\Big(\frac{W_i b_i - X_i}{W_i}\Big). \quad (4)$$

The plug-in estimate approximately satisfies unbiasedness (1).

Remark 1. *The proof of $\mathbb{E}[S_A] \geq \mathbb{E}[T_A]$ and quantification of the bias follows under the assumption of known σ_i. Although the estimated standard deviation W_i depends on the data, when W_i and X_i are independent, the derivations are still valid. One popular example is the unpaired t-test, where the sample mean and variance are independent under the Gaussian assumption. There are cases where W_i and X_i are not independent, for example when the data is binary, and X_i is defined as the conversion rate. Still, $\hat{\beta}$ is a consistent estimate of the bias β.*

3.3 Discussion on the expected total true effect

Our quantity of interest $\mathbb{E}[T_A]$ should be distinguished from the target quantities in Zhong and Prentice [25], Efron [7], and Xu, Craiu and Sun [23]. They estimate a_i given experiment i is selected and correct the bias from conditioning. Regarding the total effect, their bias correction would be targeting

$$T_A \mid A.$$

Noting that $\mathbb{E}[X_i | i \in A] > a_i$, estimating $T_A|A$ by S_A also suffers from winner's curse, and debiasing can be done similarly. For example, in the Gaussian case,

$$\mathbb{E}[X_i - a_i \mid i \in A] = \frac{\mathbb{E}[(X_i - a_i) I(i \in A)]}{\mathbb{P}[i \in A]} = \sigma_i \frac{\phi\left(\frac{\sigma_i b_i - a_i}{\sigma_i}\right)}{1 - \Phi\left(\frac{\sigma_i b_i - a_i}{\sigma_i}\right)}.$$

For GWAS studies, where we are interested in the mean of each selected gene, this type of bias is what we need to take care of. Thus, if $i \in A$, then the conditionally unbiased (plug-in) estimate of a_i is

$$\hat{a}_i = X_i - W_i \frac{\phi\left(\frac{W_i b_i - X_i}{W_i}\right)}{1 - \Phi\left(\frac{W_i b_i - X_i}{W_i}\right)}.$$

Zhong and Prentice [25] shows that this unbiased estimator conditioned by $i \in A$ is also the MLE of the conditional likelihood of a_i given $i \in A$. To differentiate with our estimate in (4), we define

$$\hat{T}_{A,\text{cond}} = \sum_{i \in A} \hat{a}_i = S_A - \sum_{i \in A} W_i \frac{\phi\left(\frac{W_i b_i - X_i}{W_i}\right)}{1 - \Phi\left(\frac{W_i b_i - X_i}{W_i}\right)}. \quad (5)$$

The estimator $\hat{T}_{A,\text{cond}}$ can be used to estimate the sum of means of selected experiments, given that the experiments are selected due to their statistical significance.

Note that the bias corrections in Garner [12], Zhong and Prentice [25], and (5) depend only on selected X_is. The derivation of the bias in these early studies share the same idea of using the truncated Gaussian distribution. The result is different from our estimator $\hat{T}_A$ because we integrate out the conditioning by A.

However, when one is interested in the aggregated effects from selected experiments, T_A should be estimated. The quantity $T_A|A$ does not fully remove the selection bias, because of the conditioning. For instance, conditioning by A can suffer heavily from false positives. In the analysis of online experiments, people usually measure the aggregated difference in a metric to measure the improvement. Integrating out the selection and measuring the bias of the process as a whole provide us with a more fair estimate of the improvement.

The following intuitive example shows the difference between estimating T_A and $T_A|A$. Let us consider the case that a data scientist ran 1000 experiments under significance level 0.001. Suppose that only the jth experiment was statistically significant with $X_j = 1$ and $\hat{\sigma_j} = 0.3$, which leads to a p-value of 0.00043. Let us also assume that the other 999 experiments had effects $X_i \approx 0$ and $\hat{\sigma}_i = 0.3$. We intuitively see that there is a high risk of false discovery in this setup. In this case,

$$\hat{T}_{A,\text{cond}} = \hat{a}_j \approx 0.77 \text{ and}$$
$$\hat{T}_A \approx 0.35.$$

Our bias correction takes account of the false discovery risk.

