Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions

Alex Deng* Jiannan Lu* Jonathan Litz
Microsoft Corporation
{alexdeng,jiannl,jolitz}@microsoft.com

ABSTRACT

A/B tests (or randomized controlled experiments) play an integral role in the research and development cycles of technology companies. As in classic randomized experiments (e.g., clinical trials), the underlying statistical analysis of A/B tests is based on assuming the randomization unit is independent and identically distributed (*i.i.d.*). However, the randomization mechanisms utilized in online A/B tests can be quite complex and may render this assumption invalid. Analysis that unjustifiably relies on this assumption can yield untrustworthy results and lead to incorrect conclusions. Motivated by challenging problems arising from actual online experiments, we propose a new method of variance estimation that relies only on practically plausible assumptions, is directly applicable to a wide of range of randomization mechanisms, and can be implemented easily. We examine its performance and illustrate its advantages over two commonly used methods of variance estimation on both simulated and empirical datasets. Our results lead to a deeper understanding of the conditions under which the randomization unit can be treated as *i.i.d.* In particular, we show that for purposes of variance estimation, the randomization unit can be approximated as *i.i.d.* when the individual treatment effect variation is small; however, this approximation can lead to variance under-estimation when the individual treatment effect variation is large.

Keywords: Causal inference; randomization unit; random effect; delta method; asymptotic variance

1. INTRODUCTION

The statistical concept of randomization is nearly one hundred years old [26; 12; 13]. Randomized controlled experiments are often considered the gold standard of causal inference [30]. Online controlled experiments (or A/B tests) have long been utilized by technology companies (e.g., Amazon, Facebook, Google, LinkedIn, Microsoft, Netflix, Pandora, Twitter, Uber, etc.) to aid in data-driven decision making [19; 33; 3; 14; 35; 22]. To quote [18], "unlike most data mining techniques for finding correlational patterns, controlled experiments allow establishing a causal relationship [between a treatment and an outcome of interest] with high probability." Indeed, as pointed out in [8], A/B testing is widely recognized as "a basic pillar of Data Science."

In A/B test platforms at scale, the entire life cycle of an experiment (including traffic allocation, randomization, data collection, and analysis) is streamlined using an automated pipeline. It is of great importance that each component of this pipeline is trustworthy: Faulty design, software bugs, or flawed analysis can cast doubt on the validity of all results produced by the platform. In this paper we focus on making the statistical analysis of A/B tests more trustworthy by proposing a generalized method of variance estimation that is largely independent of the randomization mechanism.

A metric $M = \overline{Y}^{\text{obs}}$ is typically defined as the average of observations $Y_i, i = 1, \ldots, N$.[1] In an A/B test, traffic is assigned to either the treatment or the control. Let $\overline{Y}_T^{\text{obs}}$ and $\overline{Y}_C^{\text{obs}}$ be the point estimates of a metric from the treatment and the control, respectively. $\widehat{\tau} = \overline{Y}_T^{\text{obs}} - \overline{Y}_C^{\text{obs}}$ is the estimated difference in the metric between the treatment and the control. Because proper randomization guarantees *Ceteris paribus* (other things equal), it is intuitive (and rigorously provable under Rubin's causal model [16]) that $\widehat{\tau}$ is an unbiased estimator for the average treatment effect τ:

$$\tau = \mathbb{E}(\widehat{\tau}) = \mathbb{E}\left(\overline{Y}_T^{\text{obs}} - \overline{Y}_C^{\text{obs}}\right).$$

A fundamental problem in A/B testing is to determine if $\tau \neq 0$, i.e. to determine if there is a true nonzero treatment effect. On the surface, an A/B test is a natural extension of a classic completely randomized experiment, for which two-sample t-tests are used [16]. However, there are two major differences. First, online A/B tests typically have much larger sample sizes than classic randomized experiments, and as a result the central limit theorem [34; 7; 17] guarantees $\widehat{\tau}$ will approximately follow a normal distribution. Therefore statistical tests based on an asymptotic distribution of $\widehat{\tau}$ focus on estimating its asymptotic variance, and a z-test can substitute for a t-test in analyses. The second difference is more fundamental. In classic completely randomized experiments, observations Y_i are assumed to be independent and identically distributed (*i.i.d.*)[2], and observations from the treatment and control are assumed to be independent as well. Estimating the variance of $\widehat{\tau}$ is then straightforward: The variance is simply the sum of the variances of $\overline{Y}_T^{\text{obs}}$ and $\overline{Y}_C^{\text{obs}}$, both of which can be estimated using

*The first two authors contributed equally to this work.

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018677

[1] Most metrics used in A/B tests are averages, although percentile-based metrics are common in areas such as performance. This paper only focuses on average-based metric.

[2] This is the super population perspective in which observations are assumed to be drawn from a super population. The finite population viewpoint is slightly different but provides similar result. We adopt the super population perspective in this paper. See [1; 16] for comparison and discussion.

the standard sample variance formula.

$$\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\overline{Y})^2. \quad (1)$$

Online A/B tests can be much more complex. Observations Y_i can be correlated, and, depending on the randomization mechanism, observations from treatment and control can also be correlated, rendering the problem of variance estimation challenging. Under what conditions can we assume Y_i or observations at a certain unit are *i.i.d.* for purposes of variance calculation? What justifies the *i.i.d.* assumption and what invalidates it? Does randomization play a role here? If so, how? What can be done if the randomization mechanism is unknown?

This paper is motivated by the above challenges with a hope that a few simple results can be applied to a wide range of variance estimation problems that arise from online A/B tests. We prefer analytical and closed-form solution to computationally intensive methods such as bootstrapping [11; 15; 5]. It is worth noting that the bootstrap method does not circumnavigate the fundamental *i.i.d.* assumption because it relies on assuming that the unit sampled with replacement is *i.i.d.* in the ground truth data-generating-process.

To structure our discussion, we present three questions in order of increasing difficulty. We provide solutions for each of them. These challenging questions arise from A/B tests brought to the authors by our colleagues.

Question 1. *Under what conditions can users be treated as i.i.d.? In user-randomized A/B tests, users are commonly treated as i.i.d., but users are correlated by many factors such as gender, age group, location, organization, etc. Does this invalidate the user-level i.i.d. assumption?*

Question 1 reveals a gap between theory and practice. In theory, we would start with a data generating process (DGP) in which users are defined as *i.i.d.* by the design of the DGP. In practice, there are always questions about the plausibility of the DGP. In online experimentation, a user is often regarded as the basic unit of autonomy and is therefore used as the randomization unit. Users are commonly treated as *i.i.d.* without much questioning. We reveal a deeper connection between question 1 and the concept of *external validity* in Section 3.

