Finite User Pool Effect in Two Sample t-test of Controlled Experiments on the Web

Shaojie Deng
Microsoft
Redmond, WA, 98052

Abstract

i.i.d. assumption is a basic assumption made in applications of two sample t-test in practice. In this paper, we show that when this assumption is not true and the sampling scheme is replaced by sample without replacement from a finite pool, the variance estimator assuming i.i.d. samples overestimates the variance. The overestimating effect is small or negaligile when the sample size is small relative to the finite pool size. However, the results in this paper is relavent to controlled experiment on the web because the capability of conducting controlled experiments using a large proportion of the whole web users.

Keywords

Controlled experiment, Experimentation, A/B testing, t-test, Variance estimation.

1 Introduction

For centuries people have been looking for ways to evaluate ideas. Controlled experiment, also called randomized test or A/B test has long established its importance as the methodology to establish a causal relationship. This paper will be focused on controlled experiment on the web. An obvious difference between controlled experiment on the web and other types of controlled experiment (for example, clinical trials) is that it is easy to collect data on web at low cost. In other words, web provides an unprecedented opportunity for us to use the power of controlled experiment to test and evaluate ideas quickly. It is our strong belief that unlocking the large amount of data on the web using the right methodology to analyze is the key toward a data driven philosophy, and controlled experiment has successfully set a standard in the industry. There are already many publications in the literature on controlled experiment. For a good and thorough survey on how to run web experiments, see Kohavi, Longbotham, Sommerfield & Henne (2009). Most of the works in the literature are focused on practical issues and best practices. To the author's knowledge few of them have been contributed on the underlying statistical methods. Part of the reason is that the related statistical method — the widely used two sample t-test are so well known, under i.i.d assumptions. In this paper, we show that when this assumption is modified in light of the fact that the total user pool is finite and sampling unit is not exactly i.i.d, the statistical analysis is more subtle and the results are different.

To better convey the idea, we need to introduce some terminology. A *experiment unit* is the unit on which the randomization is applied. The most commonly used experiment unit is user or its surrogate such as cookie. Page view has also been used in practice; see Tang, Agarwal, O'Brien & Meyer (2010). A *metric* is an statistic that stands for some concept the experiment designer wants to evaluate. Common metrics include clicks per user, sessions per user, click through rate, coverage rate, etc. A metric is naturally associated with an *analysis unit*. For example, per user metric such as clicks per user has the analysis unit user. Click through rate and coverage rate use page view as the analysis unit. The analysis unit associated with a metric is also called the *level* of the metric. The most important two types of metrics are user level metrics and page view level metrics. A *measurement* is an observation on a analysis unit. For example, the number of clicks of user i on page view j is a page view level measurement. It can be further rolled up (summed up) to user i's total number of clicks, which is a user level measurement.

In the following, we will focus on the case that user is the experiment unit. The general procedure of the statistical testing in web controlled experiment can be described as follows. Suppose we are interested in a metric S. After data was collected, we can calculate the value of this metric for both control group and treatment group, denoted by S_c and S_t. Under the assumption that user level measurement are i.i.d., central limit theorem guarantees that S_t and S_c is asymptotically normal and we can estimate the variance of S_t, S_c. Standard two-sample t-test can then be applied because $\frac{S_t - S_c}{\mathbb{V}ar(S_t - S_c)} = \frac{S_t - S_c}{\mathbb{V}ar(S_t) + \mathbb{V}ar(S_c)}$ are asymptotically normal[1]. The central player of this approach, is therefore to estimate the variances $\mathbb{V}arS_t$ (and $\mathbb{V}arS_c$).

The assumption that user level measurements are i.i.d. bears close examination. Because each user is a unique human being and the total number of users in the universe is a finite number, if we sample a fixed percentage x of total users for control or treatment, this sampling is sampling without replacement and theoretically it cannot be independent.[2] But on the other hand, the i.i.d. assumption has never been questioned because in most cases, we only sample a tiny portion of the total user pool and as the result, sampling without replacement and sampling with replacement does not differ to a degree that will change the result. For example, in clinical trial, the trial was only done to a cohort of people, hundreds or thousands. Comparing to the total population on earth which is in billion, it is safe to assert the i.i.d assumption and would actually be pedantic to question it.

For controlled experiment on the web, however, unlike the case for clinical trial, we do have the capability of experimenting on *all of the users that could show up*. It is not unusual to have a 50%/50% controlled experiment in which half of the total users received the control experience and the other half the treatment. For these cases, the finite user effect is no longer negligible and the variance calculation needs to be adjusted in light of this effect. In this paper, we show how the adjustment can be made from a simple formula. The adjusted variance estimate is smaller than the estimate under the i.i.d assumption. Therefore two sample t-test without this adjustment will be over conservative and losing power. To our knowledge, this problem has not been studied in previous publications and there are no related literatures.

The following paper is organized as follows. We first introduce notation in Section 1.1. In Section 2, asympototic results on the variance of both page level metric and

[1] When sample size is large, we could use normal distribution instead of t distribution. But we still call it t-test.