In Section 4, we will demonstrate by simulation that our bias correction is the correct way to get an accurate estimate of the total true effect of experiments.

Efron [7] considers bias correction in a more general selection scheme based on Empirical Bayes, using Tweedie's formula. Presumably, the analysis can also be applied to estimate $\mathbb{E}[T_A]$. The Empirical Bayes approach is based on assuming a prior on the true effects. Although the relevance between experiments can be controlled flexibly, the analysis may not be suited for online experiments when the number of experiments is small, and also when it is not reasonable to assume a smooth parametric prior on the true effects.

3.4 Bootstrap confidence intervals

We can construct a confidence interval of the expected total true effect using parametric bootstrap ([10]). For each i, we use X_i which is the maximum likelihood estimate of a_i, to resample the effect from a Gaussian distribution. The following is the simulation procedure.

For each b in $1, \ldots, B$,

(1) Sample $X_{i,b}$ from $N(X_i, W_i)$ for $i = 1, \ldots, n$.
(2) Find the set of significant experiments A_b from $\{(X_{i,b}, W_i)\}_{i=1}^n$.
(3) Compute the bottom-up estimate S_{A_b} and the true effect T_{A_b}, and the debiased estimate $\hat{T}_{A_b}$.

There are three ways to construct a confidence interval for the total true effect:

1) Naive-CI : $S_A \pm \Phi^{-1}(1 - \alpha/2) \sum_{i \in A} W_i$
2) Bootstrap-CI: confidence interval based on the empirical distribution of T_{A_b}
3) Debiased Bootstrap-CI: confidence interval based on the empirical distribution of $\hat{T}_{A_b}$

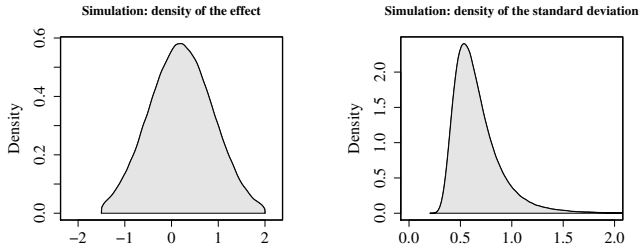
We can expect that naive-CI and bootstrap-CI will suffer from the bias. We compare the coverages empirically in the simulation section.

4 SIMLUATION FROM A HYPOTHETICAL DISTRIBUTION

We show a simulation result to show that the bias is substantial. Suppose that we have $n = 30$ experiments and measure the incremental percentages of the effects. For each i, we sample a_i from a truncated Gaussian distribution

$$a_i \overset{d}{=} Z_i | (-1.5 < Z_i < 2) \text{ where } Z_i \sim N(0.2, 0.7^2)$$

and σ_i^2 from the inverse gamma distribution with shape parameter 3 and scale parameter 1. Figure 2 shows the densities of the distributions of a_i and σ_i. Note that we chose a left-skewed distribution of the true effects, to have more positive true effects in the simulation.



(a) Distribution of a_i (b) Distribution of σ^2

Figure 2: Distributions of the true effect and the true standard deviation for simulation

We ran 1000 simulations to see the distribution of the S_A and T_A. The units are % . We assume that σ_i are known when selecting the experiments; we are essentially assuming that the estimated standard deviations are good estimates of σ_i. Table 1 is the summary table of the number of positive significant experiments ($|A|$), total estimated effect (S_A), debiased total conditional estimated effect ($\hat{T}_{A,\text{cond}}$), debiased total estimated effect ($\hat{T}_A$), total true effect (T_A), and the expectation of the bias ($\mathbb{E}[S_A - T_A]$) out of 1000 simulations. We observe that S_A and $\hat{T}_{A,\text{cond}}$ are upward biased by a substantial amount.

Table 1: Summary of 1000 simulations. $|A|$ is the number of significant positive experiments.