A more general rule of thumb says that we can always treat the randomization unit as *i.i.d.* for purposes of variance calculation. We name this the *randomization unit principle* (RUP). For example, when an experiment is randomized by user, page-view [10], cookie-day [33] or session/visit, RUP suggests it is reasonable to assume observations at each of these levels respectively are *i.i.d.*. This assumption simplifies variance estimation because when the analysis unit of a metric is defined at the same level as the randomization unit, e.g. page level metrics in page-view randomized experiments, the standard sample variance formula (1) suffices. When the analysis unit is at a level lower than the randomization unit (e.g. page level metrics in a user randomized experiment) and the randomization unit is *i.i.d.*, the delta method [34; 10] provides the correct estimation of variance. Although no previous published work has explicitly stated RUP, it is widely used in analyses of A/B tests in the technology industry, and the importance of assuming the randomization unit is *i.i.d.* has been alluded to previously [33; 5; 20]. However, there is nothing inherent to the randomization process that justifies treating randomization unit level observations as *i.i.d.* When we randomize by page-view, are page-views from a single user not correlated? This begs the question of whether RUP is generally true.

Question 2. *Under what conditions does RUP hold and how reliable is it as an approximation when these conditions are not met?*

Most server side A/B tests have simple randomization mechanisms, which are typically implemented by applying a hash function on the id strings of a chosen randomization unit [19; 3]. For mobile apps, the mechanism of randomization can be much more complicated because mobile app clients are not always connected to the internet. Client side randomization is often done via polling, a process in which a client periodically requests a new configuration from the server when connected to a low cost data source such as WiFi [32]. Although randomization on the server side can still be implemented using a hash function on the randomization unit id, the server has no control over when a client will connect and fetch the most recent treatment assignment. Moreover, after fetching the new treatment assignment, the client may not immediately comply with the new treatment experience. For example, an experiment designer might want to randomize client apps every hour as a surrogate for randomization by client session. A true client session defined from app-open to app-close might extend for days or weeks if a user keeps the app open in the background and never fully closes the app. Hourly randomization is preferred when activities are clustered by actual-usage sessions, and each usage session typically does not last longer than a hour. Of course, to prevent a sudden change of experience for end users, when a client fetches a new treatment assignment and the app is in active use, the client app should be allowed to delay applying the new configuration. Analyzing mobile A/B tests becomes challenging because there can be a huge gap between the designed behavior, e.g. randomization by hour, and the actual experience on the client side due to the network connection issue and the delayed configuration refresh issue. This complicated or perhaps unknown randomization mechanism leads to Question 3.

Question 3. *How do we estimate the asymptotic variance of $\widehat{\tau}$ when the randomization mechanism is unknown? Is this estimation feasible with only practically plausible assumptions?*

By addressing these three challenging questions, we make the following original contributions in this paper:

1. We are the first to study the RUP *i.i.d.* assumption in depth with a focus on real-life A/B tests. Our results rely only on conditions that generally hold for large-scale A/B tests.
2. We show that RUP is *exactly correct* when there is no variance in the treatment effect. However, variance under-estimation occurs when there exists significant variance in treatment effect. We give the exact formula of the variance under-estimation.
3. We propose a semi-parametric variance estimation formula that is applicable to many randomization processes, including cases where the exact randomization mechanism is unknown. This formula can be applied to most real-world cases, and is straightforward to implement.
4. Our work was motivated by problems in online experiments. We applied the general variance formula on various randomization mechanisms and provide thorough simulation and real experiment examples. We believe the answers and solutions for these questions will be useful for many people running A/B tests and enable researchers to easily extend the idea behind our general variance formula to even more complicated scenarios.

The rest of the paper is structured as follows. After reviewing related work in the literature, we present our answer of Question 1 in Section 3 where a connection between the *i.i.d.* assumption and

external validity is made. Simulation study is presented to illustrate the idea. We then propose a general variance estimation formula in Section 4 and show how this formula nicely resolves Question 2 and 3, with the latter discussed further in Section 5 using a real-life experiment example. Section 6 concludes with final remarks and points to future works.

2. RELATED WORKS

The related issue of naively assuming the analysis unit is *i.i.d.* yielding a severe under-estimation of variance when this unit is not the same as the randomization unit has been previously documented [20; 33; 5]. In web site A/B testing this problem shows up most commonly user-randomized experiments for page-view analysis unit metrics. Previous work to correct this under-estimation assumed the randomization unit (user) is *i.i.d.* and used the delta method to compute the correct variance. The randomization unit (or experiment unit) is alluded to in these papers as an important unit to reason about *i.i.d.*. However there is no formal mention of the RUP in these works nor is there an attempt to prove it. [10] studied the randomization by page-view problem under a model with some restricted assumptions and managed to prove the principle only when there is no treatment effect (e.g. AA-tests). Their results are a special case of our results in this paper.

Instead of relying on a simple *i.i.d.* data generating process, Bakshy and Eckles [2] employed a two way random effect user-item model. They used Pigeonhole bootstrap [27] to estimate the correct variance for their average treatment effect estimator. In their problem the randomization unit was user, and the authors presented simulation studies comparing the user *i.i.d.* sample variance, item *i.i.d.* sample variance, and the bootstrap method variance. Our view in Section 3 is related to the random effect model.

3. I.I.D. ASSUMPTIONS AND EXTERNAL VALIDITY

In probability theory, independence is clearly defined. In plain speech, if two events are independent, knowing that one of those events occurred in no way affects the probability of the other event occuring. In practice, independence is rarely absolute and instead depends on context. Imagine there is an urn filled with numbered balls. You pull a ball from the urn, read its number, and place it back in the urn. If you do this repeatedly, are the outcomes independent? Surprisingly, such a seemingly simple question does not have a definitive answer. If the outcomes are independent, then previous outcomes have no predictive power for the next outcome. If you have no prior knowledge of the distribution of the numbers on the balls in the urn, then as you see more balls you develop a better understanding of this distribution and therefore a better prediction for the next outcome. In this view the outcomes are dependent. Alternatively, assume you know the distribution of the numbers on the balls (e.g., uniform from 1 to 500). Then pulling balls from the urn with replacement bears no new information, and the outcomes are independent. To summarize, the observed numbers are conditionally independent given the distribution, but unconditionally dependent. See [34; 25; 4] for more on conditional independence vs. independence.

Let us now think about Question 1 and treating users as *i.i.d.*. Users can be correlated by various factors such as gender, age, occupation, etc. Let us take gender as an example. It is known that the heights of men and women follow different distributions (see Figure 1). When the proportion of men and women in the population is treated as fixed, we can create the overall adult height distribution from the two gender specific distributions by weighting by the gender ratio. From this point of view, heights of randomly sampled adults are *i.i.d.* originating from a single mixture distribution – the "All Adults" density in Figure 1. The observed heights are conditionally independent given the gender mixture. We might then extend this reasoning to other factors and convince ourselves that the *i.i.d.* assumption for users is reasonable.