[2] This will be rigorously addressed in Section 3.

user level metric under the assumption of infinite user pool are presented. The main results of this paper are in Section 3, where we exam the finite user pool effect and propose simple formula called Formula X for variance estimation for both page level and user level metric. Section 4 presents simulation study that shows the performance of Formula X. Section 5 summarizes and concludes.

1.1 Notation

Before going into the next section, we introduce notation and assumptions which will be consistently used through out this paper. Denote n the total number of unique users. Let $X_{i,j}$ be the page level measurement (e.g. number of clicks on ths page) on user i's j^{th} page view and $X_{i,j}$ has mean μ_i and variance σ_i^2. Denote K_i the total number of page views from user i and $N = \sum_{i=1}^{n} K_i$ be the total number of page views. We assume for any i, $X_{i,j}, j = 1, \ldots, K_i$ are i.i.d. and uniformly bounded above by some finite constant. But we allow (μ_i, σ_i^2) to differ from user to user. We also assume $K_i, i = 1, \ldots, n$ are i.i.d. and independent of $(\mu_i, \sigma_i^2), i = 1, \ldots, n$. This last assmption may not be always true in practice and need to be checked case by case. We have checked this assumption for some key metrics of web experiments using empirical data and this assumption is reasonable.

2 Infinite User Pool Case

In this section, we assume there are infinite number of users that could show up. A direct impact of this assumption is that we can assume $(\mu_i, \sigma_i^2), i = 1, \ldots, n$ are i.i.d. realizations from some distribution. Note that (μ_i, σ_i^2) represents the user effect on the page level measurement $X_{i,j}$, and this model is a random effect model where the user effect is the random effect. We will derive the variance of the page level metric as well as that of the user level metric and also provide some theoretical results on some common estimators of the variance. Although the results might seem to be irrelevant to the topic of this paper, we present these results for the following reasons. First, they are interesting results that reveals important insights of the variances that are crucial for t-tests. Secondly, they provide the building blocks for Section 3 where we present the main results of this paper.

2.1 Page Level Metrics

A page level metric can be denoted by:

$$\overline{X} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_i} X_{i,j}}{N}.$$

To estimate the varaince of $\overline{X}$, it is tempting to treat page level metrics $X_{i,j}$, $j = 1, \ldots, K_i$, $i = 1, \ldots, n$, as i.i.d. and $\overline{X}$ under this assumption is an average of i.i.d. samples so the variance of $\overline{X}$ can be easily estimated by

$$\frac{1}{N^2}\Big(\sum_{i=1}^{n}\sum_{j=1}^{K_i}(X_{i,j} - \overline{X})^2\Big).$$

This estimator, which we call the *naive* estimator, is not consistent because unlike the fixed effect model, where the only randomness is from the noise of $X_{i,j}$.

In our model the user effect (μ_i, σ_i^2) are also a random sample from a distribution. Nevertheless, it is true in our model that the user level measurement $(\sum_{i=1}^{K_i} X_{i,j}, K_i), i = 1, \ldots, n$ are i.i.d. By letting $Y_i = \sum_{i=1}^{K_i} X_{i,j}$ and express $\overline{X}$ as $\sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} K_i$, it is then a straightforward application of the delta method to get an asymptotically consistent estimator for $\mathbb{V}ar\overline{X}$:

$$\frac{1}{n}\Big\{\frac{1}{\widehat{\mathbb{E}K_i}^2}\widehat{\mathbb{V}arY_i} + \frac{\widehat{\mathbb{E}Y_i}^2}{\widehat{\mathbb{E}K_i}^4}\widehat{\mathbb{V}arK_i} - 2\frac{\widehat{\mathbb{E}Y_i}}{\widehat{\mathbb{E}K_i}^3}\widehat{Cov(Y_i, K_i)}\Big\}$$

where these "hatted" quantities are the sample mean, variance and covariance.

For asymptotic analysis, we will let $n \to \infty$ (so $N \to \infty$ a.s.). To normalize the naive estimator and delta method estimator, we multiply them by n so that they will converge to some nonzero numbers. We introduce the normalized naive estimator

$$\widehat{\sigma_n^2} = n\frac{1}{N^2}\Big(\sum_{i=1}^{n}\sum_{j=1}^{K_i}(X_{i,j} - \overline{X})^2\Big). \tag{1}$$

and the normalized delta method estimator

$$\widehat{\sigma_d^2} = \frac{1}{\widehat{\mathbb{E}K_i}^2}\widehat{\mathbb{V}arY_i} + \frac{\widehat{\mathbb{E}Y_i}^2}{\widehat{\mathbb{E}K_i}^4}\widehat{\mathbb{V}arK_i} - 2\frac{\widehat{\mathbb{E}Y_i}}{\widehat{\mathbb{E}K_i}^3}\widehat{Cov(Y_i, K_i)} \tag{2}$$

A natural question to ask is how biased is the naive estimator $\widehat{\sigma_n^2}$ relative to the true normalized variance $n\mathbb{V}ar\overline{X}$. This is answered in the following theorem.