	$\lvert A\rvert$	S_A	$\hat{T}_{A,\text{cond}}$	$\hat{T}_A$	T_A	β
Min	0	0	0	-3.62	-0.25	1.47
1Q	3	4.31	3.58	1.62	2.57	2.42
2Q	4	6.45	5.35	3.64	3.90	2.72
Mean	4.19	6.72	5.63	3.86	4.07	2.73
3Q	5	8.88	7.58	6.00	5.40	3.06
Max	10	19.49	17.43	16.19	11.51	4.23

Figure 3 is the scatter plot of S_A against the true effect T_A, Figure 4 is the scatter plot of $\hat{T}_{A,\text{cond}}$ against T_A, and Figure 5 is the scatter plot of $\hat{T}$ against T_A. The point size represents the number of significant positive experiments in the simulation. For most cases, the naively added estimate is overestimating the true effect, and the bias is substantial. The debiased conditional estimated effect $\hat{T}_{A,\text{cond}}$ does not fully remove the bias. We observe that our proposed bias correction is working well in Figure 5.

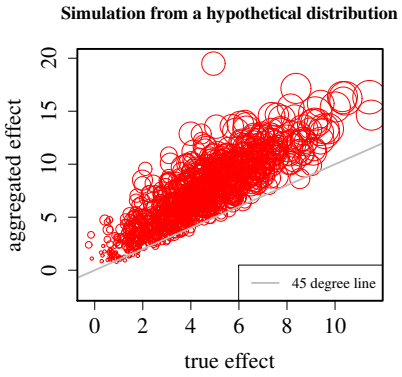


Figure 3: Scatter plot of the estimated effect S_A and the true effect T_A.

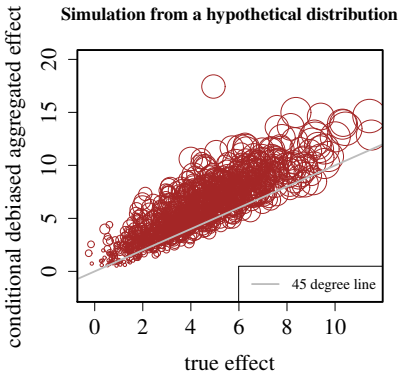


Figure 4: Scatter plot of the debiased conditional estimated effect $\hat{T}_{A,\text{cond}}$ and the true effect T_A.

We also constructed the confidence intervals of the true total effects for each simulation. We used $B = 160000$ bootstrap samples for bootstrap confidence intervals. Table 2 shows the coverage of the confidence intervals discussed in section 3.4. We observe that the coverage of the debiased bootstrap confidence intervals are closer to the target coverage than the other intervals, although the coverage is not very precise for lower target coverages. Further research on the distribution of the bias and improving bootstrapping will help us construct more precise confidence intervals of the total true effect.

5 APPLICATION AT AIRBNB

At Airbnb, controlled experiments are the fundamental building blocks of decision-making. Almost all changes to the product are

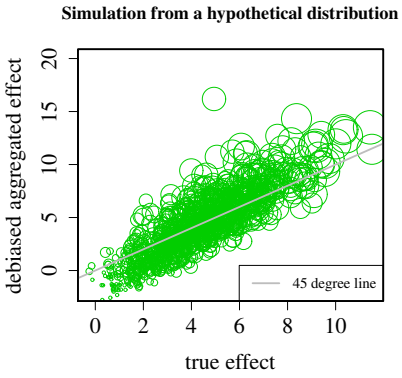


Figure 5: Scatter plot of the debiased estimated effect $\hat{T}_A$ and the true effect T_A.

Table 2: Coverage of the 95 % confidence intervals from 1000 simulations.

Target Coverage	Naive	Bootstrap	Debiased Bootstrap
0.7	0.481	0.15	0.525
0.8	0.684	0.251	0.693
0.9	0.878	0.449	0.896
0.95	0.966	0.605	0.968

tested through online experiments together with various user research studies. In this section, we give a concrete example and discuss lessons learned during the implementation stage.