But what justifies treating the gender mixture as fixed? What if we want to make inferences about subsets of the population that may have different gender ratios? Really, we want to know if treating the gender ratio as fixed will affect the external validity of inferences made from the data. In A/B tests, external validity often concerns *bias* resulting from differences between the population from which the inference was drawn and the population upon which the inference is applied. When extending externally, inferences can be made invalid due to an *under-estimation of uncertainty*. Imagine that we sample students from 2 of 20 local 7th grade classes at random to estimate the average height of all local 7th graders. We assume that the height distributions of boys and girls in the 20 schools are the same. If we also assume the gender ratios in the 20 schools are fixed, then we can assume heights of sampled students are *i.i.d.* from a single mixture distribution and form a confidence interval for the mean of height. But what if even though the gender ratio is close to 50/50 in the whole school district, there exists large differences in gender ratios at different schools? Then it is possible that the two randomly chosen schools have significantly more boys than girls, or vice versa. Under this scenario the confidence interval we get by assuming a fixed gender ratio and *i.i.d.* heights is too narrow because it does not account for the variability of gender ratios among the 20 schools. To make the result externally valid, we must treat gender ratio as a random variable.

We use the following simulation to illustrate this point. Table 1 shows female and male student counts in 20 hypothetical schools. It is constructed such that:

1. Each school has exactly 1,000 student.
2. School #1 has 690 female and 310 male students. Each incremental school has 20 fewer females and 20 more males than the previous school. School #20 has 310 female and 690 male students.
3. In total, the gender ratio is balanced with exactly 10,000 female students and 10,000 male students.

We generate male heights from a normal distribution with mean 175 cm and standard deviation 10 cm. Female heights are generated from a normal distribution with mean 160 cm and standard deviation 10 cm. This simulated data is assumed to be the true heights for each of the 20,000 students. In this dataset, the true average height over the 20,000 students is 167.46. The goal is to sample 200 students out of the 20,000 students to estimate the average height.

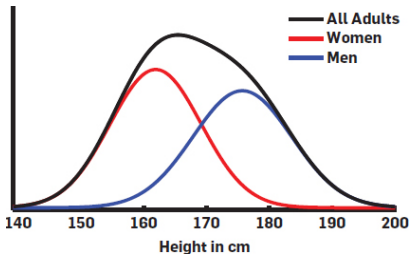

Figure 1: Mixture distribution for adult height. Source: cacm.acm.org

Table 1: Female and male students configuration in a hypothetical 20 school district.

School	Female Count	Male Count
1	690	310
2	670	330
3	650	350
4	630	370
5	610	390
...		
	...	...
15	410	590
16	390	610
17	370	630
18	350	650
19	330	670
20	310	690

Table 2: Comparison of 4 sampling mechanisms. The first is a random sample from all students; the other 3 cases select 2, 5 and 10 schools respectively and then sample students from the chosen schools. Results show that the standard variance formula (assuming *i.i.d.* samples) under-estimates the true variance and gives lower coverage than the promised 95% level.

Case	# School	Coverage	SE with standard formula	True SE
1	20	0.949	0.879	0.880
2	2	0.753	0.877	1.446
3	5	0.866	0.878	1.141
4	10	0.916	0.879	0.976

Table 2 shows the results from the 4 sampling methods. In Case 1 we randomly sample 200 students from the combined 20,000 students. For Cases 2 to 4, we use a two-stage sampling. For Case 2 we first sample 2 schools from the 20 schools, and then sample 100 students from each of the 2 chosen schools. In cases 3 and 4 we sample 5 and 10 schools, then 40 students and 20 students for each chosen school, respectively. In all cases we sample 200 students. For each case, after we sampled 200 students, we compute the sample average and also standard deviation from the standard variance formula assuming *i.i.d.* samples. We use that to compute the 95% confidence interval and record whether the true average 167.46 was within the interval. We repeated this process 2,000 times and then compute the coverage, average estimated standard error from the standard formula, and standard deviation of the 2,000 sample averages (which approximates the true standard error of the mean). Table 2 shows that case 1 has the correct coverage at 95% and the standard variance formula produces a standard error that is very close to the truth. In all other cases, coverages are all lower than the promised 95%, true standard errors are all larger than those given by standard formula, confirming variance under-estimation. In fact, in all cases the standard formula gives very similar standard errors. This is because it assumes *i.i.d.* sampled heights. However, when samples are from a small number of randomly selected schools, there is extra variation in the population due to the fluctuation of gender ratio, making the true standard error larger than under the *i.i.d.* assumption. We saw as the number of schools sampled in the first stage increased from 2 to 10, the true standard deviation decreased and coverage got better. In cases 2-4, if we assume samples are *i.i.d.* and use the standard variance formula, then we can only estimate the average height for the chosen schools and cannot extend the estimate to all 20 schools. If we want to make the extended estimation, we need to correctly account for additional "between school" variance.

To summarize, in reality users are often correlated by many factors. As long as the variation in the joint distribution of these factors is not a concern, and we understand that external validity might not hold if we apply the result to another population where the joint distribution is different, then we can treat this unit *i.i.d.* In online A/B tests, practitioners understand that using today's results to predict tomorrow's behavior is not perfect. The belief is that some small differences between the near future and the present will not be a big concern because we will keep iterating. By connecting the *i.i.d.* assumption with external validity, we now understand that the *i.i.d.* assumption is ultimately not justified by theory, but by *choice*.

There are notable exceptions in A/B tests and field experiments where *i.i.d.* assumptions for user should not be made because of external validity concerns. One example is when sampling a subset of geo-locations, or user clusters (organizations, etc.). Another example is in a user-item model where we do not want to treat item mixtures as fixed [2].[3]

4. A UNIFIED VARIANCE FORMULA

Knowing when and when not we can assume users *i.i.d.* is just the beginning. There are two important units, one is the randomization unit and the other is the analysis unit. The analysis unit is typically the denominator in a metric, e.g. page-view for page-click-rate and revenue-per-search, session for session-success-rate, etc. If we order different levels in a hierarchy, the randomization unit should always be higher or equal to the analysis unit. For example, any user-level metric would be ill-defined under page-level randomization, because the same user might be exposed to both the treatment and control. User is the most popular randomization unit because user is assumed to be a smallest unit of autonomy and randomizing by user keeps user experience consistent during the experiment. At the same time, as argued in Section 3, assuming *i.i.d.* users is reasonable in most cases. Since metrics are defined by observations at the analysis unit level, if the analysis unit is also user, then the interpretation of results is straightforward.