Theorem 1. *Let* $C = \frac{\mathbb{E}K_i^2}{(\mathbb{E}K_i)^2}$. *Then, as* $n \to \infty$,

$$n\mathbb{V}ar\overline{X} \to C\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i) \tag{3}$$

$$\widehat{\sigma_d^2} \to C\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i) \tag{4}$$

$$\widehat{\sigma_n^2} \to \frac{1}{\mathbb{E}(K_i)}(\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)). \tag{5}$$

Let $\rho := \mathbb{V}ar(\mu_i)/(\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2))$ *be the user effect coefficient(variances that explained by between user variation), then*

$$\frac{n\mathbb{V}ar(\overline{X})}{\widehat{\sigma_n^2}} \to (\mathbb{E}(K_i)C - 1)\rho + 1. \tag{6}$$

The convergence in (4) *and* (5) *are in probability.*

Proof of Theorem 1. (4) follows directly from the property of the delta method. To prove (3), we first apply conditional variance formula by conditioning on $(\mu_i, \sigma_i^2, K_i, i = 1, \ldots, n)$. This gives

$$\begin{aligned}
\mathbb{V}ar\overline{X} &= \mathbb{V}ar\Big(\mathbb{E}\Big(\frac{\sum_{i=1}^{n}\sum_{j=1}^{K_i} X_{i,j}}{N}\Big|K_i, \mu_i, \sigma_i^2, i = 1, \ldots, n\Big)\Big) \\
&+\mathbb{E}\Big(\mathbb{V}ar\Big(\frac{\sum_{i=1}^{n}\sum_{j=1}^{K_i} X_{i,j}}{N}\Big|K_i, \mu_i, \sigma_i^2, i = 1, \ldots, n\Big)\Big) \\
&=\mathbb{V}ar\Big(\frac{1}{N}\sum_{i=1}^{n} K_i\mu_i\Big) + \mathbb{E}\Big(\frac{1}{N^2}\sum_{i=1}^{n} K_i\sigma_i^2\Big).
\end{aligned}$$

Let $w_i = K_i / \sum_{i=1}^n K_i = K_i/N$. Since by assumption K_i is independent of (μ_i, σ_i^2) and $N/n \to \mathbb{E}K_i$ as $n \to \infty$, we can further simplify the right hand side. First, by applying iterative expectation(frist conditioning on $w_1, \ldots, w_n$), we have

$$n\mathbb{E}\Big(\frac{1}{N^2}\sum_{i=1}^n K_i\sigma_i^2\Big) = \sum_{i=1}^n \mathbb{E}\Big(\frac{n}{N}w_i\sigma_i^2\Big) = \frac{1}{\mathbb{E}K_i}(\sum_{i=1}^n w_i)\mathbb{E}\sigma_i^2 = \frac{\mathbb{E}\sigma_i^2}{\mathbb{E}K_i} \tag{7}$$

where the second equality is by bounded convergence theorm(since $N/n \to \mathbb{E}K_i$ and $\sum w_i\sigma_i^2$ bounded) and the last equation is from $\sum w_i = 1$. Since $(\mu_i, \sigma_i^2), i = 1, \ldots, n$ are i.i.d.,

$$n\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i) = n\mathbb{E}(\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i|w_1, \ldots, w_n)) + n\mathbb{V}ar(\mathbb{E}(\sum_{i=1}^n w_i\mu_i|w_1, \ldots, w_n)) \tag{8}$$

$$= n\mathbb{E}(\sum_{i=1}^n w_i^2\mathbb{V}ar(\mu_i)) + n\mathbb{V}ar((\sum_{i=1}^n w_i)\mathbb{E}\mu_i) = n\mathbb{E}(\sum_{i=1}^n w_i^2)\mathbb{V}ar(\mu_i) \tag{9}$$

where the last equality is from the fact that the second term is 0. By simple algebra, $n\sum_{i=1}^n w_i^2 = \frac{\overline{K_i^2}}{\overline{K_i} \times \overline{K_i}}$, where $\overline{K_i^2}$ and $\overline{K_i}$ are sample mean of K_i^2 and K_i, respectively. By strong law of large number, $\overline{K_i^2} \to \mathbb{E}K_i^2$ a.s., $\overline{K_i} \to \mathbb{E}K_i$ a.s., therefore $n\sum_{i=1}^n w_i^2 = \frac{\mathbb{E}K_i^2}{(\mathbb{E}K_i)^2}$ a.s. Combine this result with (7) and (9), we've proved (3).