5.1 Example: Experiments in Market Dynamics

In the past years, the Market Dynamics team at Airbnb has developed many different features focusing on balancing the supply and demand to optimize the efficiency and liquidity of Airbnb marketplace. The team had launched a series of experiments, to drive a target metric. These experiments were designed based on the hypothesis that the new features improve the target metric, which was related to the bookings made by guests, and have no negative impact on other business metrics.

To illustrate, we use real experiment results with anonymized experiment names. There were 24 such experiments in total, in the series of experiments. Most experiments were launched one after one, although a few overlapped with others in experimenting duration. When an experiment ended, the team determined whether to launch it or which treatment to launch according to the predetermined condition: the target metric has been lifted significantly (under level $\alpha = 0.05$) while other sanity metrics are not moved. When the project finished, 6 out of the 24 experiments were considered successful and launched to full traffic, and the team wanted to know the total impact of all the launched new features.

Figure 6 shows the estimates of the total effect from the 6 experiments. The bottom-up estimate is S_A, where A is the index set of the 6 statistically significant experiments. The holdout estimate is an estimate that measured the total effect in another set of users

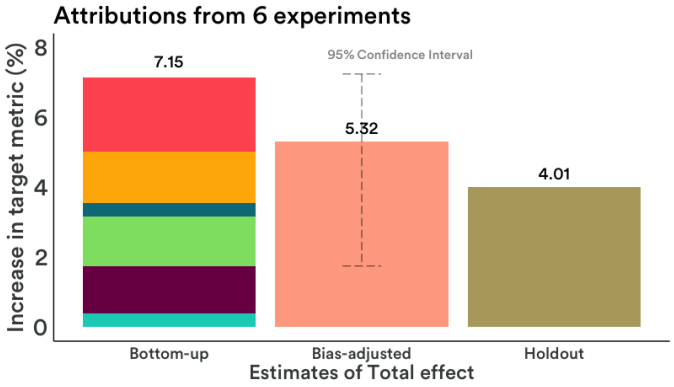


Figure 6: Comparison between different estimates

all at a single experiment, which we consider to be closer to the true effect. The bias-adjusted estimate is $\hat{T}_A$ taking account of the winner's curse bias. We observed that the bottom-up estimate was overestimating the true total effect substantially, and the winner's curs bias captures the overestimation and accurately adjusts it.

5.2 Implementation on Airbnb's experimentation platform

At Airbnb, experiments are conducted on an integrated platform called Experimentation Reporting Framework (ERF). More details can be found in [17, 18]. The platform is built and scaled up so that every product team can launch multiple experiments by self-serving UI tools. There are more than a hundred experiments starting each week, and in the meantime, the platform tracks more than 3.3k metrics that are computed both online and in the data warehouse.

Since the first iteration of ERF, the team has improved the platform in two directions. First, the team enhanced the back-end infrastructure to log data with better quality and reliability. For example, the team implemented an assignment service that mitigates the mixed-group issue for experiments on logged-out users. Second, the team enabled advanced statistical analysis on the platform and automated many popular functionalities. Winner's Curse bias correction is a good example of the latter category.

Figure 7 shows the user interface where one can input the experiment information. In this version, people pick the experiments of interest, including launched and non-launched ones. Also, one needs to specify a single metric and corresponding significance level used in their launch condition. Adjusting constants is another useful feature which we discuss in the next section.

Based on the inputs, the Winner's Curse calculator returns the results as shown in Figure 8. Intermediate values in the calculation and final aggregated values are presented. First, the web app reads from ERF's database to get the statistics of each experiment at the dates when they finished, including their relative effect (percent change), p-values and standard deviations of the relative proportions. In Figure 8, the set containing all experiments is $\{1, \ldots, 6\}$, and $A = \{3, 5, 6\}$ because their p-values are smaller than 0.05 and have positive effects. After calculation, each experiment gets the Winner's Curse bias estimate as well as the logarithm of the estimate. In the aggregated values, the bottom-up estimate S_{A_b} and

true effect estimate $\hat{T}_A$ are listed together with their difference, the bias, and the ratio. For example, u&c_experiment_2 has negative effect hence is contributing very little bias (shown as 0 as it is rounded to the fourth decimal point), and u&c_experiment_6 has relatively large bias among the selected ones for its p-value being only a bit smaller than the significance level $\alpha = 0.05$. We can see that the biases are not very large in this example.