However, the majority of metrics in A/B testing are not user based, and user is not the only randomization unit being used. For instance, page-view randomized experiments are often used when results are not expected to be consistent over the course of the experiment (e.g. ad-related experiments). Denote individual, randomization, and analysis units by $\mathcal{I}$, $\mathcal{R}$ and $\mathcal{A}$. We define the individual unit as the level at which we can assume *i.i.d.*; this is typically the user level, but is not required to be. In general these three units can all be different as long as $\mathcal{I} \geq \mathcal{R} \geq \mathcal{A}$ where $\geq$ means higher than or equal to in the hierarchy.

In this section we provide a unified variance formula that is applicable for all choices of $\mathcal{R}$ varying between $\mathcal{I}$ and $\mathcal{A}$ including when $\mathcal{R}$ is unknown. We introduce the following notation to facilitate the flow. WLOG, here we take user as $\mathcal{I}$ and page-view as $\mathcal{A}$. The same hierarchy can be straightforwardly mapped to other domains.

4.1 Variance Estimation

Let $i = 1, \ldots n$ be the index of users. We adopt the potential outcomes framework [26; 28] to formalize our data-generation process. Under the Stable Unit Treatment Value Assumption [29] there

[3]The discussion here is closely related to the fixed effect vs. random effect in statistics and econometrics. Even in [2] the user as *i.i.d.* assumption is still practically very reasonable, unless in extreme cases where between item variances are large and also highly correlated with treatment effect.

is only one version of the treatment and no treatment interference among the users ($\mathcal{I}$). Let N_i be the number of page-views ($\mathcal{A}$).

Let $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{i,N_i})'$ denote the treatment assignment vector for user i, where $W_{ij} = 1$ if occurrence j is assigned to treatment and 0 otherwise. The observed outcome of occurrence j is

$$Y_{ij}^{\text{obs}} = W_{ij}Y_{ij}(1) + (1 - W_{ij})Y_{ij}(0) \quad (j = 1, \ldots, N_i),$$

and the numbers of occurrences assigned to treated and control are

$$N_{iT} = \sum_{j=1}^{N_i} W_{ij}, \quad N_{iC} = N_i - N_{iT}.$$

Note that we did not specify how W_{ij} is generated. We want our result to be valid for general randomization without needing to know the mechanism. This will be extremely important for solving Question 3.

To estimate the average treatment effect τ, let

$$S_{iT} = \sum_{j=1}^{N_i} 1_{\{W_{ij}=1\}} Y_{ij}^{\text{obs}}, \quad S_{iC} = \sum_{j=1}^{N_i} 1_{\{W_{ij}=0\}} Y_{ij}^{\text{obs}}$$

then the treatment and control metrics $\bar{Y}_T^{\text{obs}}$ and $\bar{Y}_C^{\text{obs}}$ are

$$\bar{Y}_T^{\text{obs}} = \frac{\sum_{i=1}^n S_{iT}}{\sum_{i=1}^n N_{iT}}, \quad \bar{Y}_C^{\text{obs}} = \frac{\sum_{i=1}^n S_{iC}}{\sum_{i=1}^n N_{iC}}.$$

Let $\mu_i = \text{E}Y_{ij}(0)$ be user i's mean outcome when assigned to control. And $\tau_i = \text{E}(Y_{ij}(1) - Y_{ij}(0))$ be her individual average treatment effect. Let the average number of page-views be $N = \text{E}(N_i)$.

Using this notation, the mean outcome for control group and the average treatment effect are

$$\mu = \frac{\text{E}(N_{iC} \cdot \mu_i)}{\text{E}(N_{iC})} \quad \tau = \frac{\text{E}(N_{iT} \cdot \tau_i)}{\text{E}(N_{iT})}$$

and the point estimators of N, μ and τ are

$$\widehat{N} = n^{-1} \sum_{i=1}^n N_i, \quad \widehat{\mu} = \bar{Y}_C^{\text{obs}}, \quad \widehat{\tau} = \bar{Y}_T^{\text{obs}} - \bar{Y}_C^{\text{obs}}.$$

We derive the asymptotic variance of $\widehat{\tau}$ under three mild assumptions.

1.Randomization: $\text{E}(N_{iT}|N_i) = pN_i$ with p fixed.

2.DGP: $(S_{iT}, S_{iC}, N_{iT}, N_{iC}, \mu_i, \tau_i), i = 1, \ldots, n$ are *i.i.d.*

3.Stable Denominator: Treatment does not impact the count of analysis unit $\mathcal{A}$. This is why we do not need to consider a pair $(N_i(1), N_i(0))$ and only use N_i.

Let's see why these three assumption are naturally satisfied. In A/B tests we always have fixed traffic splitting. Even when we do not know the randomization mechanism, it is reasonable to assume that the number of page-views assigned to treatment is expected to be p of the total. For each user, if the experiment is randomized by user $N_{iT} \sim N_i\text{Bernoulli}(p)$, e.g. all or nothing with probability p. If randomized by page-view, $N_{iT} \sim \text{Binomial}(N_i, p)$. This assumption is also true for unknown randomization mechanism as in Question 3. The second assumption assumes *i.i.d.* at individual level. If users can be assumed *i.i.d.*, we can let $\mathcal{I}$ be user level. Otherwise based on our discussion in Section 3 we can usually find some other independent unit. The last assumption is not a concern because metrics defined on a certain analysis unit are designed to use analysis unit as a normalization factor to highlight changes in the numerator. If a treatment impacts the denominator, this metric should not be used in decision making [9].

We are now ready for the main theorem:

Theorem 1. *The asymptotic variance of $\widehat{\tau}$ is*

$$\text{Var}_{asy}(\widehat{\tau}) = \frac{1}{n}\text{Var}\left\{\frac{S_{iT} - N_{iT} \cdot (\mu + \tau)}{pN} - \frac{S_{iC} - N_{iC} \cdot \mu}{(1-p)N}\right\}. \quad (2)$$

Theorem 1 is related to the statistical inference of the ratio estimator in survey sampling [6], and has several theoretical and practical advantages. First, it is *universal*; it is not only applicable to a wide range of randomization mechanisms, but it is also agnostic to the actual randomization mechanism, helping reduce computational overhead. In particular, this theorem holds even when the randomization mechanism is unknown or varies among subjects. Second, it is *robust*; it depends only on the aforementioned mild assumptions. Third, it is *actionable*; it is straightforwardly estimable by its finite-sample analogue

$$\widehat{\text{Var}}_{\text{G}}(\widehat{\tau}) = \frac{1}{n}\widehat{\text{Var}}\left\{\frac{S_{iT} - N_{iT} \cdot (\widehat{\mu} + \widehat{\tau})}{p\widehat{N}} - \frac{S_{iC} - N_{iC} \cdot \widehat{\mu}}{(1-p)\widehat{N}}\right\}, \quad (3)$$

henceforth referred to as the "general" formula. By the Law of Large Numbers, as $n \to \infty$

$$n\{\widehat{\text{Var}}_G(\widehat{\tau}) - \text{Var}_{\text{asy}}(\widehat{\tau})\} \xrightarrow{a.s.} 0.$$

Theorem 1 follows easily from the following Lemma, whose proof together with proof of Theorem 1 can be found in the appendix.