We now turn to the limit of $\widehat{\sigma_n^2}$.

$$\widehat{\sigma_n^2} = n\frac{1}{N^2}\Big(\sum_{i=1}^n\sum_{j=1}^{K_i}(X_{i,j} - \overline{X})^2\Big) = \frac{n}{N^2}\Big\{\sum_{i=1}^n\sum_{j=1}^{K_i}X_{i,j}^2 - N\overline{X}^2\Big\}$$

$$\to \lim_{n\to\infty}\Big(\frac{n^2}{N^2}\Big)\mathbb{E}(\sum_{j=1}^{K_i}X_{i,j}^2) - \lim_{n\to\infty}\Big(\frac{n}{N}\Big)(\mathbb{E}\mu_i)^2.$$

The last limit is from $(1/n)\sum_{j=1}^{K_i}X_{i,j}^2 \to \mathbb{E}\big(\sum_{j=1}^{K_i}X_{i,j}^2\big)$ and $\overline{X} \to \mathbb{E}\mu_i$ a.s., both by the strong law of large number. Using bounded convergence theroem and $N/n \to \mathbb{E}K_i$, and also $\mathbb{E}\big(\sum_{j=1}^{K_i}X_{i,j}^2\big) = \mathbb{E}K_i\mathbb{E}X_{i,j}^2 = \mathbb{E}K_i(\mathbb{E}\mu_i^2 + \mathbb{E}\sigma_i^2)$, (5) follows. □

2.2 User Level Metrics

A user level metric can be denoted by

$$\widetilde{X} = \frac{\sum_{i=1}^n\sum_{j=1}^{K_i}X_{i,j}}{n} = \frac{\sum_{i=1}^n Y_i}{n}$$

A consistent estimator for $n\mathbb{V}ar\widetilde{X}$ is the sample variance of Y_i. The following theorem gives the formula of $n\mathbb{V}ar\widetilde{X}$.

Theorem 2. *As $n \to \infty$,*

$$n\mathbb{V}ar\widetilde{X} \to \mathbb{E}K_i^2\mathbb{V}ar\mu_i + (\mathbb{E}\mu_i)^2\mathbb{V}arK_i + \mathbb{E}K_i\mathbb{E}\sigma_i^2.$$

This theorem is a special case of a more general theorem in Section 3. We defer the proof to Section 3.

3 Finite User Pool Effect

In this section, we consider the case that the number of all users that could show up is finite, although the number can be very large. We label these users by index $i = 1, \ldots, M$ and suppose they have their own (μ_i, σ_i^2) fixed but unknown. For an controlled experiment on web, we sampled a portion x of users from the M total users without replacement where $0 < x \leq 1$ to treatment group and did the same for control group. WLOG, in the following we let $n = xM$ and treat it as an integer and we will only focus on one group. Note that because users are sampled without replacement, (μ_i, σ_i^2) of those n sampled users are no longer i.i.d. This sampling without replacement scheme can be equivalently modeled as the following reshuffle scheme. Suppose we first shuffle the $(\mu_i, \sigma_i^2)_{i=1,\ldots,M}$ and then relabel these from 1 to M. We then select the first $i = 1, \ldots, n$ as the users that are sampled. The purpose of using this model is that we can still use index $i = 1, \ldots, n$ in the proof rather than introducing an additional layer of index S_i denoting the index of the i^{th}selected user.

We first study the limit of the normalized variance of a page level metric $\overline{X}$ and propose a consistent estimator in the following theorem:

Theorem 3. *Let* $C = \frac{\mathbb{E}K_i^2}{(\mathbb{E}K_i)^2}$. *As* $n \to \infty$,

$$n\mathbb{V}ar\overline{X} \to (C - x)\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i) \tag{10}$$

$$\widehat{\sigma_d^2} \to C\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i) \tag{11}$$

$$\widehat{\sigma_n^2} \to \frac{1}{\mathbb{E}(K_i)}(\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)), \tag{12}$$

where the measure for (μ_i, σ_i^2) *is the law corresponding to sample one* (μ_i, σ_i^2) *from* $(\mu_i, \sigma_i^2), i = 1, \ldots, M$. *Moreover,*

$$\widehat{\mathbb{V}ar\mu_i} = \frac{\widehat{\sigma_d^2} - \widehat{\sigma_n^2}}{\frac{\widehat{\mathbb{E}Ki^2}}{\widehat{\mathbb{E}K_i}^2} - \frac{1}{\widehat{\mathbb{E}K_i}}}, \quad \widehat{\mathbb{E}\sigma_i^2} = \widehat{\mathbb{E}K_i}\Big(\widehat{\sigma_d^2}^2 - \frac{\widehat{\sigma_d^2} - \widehat{\sigma_n^2}}{1 - \frac{\widehat{\mathbb{E}K_i}}{\widehat{\mathbb{E}K_i^2}}}\Big) \tag{13}$$

are consistent estimator for $\mathbb{V}ar\mu_i$ *and* $\mathbb{E}\sigma_i^2$ *respectively. Therefore a consistent estimator of* $n\mathbb{V}ar\overline{X}$ *is:*

$$(\widehat{C} - x)\widehat{\mathbb{V}ar\mu_i} + \widehat{\mathbb{E}\sigma_i^2}/\widehat{\mathbb{E}(K_i)} = \widehat{\sigma_d^2} - x\widehat{\mathbb{V}ar\mu_i}. \tag{Formula X}$$