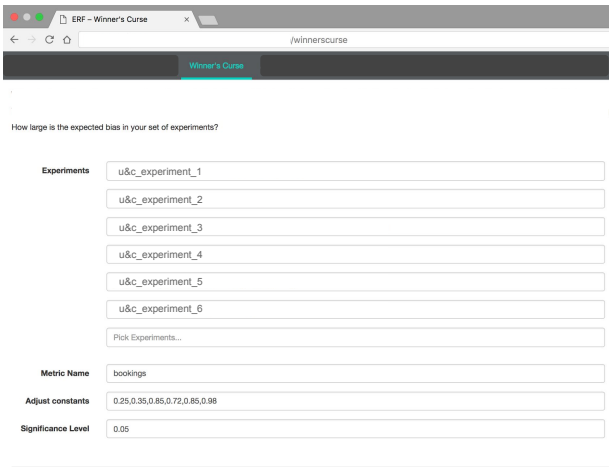


Figure 7: ERF Winner's Curse UI for inputs.

5.3 Implementation considerations

There are many situations at Airbnb facing a similar question and solution to the above example, and we found it common in other web companies as well. Now we deep dive into considerations and lessons learned when we designed and implemented Winner's Curse calculator in ERF, which is motivated by the practical needs occurred in the daily work with experimentation at Airbnb. They should be taken into account or adjusted accordingly when the experimentation platform has different advantages and challenges.

5.3.1 Adjusting with global coverage. Many new changes only affect a smaller group of users. For example, guests only see the urgency and commitment messages when they land on the search page or a listing page, but the messages do not impact guests who only stay on the front page. When the effects of messages are tested through A/B testing, we should only assign eligible users into the experiment so that the measured effect is not diluted by users that are not affected at all. It is a better experimental design because variance among assigned users can be reduced. For example, [4] discusses ways to mitigate this dilution. Suppose that the subset contains $\theta\%$ of total users, which is called the *global coverage* of the experiment. If a new feature leads to $a\%$ increase on a given metric, then when it is launched fully to the eligible subset globally the metric can only see $a\theta\%$ increase. It is also described in the Appendix of [24]. Therefore, we can easily adjust the estimate of the effect measured from each experiment using the global coverage as a linear constant for each term in (3) and (4). In the example above, we have implemented the function of adjusting constants as shown in Figure 7, which by default is set as the global coverage of each experiment. In other cases, we can also adjust the constants using prior knowledge to treat each experiment individually when the estimate is known to be over-estimated or under-estimated systematically.

5.3.2 Choosing an appropriate attribution set of experiments. As pointed out in (2) and further in Section 3.3, each experiment in the attribution set $\{1, \ldots, n\}$ contributes to the total bias. It is critical to determine the attribution set beforehand so that it does not depend on the results after running them, which can introduce another layer of bias not addressed by our method. In the above example, it is a natural choice to look at a group of experiments belonging to the same project, yet in practice there exist much more complicated details. As a rule of thumb to choose the attribution set appropriately, we recommend people to always refer to the hypothesis made a priori for each experiment: an experiment should be included if and only if that its launch condition is to have a meaningful lift on the target metric. For instance, sometimes a change is considered to be beneficial for the back-end infrastructure and will always be launched as long as it does not hurt top-line metrics. Then, it should not be included in the attribution set because the underlying hypothesis for this change is that it is most likely to be neutral with no expectation to lift the target metric.

Another situation is when an experiment has multiple treatments. Usually, the hypothesis for such an experiment is to launch only one of the treatments; if not, it should be designed and considered as an experiment with multiple factors or multiple experiments. Therefore, it is sufficient to only account for the single group in the multi-treatment experiment. We are not addressing the bias introduced by selecting from the multiple treatments, see [6] for further discussion. In ERF, the process has been designed in a way that anyone can pick a team, and the start date and end date. The Winner's Curse calculator returns all the experiments created by the team during that period, while one can edit and remove individual experiments.