Lemma 2. *Let*

$$\widetilde{\tau} = \frac{\sum_{i=1}^n \sum_{j:W_{ij}=1}(Y_{ij}^{obs} - \mu - \tau)}{npN} - \frac{\sum_{i=1}^n \sum_{j:W_{ij}=0}(Y_{ij}^{obs} - \mu)}{n(1-p)N} + \tau,$$

then $\sqrt{n}(\widetilde{\tau} - \widehat{\tau}) \xrightarrow{D} 0$.

4.2 Existing Methods and the Randomization Unit Principle

Theorem 1 is a unified formula for different choices of the randomization unit $\mathcal{R}$. There are established variance estimation methods used in industry, namely standard sample variance formula and the "delta-method". The standard sample variance formula sees $\bar{Y}_T^{\text{obs}}$ and $\bar{Y}_C^{\text{obs}}$ as the sample averages of two *i.i.d.* random variables independent of each other.

$$\widehat{\text{Var}}_{\text{S}}(\widehat{\tau}) = \lambda_T^2 + \lambda_C^2, \quad (4)$$

where

$$\lambda_T^2 = \frac{1}{(\sum_{i=1}^n N_{iT})^2} \sum_{i=1}^n \sum_{j:W_{ij}=1} (Y_{ij}^{\text{obs}} - \bar{Y}_T^{\text{obs}})^2,$$

$$\lambda_C^2 = \frac{1}{(\sum_{i=1}^n N_{iC})^2} \sum_{i=1}^n \sum_{j:W_{ij}=0} (Y_{ij}^{\text{obs}} - \bar{Y}_C^{\text{obs}})^2;$$

The assumption behind the standard sample variance formula is satisfied when $\mathcal{I} = \mathcal{R} = \mathcal{A}$ because we assume $\mathcal{I}$ is *i.i.d.*. In practice it is common to set $\mathcal{I} = \mathcal{R} =$ user. However for many metrics $\mathcal{A}$ is different from user, e.g. page-view. In this case, using the standard formula will under-estimate the variance because page-views are wrongly assumed to be *i.i.d.*. When the randomization unit is strictly at lower in the hierarchy than user, we need to use the "delta-method"

$$\widehat{\text{Var}}_{\text{D}}(\widehat{\tau}) = n^{-1}(\xi_T^2 + \xi_C^2), \quad (5)$$

where

$$\xi_T^2 = \frac{1}{(\widehat{\mathrm{E}}N_{iT})^2}\widehat{\mathrm{Var}}(S_{iT}) + \frac{(\widehat{\mathrm{E}}S_{iT})^2}{(\widehat{\mathrm{E}}N_{iT})^4}\widehat{\mathrm{Var}}(N_{iT}) - 2\frac{\widehat{\mathrm{E}}S_{iT}}{(\widehat{\mathrm{E}}N_{iT})^3}\widehat{\mathrm{Cov}}(S_{iT}, N_{iT})$$

and ξ_C^2 defined similarly by replacing T with C.

As we can see, these two existing methods only covers two cases: 1) $\mathcal{I} = \mathcal{R} = \mathcal{A}$ and 2) $\mathcal{I} = \mathcal{R} > \mathcal{A}$ where user is commonly chosen as both $\mathcal{I}$ and $\mathcal{R}$. In practice, choosing a randomization unit other than user is becoming more common because user experience consistency is oftentimes not a design requirement. Exposing the same user to both treatment and control achieves an effect of "pairing" (as in a paired two-sample test) and thus produces higher statistical power. This lead us to the third case: 3) $\mathcal{I} > \mathcal{R} = \mathcal{A}$. The standard sample variance formula (4) is widely used in industry for this case. Because the standard sample variance requires $\mathcal{A}$ to be *i.i.d.*, the use of (4) for the third case is based on a rule of thumb that the randomization unit can be treated as *i.i.d.*. We refer to this rule of thumb as RUP – the *Randomization Unit Principle*.

Armed with Theorem 1, we have a general formula (3) that handles all three of the above cases and more. The following Corollary shows (3) and the "delta method" (5) are equivalent for case 2).

Corollary 3. *When $\mathcal{I} = \mathcal{R} > \mathcal{A}$, as $n \to \infty$,*

$$n\{\widehat{\mathrm{Var}}_D(\widehat{\tau}) - \widehat{\mathrm{Var}}_G(\widehat{\tau})\} \xrightarrow{a.s.} 0.$$

We see (3) and (5) asymptotically give the same estimation of variance even though their appearances are quite different. More interesting is the next result in which we evaluate the RUP.

Corollary 4. *When $\mathcal{I} > \mathcal{R} = \mathcal{A}$, as $n \to \infty$,*

$$n\{\widehat{\mathrm{Var}}_S(\widehat{\tau}) - \widehat{\mathrm{Var}}_G(\widehat{\tau})\} \xrightarrow{a.s.} -\mathrm{E}\{N_i(N_i-1)(\tau_i-\tau)^2\}/N^2. \quad (6)$$

This result is very interesting and surprising at first. What Corollary 4 says is that the RUP is *wrong*! There will always be an *under-estimation* of variance when we treat page-views as *i.i.d.* even when we randomize by page-view. This is surprising because the standard sample variance formula (4) has been used in practice with a long history. If there is a variance under-estimation, should it not already be known? Variance under-estimation leads to Type-I error inflation; how did experiments analyzed assuming RUP pass AA tests?

We make the following observations about the correction term $\mathrm{E}\{N_i(N_i-1)(\tau_i-\tau)^2\}/N^2$ in (6):

1. τ_i is the individual treatment effect. When there is no variance between individuals, this term is 0. As a special case, when there is no treatment effect, this term is 0.
2. With extra derivation omitted here, we can show the asymptotic variance of $\widehat{\tau}$ is on the order of $\mathrm{E}\{N_i(\mu_i-\mu)^2\}/N^2$. Comparing this to the correction term, we see that if the variance of τ_i is much smaller than the variance of μ_i, then the correction term relative to the true variance is small.
3. Assuming a constant multiplicative treatment effect of $x\%$, the variance of τ_i is $x\%^2$ the variance of μ_i. A rough estimation of the under-estimation is $x\%^2 \times \mathrm{E}(N_i)$.

These observations are very useful. It shows that even though RUP is not true in theory, in practice using the standard formula motivated by RUP when $\mathcal{R} = \mathcal{A}$ generates a variance close to the true one as long as the treatment effect is small. It is worth noting that when there is a large treatment effect, rejecting the null hypothesis is the desired outcome and variance under-estimation does not lead to false positive. Nevertheless, there might be cases where the average treatment effect τ is 0 or close to 0 but τ_i has large variance. In this case under-estimating variance by assuming RUP will lead to more false positive. In the next section we use simulation studies to further support our results.