Theorem 3 shows that under the finite user pool assumption, the normalized variance $n\mathbb{V}ar\overline{X}$ decreases linearly as x increases from 0 to 1. On one extreme, when x close to 0, the distinction between sampling with replacement and sampling without replacement is so small that it is safe to treat (μ_i, σ_i^2) as i.i.d. just as in the case of infinite user pool case treated in Section 2. On the other extreme, when $x = 1$, we have observed all the users in one group, the user effect is reduced because we will always have the same set of users (same set of $(\mu_i, \sigma_i^2), i = 1, \ldots, n$). In controlled experiment, since we have at least two groups, we will not let x close to 1. But there are cases that we allocated 50% of traffic to both control and treatment group, corresponds to the case $x = 0.5$. If we ignore the finite user pool effect, we will be using the delta method estimator $\widehat{\sigma_d^2}$ to estimate the normalized variance, which by theorem 3 is upward biased by a factor of $x\mathbb{V}ar\mu_i$. We name the consistent estimator Formula X where the "X" stands for its relationship with the parameter x and also stands for the fact that this estimator is a hybrid of $\widehat{\sigma_d^2}$ and $\widehat{\sigma_n^2}$.

Proof of Theorem 3. First note that the delta method estimator will converge to the limit of $n\mathbb{V}ar\overline{X}$ as if (μ_i, σ_i^2) were i.i.d. from the law corresponding to sample one (μ_i, σ_i^2) from $(\mu_i, \sigma_i^2), i = 1, \ldots, M$. In this case from Theorem 1, (11) follows. (12) follows from a similar argument as in the proof of Theorem 1, with the use of strong law of large numbers replaced by using ergodic theory on an infinite exchangeable sequence, which is strictly stationary; see DasGupta (2008, Chapter 10) and *Wikipedia: Exchangeable random variables* (n.d.).

To prove 10, first by conditional variance formula by conditioning on $(\mu_i, \sigma_i, K_i), i = 1, \ldots, M$, $\mathbb{V}ar\overline{X} = \mathbb{V}ar\Big(\frac{1}{N}\sum_{i=1}^n K_i\mu_i\Big) + \mathbb{E}\Big(\frac{1}{N^2}\sum_{i=1}^n K_i\sigma_i^2\Big)$. The second term after multiplied by n converges to $\mathbb{E}\sigma_i^2/\mathbb{E}K_i$ for the same argument in (7). For the first term,

$$n\mathbb{V}ar\Big(\frac{1}{N}\sum_{i=1}^n K_i\mu_i\Big) = n\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i) = n\mathbb{E}(\sum_{i=1}^n w_i\mu_i|w_1, \ldots w_n) + n\mathbb{V}ar(\mathbb{E}(\mu_i)\sum_{i=1}^n w_i)$$
$$= n\mathbb{E}\big(\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i|w_1, \ldots w_n)\big).$$

Since now μ_i are not independent, we need to consider their covariances:

$$\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i|w_1, \ldots w_n) = \sum_{i=1}^n w_i^2\mathbb{V}ar(\mu_i) + \sum_{i\neq j} w_iw_jCov(\mu_i, \mu_j).$$

It is trivial that $\mathbb{V}ar\big((\sum_{i=1}^n w_i)^2\big) = 0$ because $\sum w_i = 1$.

$$0 = \mathbb{V}ar\big((\sum_{i=1}^n w_i)^2\big) = \sum_{i=1}^n \mathbb{V}ar(w_i) + \sum_{i\neq j} Cov(w_i, w_j).$$

Hence $Cov(w_i, w_j) = -\mathbb{V}ar(w_i)/n$.

Similarly $\mathbb{V}ar(\sum_{i=0}^M \mu_i) = 0$ since $\sum_{i=1}^M \mu_i$ is invariant against shuffling, and the same argument implies $Cov(\mu_i, \mu_j) = -\mathbb{V}ar(\mu_i)/M$.

$$n\mathbb{E}\big(\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i|w_1, \ldots w_n)\big) = n\mathbb{E}(\sum_{i=1}^n w_i^2)\mathbb{V}ar(\mu_i) + n\sum_{i\neq j}\mathbb{E}(w_iw_j)Cov(\mu_i, \mu_j)$$
$$= n\mathbb{E}(\sum_{i=1}^n w_i^2)\mathbb{V}ar(\mu_i) + n \times n(n-1)\mathbb{E}(w_iw_j)(-\mathbb{V}ar(\mu_i)/M)$$

The first term converges to $C\mathbb{V}ar(\mu_i)$. For the second term,

$$n \times n(n-1)\mathbb{E}(w_iw_j)(-\mathbb{V}ar(\mu_i)/M) = n \times n(n-1)(\mathbb{E}^2(w_i) + Cov(w_i, w_j))(-\mathbb{V}ar(\mu_i)/M)$$
$$= n \times n(n-1)(\frac{1}{n^2} - \mathbb{V}ar(w_i)/n)(-\mathbb{V}ar(\mu_i)/M) \to -x\mathbb{V}ar(\mu_i)$$

where the last step uses the fact that $n\mathbb{V}ar(w_i) \leq n\mathbb{E}(w_i^2) \to 0$ since $n\mathbb{E}(\sum_{i=1}^n w_i^2) \to C < \infty$. (10) is proved by combining results together. The rest of the theorem follows by (10),(11) and (12) with straightforward algebra. □

For user level metric $\widetilde{X}$, we have the following theorem that gives the limit of the normalized variance and a consistent estimator which we also call Formula X.