5.3.3 Setting up hold-out group. In the above example, it is easy to propose launching a new experiment with all the selected changes combined as a straightforward way to measure the total effect. Indeed, a two-stage framework can be useful in removing the selection bias. However, it works only if we reserve a proportion of users as a hold-out group. The hold-out group should be set up ahead of time, and its users should not be assigned into any of the experiments. Since the second stage experiment only runs for the measurement purpose, it is quite expensive regarding not only taking up experimentation bandwidth and resources but also delaying the delivery of those features which are already known to be beneficial to users through the first stage experimentation.

On the other hand, we consider the hold-out group as one of the best practices and highly recommend setting it up on a manageable scale. Besides reducing measurement bias for launching experiments, it can also help us understand longer term performance of the product and eliminate seasonality effect. At Airbnb, we have set up hold-out groups for experiments launched by a few product teams, and they are updated on appropriate cadence.

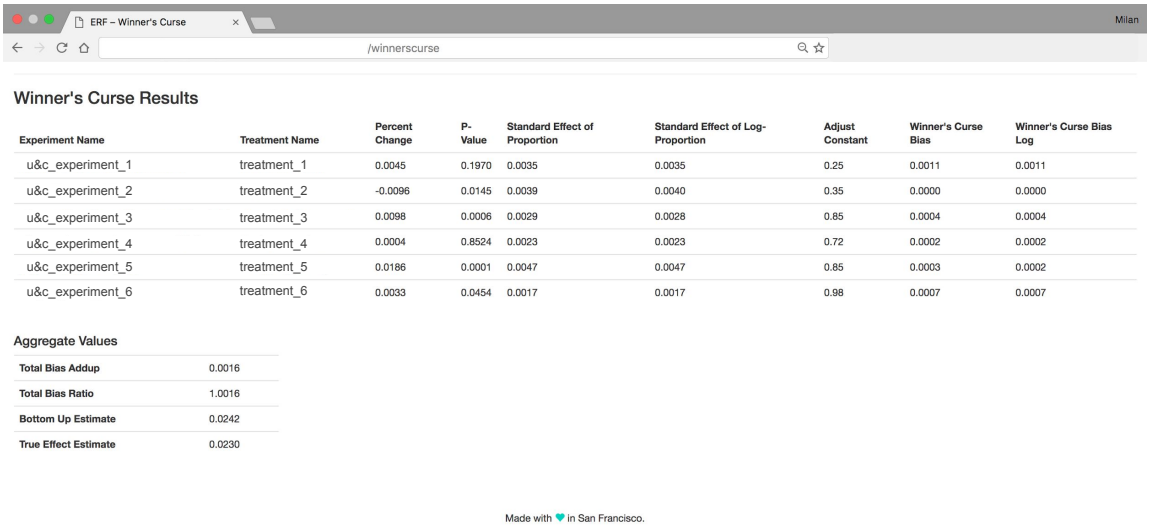


Figure 8: Winner's Curse Calculation results.

There are many more best practices discussed in the literature for online experimentation. Related to this work, we want to emphasize the importance of preventing cherry picking and setting up experiments with clarified hypotheses. In practice, it is common that a set of experiments have the bottom-up estimate more than double of the estimated true effects, which is a consequence of selecting a series of experiments just meeting the significance level. As shown in Figure 1, when the p-value is just cutting through the significance level, or in general when the experiment result is only making to the minimum launching condition, it contributes to the bias the most. There are various methods that mitigate this issue. For example, sequential testing and multiple testing can be easily applied and built into experimentation platforms.

6 CONCLUSIONS AND DISCUSSIONS

Measurement plays a crucial role in data-driven decision making. When online experiments are costly and have to be performed efficiently, it is inevitable to carry out measurements on the same data used for inference and model selection. We explicitly showed that the selection procedure from hypothesis testing introduces an upward bias in the aggregated estimate, in multiple online controlled experiment setting. We proposed a bias-corrected estimate and confidence intervals for the total effect of selected experiments.