4.3 Simulation Study

To examine the finite-sample performance of our methodology, we conduct a series of simulation studies. In this study, we set $\mathcal{I}$ to be user and $\mathcal{A}$ to be page-view. We introduce an additional level "session" so the randomization unit $\mathcal{R}$ can vary from user to session and to page-view. This additional session level serves as an unknown randomization unit, e.g. by-hour randomization with extra complexities as motivated by an actual mobile app experiment.

We let $p = 1/2$ so treatment and control are 50/50, and we let the sample size be $n = 10000$. For user $i = 1, \ldots, n$ let the number of sessions be $L_i \sim 1 + \mathrm{Pois}(3)$, and for each session, simulate the number of page-views from

$$R_{il} \sim 1 + \mathrm{Pois}(3) \quad (l = 1, \ldots, L_i).$$

We consider six simulation cases:

1. Cases 1–3 are indexed by the parameter $\sigma \in \{0, 1/2, 1\}$. We generate the potential outcomes by a Normal model, where $\mu_i \sim N(0,1)$ and $\tau_i \sim N(0, \sigma^2)$, and

$$Y_{ij}(1) \mid \mu_i, \tau_i \overset{iid}{\sim} N(\mu_i + \tau_i, 1), \quad Y_{ij}(0) \mid \mu_i, \tau_i \overset{iid}{\sim} N(\mu_i, 1)$$

2. Cases 4–6 are indexed by the parameter $\xi \in \{0, 1/4, 1/2\}$. We generate the potential outcomes by a Bernoulli model. Let $\mu_i \sim Unif(0, 1/2)$ and $\tau_i \sim Unif(0, \xi)$, and

$$Y_{ij}(1) \mid \mu_i, \tau_i \overset{iid}{\sim} \mathrm{Bern}(\mu_i + \tau_i), \quad Y_{ij}(0) \mid \mu_i, \tau_i \overset{iid}{\sim} \mathrm{Bern}(\mu_i).$$

The parameters σ and ξ quantify the variation of the treatment effect, formally defined as $\mathrm{TEV} = \mathrm{sd}(\tau_i)/\mathrm{sd}(\mu_i)$. To be specific, for Cases 1–3 we have TEV= σ, and for Cases 4–6 we have TEV= 2ξ. In particular, Case 1 ($\sigma = 0$) and Case 4 ($\xi = 0$) correspond to A/A tests. For each case, we consider the three randomization units of page, session, and user. For each randomization mechanism, we evaluate the performances of the three variance formulas (4), (5) and (3) over 3000 repeated samplings. To be specific, we obtain 3000 point estimates of the treatment effect, whose empirical standard deviation resembles the "true" standard deviation of the treatment effect estimator. In the meanwhile, by applying each variance formula we obtain 3000 estimated standard deviations, whose average characterizes the performance of the corresponding formula.

The simulation results are in Table 3, from which we can draw several conclusions. First, the standard formula works well in A/A tests when randomizing by page (RUP works), but under-estimates the true variance and under-covers the true parameter in other cases, confirming Corollary 4. The under-coverage of the confidence interval is more severe for the last case in each sub-table corresponding to a treatment effect with large variation, and is minor in the middle case when the treatment effect variation is smaller. Second, the delta-method formula works well for A/A and A/B tests when randomizing by user, but over-estimates the true variance and over-covers the true parameter in other cases, confirming Corollary 3. Third, the general formula correctly estimates variances and produces desirable coverage rates, for both potential outcome models, all randomization mechanisms, and both A/A and A/B tests.

Table 3: Simulation results for Normal and Bernoulli potential outcome models. In each sub-table, the first four columns contain the case label, randomization mechanism, treatment effect variation (TEV) and "true" standard deviation evaluated by 3000 repeated samplings, the next three columns contain the average standard errors of the point estimator of the standard ("S"), general ("G") and delta-method ("D") formula by the 3000 repeated samplings, and the last three columns contain the coverage rates of the 95% confidence intervals for the parameter τ. Sub-cases in each sub-table are ordered by the magnitude of treatment effect variation, with the first case being A/A, the second a treatment effect with small variation and third a treatment effect with large variation.

(a) Normal

Case	randomization	TEV	$\text{sd}(\widehat{\tau})$	asd_S	asd_G	asd_D	cover_S	cover_G	cover_D
1	Page	0.000	0.007	0.007	0.007	0.017	0.951	0.950	1.000
1	Session	0.000	0.012	0.007	0.012	0.018	0.750	0.952	0.997
1	User	0.000	0.023	0.007	0.023	0.023	0.467	0.943	0.943
2	Page	0.500	0.009	0.007	0.009	0.018	0.876	0.943	1.000
2	Session	0.500	0.013	0.007	0.013	0.019	0.711	0.949	0.995
2	User	0.500	0.024	0.007	0.024	0.024	0.444	0.952	0.952
3	Page	1.000	0.014	0.008	0.013	0.020	0.750	0.951	0.996
3	Session	1.000	0.017	0.008	0.017	0.022	0.635	0.947	0.988
3	User	1.000	0.028	0.008	0.028	0.028	0.419	0.955	0.955

(b) Bernoulli

Case	Randomization	TEV	$\text{sd}(\widehat{\tau})$	asd_S	asd_G	asd_D	cover_S	cover_G	cover_D
4	Page	0.000	0.002	0.002	0.002	0.003	0.954	0.955	0.994
4	Session	0.000	0.003	0.002	0.003	0.003	0.898	0.948	0.990
4	User	0.000	0.004	0.002	0.004	0.004	0.742	0.952	0.952
5	Page	0.500	0.002	0.002	0.002	0.003	0.938	0.953	0.996
5	Session	0.500	0.003	0.002	0.003	0.003	0.885	0.949	0.985
5	User	0.500	0.004	0.002	0.004	0.004	0.739	0.954	0.954
6	Page	1.000	0.003	0.002	0.003	0.004	0.898	0.950	0.983
6	Session	1.000	0.003	0.002	0.003	0.004	0.849	0.955	0.978
6	User	1.000	0.004	0.002	0.004	0.004	0.697	0.953	0.953

5. A/B TESTING WITH UNKNOWN RANDOMIZATION MECHANISM

We apply our newly-proposed method to a series of experiments from Skype, a leading messaging and VoIP mobile app by Microsoft. We focus on two real metrics that are important measures of user engagement and product quality: call duration, i.e., length of a call (in seconds), and call-dropped-rate, i.e., whether the call is disconnected due to software related issues. As mentioned in Question 3, hourly polling on the client side is implemented to fetch treatment assignment, along with numerous additional complexities to allow the client to delay configuration refresh. The analysis unit here is the call[4] and the randomization unit is obscured. These complexities pose no challenges to Theorem 1 since the result is general enough to work for an unspecified randomization unit $\mathcal{R}$.