Theorem 4. *As $n \to \infty$,*

$$n\mathbb{V}ar\widetilde{X} \to \mathbb{E}K_i^2\mathbb{V}ar\mu_i - x(\mathbb{E}K_i)^2\mathbb{V}ar\mu_i + (\mathbb{E}\mu_i)^2\mathbb{V}arK_i + \mathbb{E}K_i\mathbb{E}\sigma_i^2,$$

where the measure for (μ_i, σ_i^2) are the law corresponding to sample one (μ_i, σ_i^2) from $(\mu_i, \sigma_i^2), i = 1, \ldots, M$. Moreover, using the consistent estimator of $\mathbb{V}ar\mu_i$ and $\mathbb{E}\sigma_i^2$ in Theorem 3 and $\widehat{\mathbb{E}\mu_i} := \overline{X}$ being a consistent estimator of $\mathbb{E}\mu_i$, a consistent estimator of $n\mathbb{V}ar\widetilde{X}$ is

$$(\widehat{\mathbb{E}K_i^2} - x\widehat{\mathbb{E}K_i}^2)\widehat{\mathbb{V}ar\mu_i} + \widehat{\mathbb{E}\mu_i}^2\widehat{\mathbb{V}arK_i} + \widehat{\mathbb{E}K_i}\widehat{\mathbb{E}\sigma_i^2}. \qquad \text{(Formula X)}$$

Proof. The proof is similar to the proof of Theorem 3. The key ingredient of the proof is the observation that for $i \neq j$, $Cov(\mu_i, \mu_j) = -\mathbb{V}ar(\mu_i)/M$. □

4 Simulation Study

In this section we use two simulation studies to show the performance of Formula X in Theorem 3 and Theorem 4.

To reflect the finite user pool assumption, we prescribed $M = 1,000,000$ number of users. Each of them have a fixed click through rate p_i. We generate these M p_i from a $Beta(0.1, 0.5)$ distribution. Figure 1 shows the shape of this distribution. Once these p_i's are sampled, they will be fixed throughout the whole simulation study.

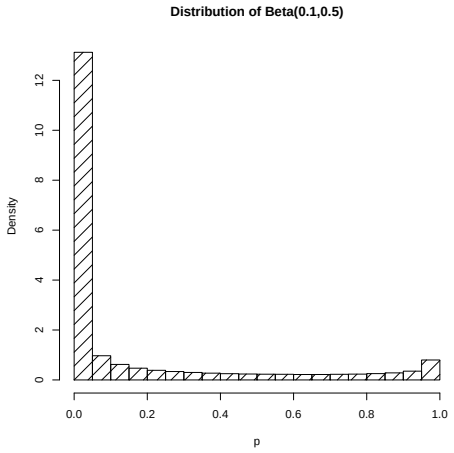

Figure 1: The distribution of users' page click rate parameter p_i from $Beta(0.1, 0.5)$ distribution.

We first test Formula X for page level metric $\overline{X}$. We use page click rate (PCR) as an example. The page level measurement $X_{i,j}$ for page click rate is binary with 1 stands for clicked and 0 not clicked. Metrics with the exact same mathematical form includes ad click rate, ad coverage rate, etc.

For a fixed x, each time, we sample a proportion x of users as well as their p_i from the M users without replacement. Then we simulate the page view K_i for each selected users from $Poisson(5)$. The rest of the simulation is straightforward, we simulate the number of page clicked by the user i from $Binomial(K_i, p_i)$. For this

simulated data, we then compute empirical PCR, $\widehat{\sigma_n^2}$, $\widehat{\sigma_d^2}$, and also apply Formula X in Theorem 3. On the other hand, to empirically estimate the true variance, we repeat this step for 1000 times and calculate sample variances of the 1000 empirical PCRs. Note that for each of the 1000 run we have an estimate from $\widehat{\sigma_n^2}$, $\widehat{\sigma_d^2}$ and Formula X. We take average of those 1000 estimates respectively to represent the final output of $\widehat{\sigma_n^2}$, $\widehat{\sigma_d^2}$ and Formula X. The last step is to repeat the preceding step for a grid of x varying from 0 to 1. We started with $x = 0.02$ and linearly grow x to 1 with step size 0.02.