In practice, various characteristics of online experiments make application of the theoretical work very challenging. Not every decision making follows the same rule. We covered a few implementation considerations in online experiments, with an application example of our bias correction in Airbnb's ERF platform. With our formulation, we can apply more specific implementation considerations to obtain a bias corrected measurement.

ACKNOWLEDGMENTS

We would like to thank Airbnb's Data Science team for their support on this work. We also appreciate Airbnb's ERF team for collaborating with the implementation and providing us the examples.

REFERENCES

[1] Theodore Alfonso Bancroft. 1944. On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics* 15, 2 (June 1944), 190–204.
[2] Theodore Alfonso Bancroft. 1964. Analysis and inference for incompletely specified models involving the use of preliminary test (s) of significance. *Biometrics* 20, 3 (Sept. 1964), 427–442.
[3] Edward C Capen, Robert V Clapp, William M Campbell, and others. 1971. Competitive bidding in high-risk situations. *Journal of petroleum technology* 23, 06 (June 1971), 641–653.
[4] Robert Chang. 2015. Detecting and avoiding bucket imbalance in A/B tests. (Dec. 2015). Retrieved February 16, 2017 from https://blog.twitter.com/2015/detecting-and-avoiding-bucket-imbalance-in-ab-tests
[5] 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.
[6] Alex Deng, Tianxi Li, and Yu Guo. 2014. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *Proceedings of the 23rd international conference on World wide web*. ACM, 609–618.
[7] Bradley Efron. 2011. TweedieâĂŹs formula and selection bias. *J. Amer. Statist. Assoc.* 106, 496 (Dec. 2011), 1602–1614.
[8] Bradley Efron. 2012. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, Cambridge.
[9] Bradley Efron. 2014. Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* 109, 507 (July 2014), 991–1007.
[10] Bradley Efron and Robert J Tibshirani. 1993. *An introduction to the bootstrap*. Chapman and Hall, London.
[11] William Fithian, Dennis Sun, and Jonathan Taylor. 2014. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597* (Oct. 2014).
[12] Chad Garner. 2007. Upward bias in odds ratio estimates from genome-wide association studies. *Genetic epidemiology* 31, 4 (May 2007), 288–295.
[13] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1168–1176.
[14] Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 959–967.
[15] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (Feb. 2009), 140–181.
[16] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, and others. 2016. Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44, 3 (April 2016), 907–927.

[17] Will Moss. 2014. Experiment reporting framework. (May 2014). Retrieved February 16, 2017 from http://nerds.airbnb.com/experiment-reporting-framework
[18] Jan Overgoor. 2014. Experiments at Airbnb. (May 2014). Retrieved February 16, 2017 from http://nerds.airbnb.com/experiments-at-airbnb
[19] Lei Sun and Shelley B Bull. 2005. Reduction of selection bias in genomewide studies by resampling. *Genetic epidemiology* 28, 4 (May 2005), 352–367.
[20] 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.
[21] Richard H Thaler. 1988. Anomalies: The winner's curse. *The Journal of Economic Perspectives* 2, 1 (Jan. 1988), 191–202.
[22] Rui Xiao and Michael Boehnke. 2009. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic epidemiology* 33, 5 (2009), 453–462.
[23] Lizhen Xu, Radu V Craiu, and Lei Sun. 2011. Bayesian methods to overcome the winner's curse in genetic studies. *The Annals of Applied Statistics* (2011), 201–231.
[24] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2227–2236.
[25] Hua Zhong and Ross L Prentice. 2008. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 9, 4 (Oct. 2008), 621–634.
[26] Hua Zhong and Ross L Prentice. 2010. Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic epidemiology* 34, 1 (Jan. 2010), 78–91.
[27] Sebastian Zöllner and Jonathan K Pritchard. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics* 80, 4 (April 2007), 605–615.