120 real A/A experiments were run, with durations of two weeks, and the numbers of randomization units (i.e., users) more than one million. To thoroughly examine the impacts of the three variance estimation methods on p-value calculation and decision-making, we adopt the variance formulas (4), (5) and (3) to calculate the three variance estimates of the treatment effect estimators and corresponding p-values for the 120 experiments. Assuming no treatment effect and accurate variance estimation, the p-values would follow a standard uniform distribution. Therefore, we conduct K-S tests [31] to compare the empirical densities of the p-values by the standard, delta-method, and general formulas against the standard uniform density, and calculate the proportion of p-values less than 0.05.

[4]To be precise it is a call-leg as each call has both a caller and receiver. By only considering the call-legs, we can by-pass the more complex social network setting.

The results suggest that the general formula is optimal. To be specific, for *call duration*, the proportion of p-values by the standard, delta-method and general formulae that are smaller than 0.05 are 8.33%, 1.67%, and 5.00% respectively, and the K-S tests yield p-values of 0.144, 0.101, and 0.799 respectively. For *is call dropped*, the proportions of p-values by the standard, delta-method, and general formulae smaller than 0.05 are 15.83%, 1.67%, and 4.17% respectively, and the K-S tests yield p-values of 0.000, 0.132, and 0.279 respectively.

The above observations are not limited to only the two metrics described. Figure 2 shows the histogram of p-values for 10 different metrics from those 120 A/A experiments. We can see the standard sample variance formula (4) produces too many small p-values, while the "delta method" (5) gives p-value skewed to the right. The general formula (3) shows a reasonable shape resembling a sample histogram from uniform distribution. The K-S tests yield p-values of 0 for the standard formula, 0.003 for the "delta method" formula, and 0.152 for the general formula. In other words, only the standard formula produces p-values that are truly uniform.

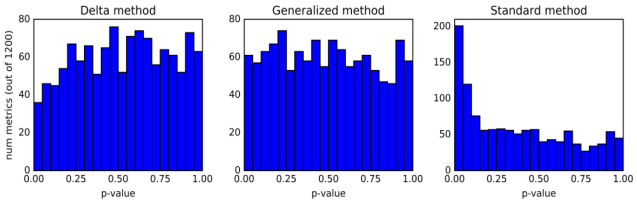

Figure 2: Histogram of p-values for 20 metrics in 120 A/A experiments combined. From left to right p-values are computed with variance estimation using the "delta method" (5), the general variance formula (3), and the standard sample variance formula (4).

6. CONCLUDING REMARKS

Many A/B tests are analyzed assuming by-user randomization guarantees *i.i.d.* samples. The standard sample variance formula is used for variance estimation, and the classic two-sample test is directly applied. However in real-life industrial A/B testing randomization mechanisms are more complex, and randomization units different from user are widely used, making variance estimation for the two-sample test challenging. Mobile experimentation adds more difficulty to this problem by uncontrollable client side behavior that can obscure the actual randomization mechanism. Above all these challenges, even the very basic user *i.i.d.* assumption has not been well justified in previous works.

Motivated by the need for a deeper understanding of the issue and a "unified" solution, we proposed a variance estimation method that only relies on practically plausible assumptions and is applicable to a wide range of randomization mechanisms. We also made a deep connection between the *i.i.d.* assumption and external validity – two seemingly unrelated concepts. We illustrate our thoughts and the unified variance formula by answering 3 challenging questions arise from real-life A/B testing. Detailed simulation and empirical results are included for each solution.

There are multiple possible future directions based on our current work. First, we can extend our current framework to re-randomization [23; 24]. Second, it is possible to derive parallel results for multi-arm experiments or factorial designs. Third, it would be interesting to incorporate the idea of covariate adjustment [6; 21] into our current framework. Fourth, we need to propose the Bayesian counter-

part of the current framework. All of the above are our ongoing or future research projects.

Acknowledgement

The authors thanks several colleagues at Microsoft, especially Yasaman Hosseinkashi and Jennifer Perret, for providing the Skype experimentation data sets, and Professor Peng Ding at UC Berkeley for several helpful suggestions.

APPENDIX

We provide proofs of Lemma 2, Theorem 1.

Proof of Lemma 2. First note that

$$\widehat{\tau} = \frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=1}(Y_{ij}^{\text{obs}} - \mu - \tau)}{\sum_{i=1}^{n} N_{iT}} - \frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=0}(Y_{ij}^{\text{obs}} - \mu)}{\sum_{i=1}^{n} N_{iC}} + \tau.$$

Therefore

$$\begin{aligned}\sqrt{n}(\widetilde{\tau} - \widehat{\tau}) &= \sqrt{n}\frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=1}(Y_{ij}^{\text{obs}} - \mu - \tau)}{\sum_{i=1}^{n} N_{iT}}\left(\frac{\sum_{i=1}^{n} N_{iT}}{npN} - 1\right) \\ &- \sqrt{n}\frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=0}(Y_{ij}^{\text{obs}} - \mu)}{\sum_{i=1}^{n} N_{iC}}\left\{\frac{\sum_{i=1}^{n} N_{iC}}{n(1-p)N} - 1\right\}.\end{aligned}$$

Since

$$E(S_{iT}) = E\{N_{iT}(\mu_i + \tau_i)\} = pE\{N_i(\mu_i + \tau_i)\}.$$

and $E(N_{iT}) = E\{E(N_{iT} \mid N_i)\} = pN$, by Law of Large Numbers

$$\frac{\sum_{i=1}^{n} S_{iT}}{\sum_{i=1}^{n} N_{iT}} \xrightarrow{a.s.} \mu + \tau,$$

as $n \longrightarrow \infty$, which implies that

$$\frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=1}(Y_{ij}^{\text{obs}} - \mu - \tau)}{\sum_{i=1}^{n} N_{iT}} \xrightarrow{a.s.} 0.$$

Additionally,

$$\begin{aligned}\sqrt{n}\left(\frac{\sum_{i=1}^{n} N_{iT}}{npN} - 1\right) &= \frac{\sqrt{n}}{pN}\left(n^{-1}\sum_{i=1}^{n} N_{iT} - pN\right) \\ &\xrightarrow{D} N\left\{0, \text{Var}(N_{iT})/(pN)^2\right\}.\end{aligned}$$