To test Formula X for user level metric $\widetilde{X}$, we use the same page level measurement $X_{i,j}$ as above but roll up to user level. The simulation procedure is the same except for each fixed x, for each of the 1000 simulation run we calculate $\widetilde{X}$ and use the Formula X in Theorem 4. The results are summarized in Figure 2. The results

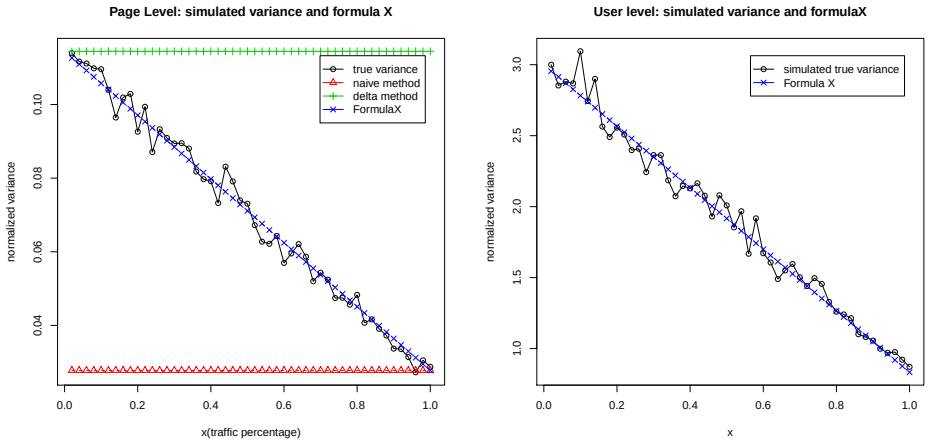

Figure 2: Left: Plot of 3 estimators for normalized variances for $\overline{X}$ (naive method, delta method, Formula X) and the estimated true variances. The formula X predicts linearly decreasing variances as x increase from 0 to 1. Right: Plot of the simulated variance of $\widetilde{X}$(after normalization) and the estimated value from Formula X.

show that Formula X performs quite well in both cases. The empirically estimated true normalized variance indeed varies around the value Formula X predicted. The prominent linearly decreasing trend agrees with Theorem 3 and 4.

5 Summary and Conclusions

In this paper we showed, for both page level metrics and user level metrics, the limiting behavior of the normalized variance under the i.i.d assumptions(infinite user pool) and under the finite user pool model. Moreover, we gave consistent estimators for both type of metrics which in both cases is a linear function of x— the percentage of users sampled per group. Simulation studies supports the theoretical results and highlighted the decreasing trend of the true limiting variances as x increases.

To see the implication of this finding, let's consider the page level metric $\overline{X}$. For general controlled experiments, it is a natural and general practice to assume an infinite pool of users and based on this very assumption people treat users as independent even though we know each individual is a unique human being.

The delta method is consistent if this assumption is right and a relatively common mistake people make is to use the naive formula which treats page level measurement as i.i.d. But Formula X shows that if we assume a finite user pool and let x goes to 1, the true normalized variances will decrease from the delta method estimate linearly to a smaller value that is closer to the estimate from the naive formula[3]. The value of Formula X when x equal to 1 is not exactly equal to the naive formula, but they are close for many cases and we can actually quantify it by comparing Theorem 1 and 3. Using delta method will only over-estimate the variance. Hence the type-I error will still get controlled. And when x is close to 0, true variance and the delta method(or corresponding bootstrap estimates, if bootstrap user with replacement was used) are very close. So there is little gain of statistical power to actually use Formula X for the calculation of variance. If $x = 0.5$, the effect of using Formula X can really increase the power. For example, in the simulation study in Section 4, the normalized delta method variance is about 0.114, the true variance was estimated to be 0.071. The true variance is about 62% of the delta method estimate. How will this affect the statistical power? Using the following approximated relationship[4] between power β, Type-I error α, the coefficient of variation (CV), the expected percentage change (treatment effect) $\delta\%$ and the sample size n_1 and n_2 of two groups,

$$\frac{1}{1/n_1 + 1/n_2} = cv^2\Big(\frac{\Phi_{\alpha/2} + \Phi_\beta}{\delta\%}\Big)^2, \tag{14}$$

if we further assume $n_1 = n_2 = n$ and $\delta\%$ fixed. Decreasing CV^2 by 62% is equivalent to increasing sample size n by $1/62\% = 1.61$. This is a significant gain because the experiments that we really need to allocate 50% to a group in the first place are generally those that we suffer from low sample size.

Theorem 3 and 4 assumes there is a static pool of users that the experiment designer want to make inference upon. It assumes we can collect data from all of them even if they didn't visit during the experiment and for these users by definition the number of page views K_i is 0. But in many cases, we are only interested in those users that are active, or at least have one visit during the experiment. More importantly, if the user identification is based on cookie, based on the data collected during the experiment alone, we might not know there exists a user unless the user showed up at least once during the experiment. In Appendix A, we show a simple adjustment on Formula X can be made for this cases. Another dependency we didn't cover in this paper is that in the model of finite user pool, the metric for control and treatment groups S_t and S_c are no longer independent. But we believe this dependency is week and summing up Formula X estimates for control and treatment groups will be very close to the variance of $S_t - S_c$.

Acknowledgements

The author wants to thank Roger Longbotham for the insightful discussion, consistent encouragement and support on this project. He also want to thank his colleague Ya Xu and Toby Walker for proposing this problem and their continuous help, discussion and suggestion during the project.