By Slutsky's Theorem,

$$\sqrt{n}\frac{\sum_{i=1}^{n}\sum_{j:W_{ij}=1}(Y_{ij}^{\text{obs}} - \mu - \tau)}{\sum_{i=1}^{n} N_{iT}}\left(\frac{\sum_{i=1}^{n} N_{iT}}{npN} - 1\right) \xrightarrow{D} 0,$$

and similarly

$$\sqrt{n}\frac{\sum_{i=1}^{n}\sum_{W_{ij}=1}^{N_i}(Y_{ij}^{\text{obs}} - \mu)}{\sum_{i=1}^{n} N_{iC}}\left\{\frac{\sum_{i=1}^{n} N_{iC}}{n(1-p)N} - 1\right\} \xrightarrow{D} 0,$$

which completes the proof. □

Proof of Theorem 1. By Lemma 2 we have $n\text{Var}(\widehat{\tau}) = n\text{Var}(\widetilde{\tau})$. Then note that

$$\widetilde{\tau} = n^{-1}\sum_{i=1}^{n}\left\{\frac{S_{iT} - N_{iT}(\mu+\tau)}{pN} - \frac{S_{iC} - N_{iC}\mu}{(1-p)N}\right\} + \tau.$$

and the fact that

$$\frac{S_{iT} - N_{iT}(\mu+\tau)}{pN} - \frac{S_{iC} - N_{iC}\mu}{(1-p)N} \quad (i = 1, \ldots, n)$$

are i.i.d. random variables, because all the components, i.e., S_{iT}'s, S_{iC}'s, N_{iT}'s and N_{iC}'s are i.i.d. Therefore the proof is completed. □

References

[1] Athey, S. and Imbens, G. [2016], 'The econometrics of randomized experiments', *arXiv:1607.00698* .

[2] Bakshy, E. and Eckles, D. [2013], Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods, *in* 'Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.', ACM, pp. 1303–1311.

[3] Bakshy, E., Eckles, D. and Bernstein, M. S. [2014], Designing and deploying online field experiments, *in* 'Proceedings of the 23rd international conference on World wide web', ACM, pp. 283–292.

[4] Barber, D. [2012], *Bayesian reasoning and machine learning*, Cambridge University Press.

[5] Chamandy, N., Muralidharan, O. and Wager, S. [2015], 'Teaching statistics at google-scale', *The American Statistician* **69**(4), 283–291.

[6] Cochran, W. G. [1977], *Sampling Techniques, Third Edition*, New York: W.W. Norton.

[7] DasGupta, A. [2008], *Asymptotic Theory of Statistics and Probability*, Springer.

[8] Deng, A., Lu, J. and Chen, S. [2016], Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing, *in* 'Proceedings of the 3rd IEEE International Conference on Data Science and Advanced Analytics'.

[9] Deng, A. and Shi, X. [2016], Data-driven metric development for online controlled experiments: Seven lessons learned, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'.

[10] Deng, S., Longbotham, R., Walker, T. and Xu, Y. [2011], 'Choice of the Randomization Unit in Online Controlled Experiment', *JSM Proc.* .

[11] Efron, B. and Tibshirani, R. J. [1994], *An introduction to the bootstrap*, CRC press.

[12] Fisher, R. A. [1925], *Statistical Methods for Research Workers, First Edition*, Edinburgh: Oliver and Boyd.

[13] Fisher, R. A. [1935], *The Design of Experiments, First Edition*, Edinburgh: Oliver and Boyd.

[14] Gomez-Uribe, C. A. and Hunt, N. [2016], 'The netflix recommender system: Algorithms, business value, and innovation', *ACM Transactions on Management Information Systems (TMIS)* **6**(4), 13.

[15] Hesterberg, T. C. [2015], 'What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum', *The American Statistician* **69**(4), 371–386.

[16] Imbens, G. W. and Rubin, D. B. [2015], *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press.

[17] Kohavi, R., Deng, A., Longbotham, R. and Xu, Y. [2014], Seven rules of thumb for web site experimenters, *in* 'Proc. 20th Conf. Knowl. Discov. Data Min.', KDD '14, New York, USA, pp. 1857–1866.

[18] Kohavi, R. and Longbotham, R. [2015], 'Online controlled experiments and A/B tests', *Encyclopedia of Meaning Learning and Data Mining* .

[19] Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R. M. [2009], 'Controlled experiments on the web: Survey and practical guide', *Data Mining and Knowledge Discovery* **18**, 140–181.

[20] Kohavi, R., Longbotham, R. and Walker, T. [2010], 'Online experiments: Practical lessons', *Computer (Long. Beach. Calif).* **43**(9), 82–85.

[21] Lin, W. [2013], 'Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique', *The Annals of Applied Statistics* **7**, 295–318.

[22] Lu, J. and Deng, A. [2016], 'Demystifying the bias from selective inference: A revisit to Dawid's treatment selection problem', *Statistics and Probability Letters* **118**, 8–15.

[23] Morgan, K. L. and Rubin, D. B. [2012], 'Rerandomization to improve covariate balance in experiments', *The Annals of Statistics* **40**, 1263–1282.

[24] Morgan, K. L. and Rubin, D. B. [2015], 'Rerandomization to balance tiers of covariates', *Journal of the American Statistical Association* **110**, 1412–1421.

[25] Murphy, K. P. [2012], *Machine learning: a probabilistic perspective*, MIT press.

[26] Neyman, J. [1923], 'On the application of probability theorey to agricultural experiments. Essay on principals. Section 9.', *Statistical Science* **5**, 465 – 480. [Translated by D. Dabrowska and T. Speed].

[27] Owen, A. B. et al. [2007], 'The pigeonhole bootstrap', *The Annals of Applied Statistics* **1**(2), 386–411.

[28] Rubin, D. B. [1974], 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of Educational Psychology* **66**, 688–701.

[29] Rubin, D. B. [1980], 'Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu.', *Journal of the American Statistical Association* **75**, 591–593.

[30] Rubin, D. B. [2008], 'For objective causal inference, design trumps analysis', *The Annals of Applied Statistics* **2**, 808 – 840.

[31] Smirnov, N. [1948], 'Table for estimating the goodness of fit of empirical distributions', *The Annals of Mathematical Statistics* **19**, 279–281.

[32] Tang, C., Kooburat, T., Venkatachalam, P., Chander, A., Wen, Z., Narayanan, A., Dowell, P. and Karl, R. [2015], Holistic configuration management at facebook, *in* 'Proceedings of the 25th Symposium on Operating Systems Principles', ACM, pp. 328–343.

[33] Tang, D., Agarwal, A., O'Brien, D. and Meyer, M. [2010], 'Overlapping Experiment Infrastructure: More, Better, Faster Experimentation', *Proc. 16th Conf. Knowl. Discov. Data Min.* .

[34] Wasserman, L. [2003], *All of Statistics: A Concise Course in Statistical Inference*, Springer.

[35] Xu, Y., Chen, N., Fernandez, A., Sinno, O. and Bhasin, A. [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', pp. 2227–2236.