[3] In Figure 2, it looks like Formula X gives approximately the same number as the naive formula (the triangle points), this is just a coincidence. But in general, it is safe to say that Formula X when x goes to 1 gives a value that is much closer to the naive formula value than the delta formula value.

[4] Assume equal variance between control and treatment.

Appendices

A Adjustment for Inference on Active Users

The results in Section 3 are quite general. For example, it allows K_i to be 0, which means user i had no page view during the experimentation. For web experimentation, if the user identification is based on cookie assignment, then when the user never came during the experiment, we might not know whether there exists such a user. In practice, many analysis was done on those users actually appeared in the experiment, *i.e.*, with $K_i > 0$. Theorem 3 and 4 can be easily adjusted for this case.

To see how the adjustment can be made for non user level metrics $\overline{X}$, assume that we know the appearance rate for a fixed duration of the experiment is r. There are many ways to estimate this rate, the web site owner can get the idea of the number of active users by counting the stable users for a relatively long period of time, months, for example. r can be estimated by dividing the number of appeared users in the experiment by the total number of active users (scaled by the percentage of traffic that was allocated to the experiment).

An important observation is that $\overline{X}$ can be calculated using only appeared users since those "hidden" users will have zero page views and zero $X_{i,j}$'s that will not contribute to $\overline{X}$. The only thing changes is the estimation of $\mathbb{E}K_i$ and $\mathbb{E}K_i^2$. Suppose n_a is the number of users appeared, then the sample mean of K_i and K_i^2 for those appeared users is actually unbiased estimator for $\mathbb{E}(\widehat{K_i|K_i > 0})$ and $\mathbb{E}(\widehat{K_i^2|K_i > 0})$. Simple algebra reveals that $\widehat{\mathbb{E}K_i} = \mathbb{E}(\widehat{K_i|K_i > 0})r$ and $\widehat{\mathbb{E}K_i^2} = \mathbb{E}(\widehat{K_i^2|K_i > 0})r$. Hence $C_a := \frac{\mathbb{E}(\widehat{K_i^2|K_i>0})}{\mathbb{E}(\widehat{K_i|K_i>0})^2} = r\frac{\widehat{\mathbb{E}K_i^2}}{\widehat{\mathbb{E}K_i}^2} = rC$

If the analysis was done on those appeared users, then the normalized variance is defined as $n_a\mathbb{V}ar\overline{X}$. Therefore, from Theorem 3,

$$\begin{aligned} n_a\mathbb{V}ar\overline{X} = r \times n\mathbb{V}ar\overline{X} &\to r\big\{(C-x)\mathbb{V}ar\mu_i + \mathbb{E}\sigma_i^2/(\mathbb{E}(K_i|K_i>0)r)\big\}. \\ &= (C_a - xr)\mathbb{V}ar\mu_i + \mathbb{E}\sigma_i^2/\mathbb{E}(K_i|K_i>0) \end{aligned}$$

Also, note that if we apply delta method or naive method on those appeared users, we are essentially conditioned on $K_i > 0$ and we can estimate $\mathbb{V}ar(\mu_i|K_i > 0)$ and $\mathbb{E}(\sigma_i^2|K_i > 0)$ using (13) with the estimators all conditioned on $K_i > 0$, *i.e.*, use data from appeared users. Since (μ_i, σ_i^2) independent of K_i, this actually gives us consistent estimators of $\mathbb{V}ar\mu_i$ and $\mathbb{E}\sigma_i^2$. In another word, we don't need to know these "hidden" users to estimate $\mathbb{V}ar\mu_i$ and $\mathbb{E}\sigma_i^2$. The preceding argument shows that if we know the appearance rate r, the only adjustment we need to make upon Theorem 3 is that we need to replace x to rx and replace all other estimators by the conditioned version, *i.e.*, using data from appeared users. The same argument applies to user level metric. Define $\widetilde{X}_a = (\sum_{i=1}^{n_a}\sum_{j=1}^{K_i} X_{i,j})/n_a = (\sum_{i=1}^{n}\sum_{j=1}^{K_i})/n_a$. It can be shown that

$$\begin{aligned} n_a\mathbb{V}ar\widetilde{X}_a \to \big(\mathbb{E}(\widehat{K_i^2|K_i>0}) - rx\mathbb{E}(\widehat{K_i|K_i>0})^2\big)\widehat{\mathbb{V}ar\mu_i} \\ + \widehat{\mathbb{E}\mu_i}^2\mathbb{V}ar(\widehat{K_i|K_i>0}) + \mathbb{E}(\widehat{K_i|K_i>0})\widehat{\mathbb{E}\sigma_i^2}. \end{aligned}$$

The intuition for the adjustments is that the "effective" sampling percentage is xr rather than x. When the appearance rate r is small, the finite user effect is again negligible because xr is small.

References

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

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

Tang, Diane, Ashish Agarwal, Deirdre O'Brien & Mike Meyer (2010), 'Overlapping experiment infrastructure: More, better, faster experimentation', *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* .

Wikipedia: Exchangeable random variables (n.d.). `http://en.wikipedia.org/wiki/Exchangeable_random_variables`.