Choice of the Randomization Unit in Online Controlled Experiment

Shaojie Deng* Roger Longbotham† Toby Walker‡ Ya Xu§

Abstract

Controlled experiment has been used widely to support data driven decision making for on-line businesses. By applying appropriate randomization of the experiment units, causal inference can be established. The choice of the experiment unit for randomization can vary. User and page view are two mostly used units. Moreover, the analysis unit is sometimes different from the experiment unit. There are pros and cons in choosing which experiment unit to use and the choice affects the downstream statistical analysis. Generally for page level metrics, randomization by page will have an edge in power due to variance reduction. In this paper, we compare the two experiment units and provide a method to correctly analyze a page view randomization experiment in a two layer randomization framework.

Keywords

Controlled experiment, Experimentation, A/B testing, randomization unit, Variance estimation.

1. Introduction

For centuries people have been looking for ways to evaluate ideas. Controlled experiment, also called randomization 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 controlled experiment on the web and other types of controlled experiment (for example, clinical trials) is that it is easy and also with low cost to collect data on web. 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 unlock the huge data on the web and use the right methodology to analyze it 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,

*Microsoft, Redmond, WA 98052, USA, email: alexdeng@microsoft.com.
†Microsoft, Redmond, WA 98052, USA, email: roger.longbotham@microsoft.com.
‡Microsoft, Redmond, WA 98052, USA, email: toby.walker@microsoft.com.
§Microsoft, Redmond, WA 98052, USA, email: yaxu@microsoft.com.

we focus on the randomization step itself, and show how to analyze a randomized experiment where page view is used to randomize traffic.

We will be consistently using the following notations and assumptions through out this paper. Let n be the total number of unique users. Let $X_{i,j}$ be the per-page measurement (e.g. number of clicks on the 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. In particular, 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 assumption is for the purpose of theoretical investigation and need to be checked case by case in practice. We have checked this assumption for some key metrics of web experiments using empirical data and this assumption is reasonable. In a randomized experiment, we call the unit on which randomization is performed the randomization unit. In analysis phase, a metric will be naturally associated with an unit, which we call it analysis unit. For instance, a user level metric such as clicks per user is associated with user as the analysis unit while page view level metrics such as click through rate is associated with page view.

The following paper is organized as follows. In Section 2, we first briefly review the case where user is used as the randomization unit. In particular we are interested in page view level metrics and show how delta method should be used. We also give a formula for the bias introduced should we fail to use delta method. In Section 3, we shift the gear to the case that the randomization unit is page view. We present an asymptotically consistent variance formula of page view level metrics under a two layer randomization framework where a group of users are first recruited and then their page views are randomized into treatment and control. Section 4 presents simulation results. In Section 5, we discuss the pros and cons of using page view as randomization unit. We suggested extensions and concludes.

2. User as Randomization Unit

In this section we focus on the case where the randomization unit is user. In practice users are identified with their login ids or simply cookies stored by the browser. A detailed discussion of user tracking is far beyond the scope of this paper. For this paper, we assume user can be identified perfectly.

Traditional applications of user randomized experiment is to study movement of user level metrics. From a statistical perspective, this is a vanilla application of two sample t-test to test the null hypothesis that treatment group and control group are the same under the i.i.d assumption. To be more specific, a user level metric is a sample mean of user level measurements. Since users are randomized into control and treatment group, it is safe to assume user level measurements are independent and identically distributed. Under the hierarchical model in Section 1, this means we draw (μ_i, σ_i^2) independently for each user and then user level measurements for that particular user is drawn from corresponding distribution. Denote the user level metrics in control and treatment by $\widetilde{X}_T$ and $\widetilde{X}_C$, it is clear that they are independent. By central limit theorem,

$$\frac{\widetilde{X}_T - \widetilde{X}_C}{\sqrt{\mathbb{V}ar\{\widetilde{X}_T - \widetilde{X}_C\}}} \to Z, \tag{1}$$

where Z is standard normal. The two sample t-test is to replace $\mathbb{V}ar\{\widetilde{X}_T - \widetilde{X}_C\}$ by its estimate. In this particular case since $\widetilde{X}_T$ and $\widetilde{X}_C$ independent, $\mathbb{V}ar\{\widetilde{X}_T - \widetilde{X}_C\} = \mathbb{V}ar\widetilde{X}_T + \mathbb{V}ar\widetilde{X}_C$ and both terms in the right hand side can be simply estimated via sample variances. Also, for web experiment, n is large (much much larger than most applications of two sample t-test). Therefore we might as well treat t-statistics as standard normal.

For page view level metrics, (1) still holds when $\widetilde{X}_T$ and $\widetilde{X}_C$ are replaced by page view level metrics $\overline{X}_T$ and $\overline{X}_C$. However, $\mathbb{V}ar\{\overline{X}_T - \overline{X}_C\}$ can not be estimated by sample variances. Delta method is needed for variance estimation of $\mathbb{V}ar\overline{X}_g$, $g = T, C$; see Section 2.1. Moreover, when randomization unit is not user, $\overline{X}_T$ and $\overline{X}_C$ are not even independent and we need to estimate $\mathbb{V}ar\{\overline{X}_T - \overline{X}_C\}$ directly. This will be our topic in Section 3.

2.1 Page Level Metrics and Delta Method

Under our hierarchical model introduced in Section 1, a page level metric can be denoted by:

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

When user is the randomization unit, $\overline{X}_T$ and $\overline{X}_C$ are independent and $\mathbb{V}ar\{\overline{X}_T - \overline{X}_C\} = \mathbb{V}ar\overline{X}_T + \mathbb{V}ar\overline{X}_C$. Therefore we only need to focus on estimating $\mathbb{V}ar\overline{X}$. To this end, 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 in our model the user effect (μ_i, σ_i^2) are also a random sample from a distribution and page view level measurements $X_{i,j}$ of the same user are only independent conditioned on (μ_i, σ_i^2). 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 or 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

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

and the normalized delta method estimator

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

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* $n \to \infty$,

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

$$\widehat{\sigma_d^2} \to C\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i) \tag{5}$$

$$\widehat{\sigma_n^2} \to \frac{1}{\mathbb{E}(K_i)}(\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)). \tag{6}$$

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*

$$\frac{n\mathbb{V}ar(\overline{X})}{\widehat{\sigma_n^2}} \to (\mathbb{E}(K_i)C - 1)\rho + 1. \tag{7}$$

The convergence in (5) *and* (6) *are in probability.*

Proof of Theorem 1. (5) follows directly from the property of the delta method. To prove (4), 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 K_i 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. First, by applying iterative expectation(frist conditioning on $w_1, \ldots, w_n$), we have

$$n\mathbb{E}\Big(\frac{1}{N^2}\sum_{i=1}^n K_i\sigma_i^2\Big) = \sum_{i=1}^n \mathbb{E}\Big(\frac{n}{N}w_i\sigma_i^2\Big) = \frac{1}{\mathbb{E}K_i}(\sum_{i=1}^n w_i)\mathbb{E}\sigma_i^2 = \frac{\mathbb{E}\sigma_i^2}{\mathbb{E}K_i} \tag{8}$$

where the second equality is by bounded convergence theorem(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 (μ_i, σ_i^2) are i.i.d.,

$$n\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i) = n\mathbb{E}(\mathbb{V}ar(\sum_{i=1}^n w_i\mu_i|w_1, \ldots, w_n)) + n\mathbb{V}ar(\mathbb{E}(\sum_{i=1}^n w_i\mu_i|w_1, \ldots, w_n)) \tag{9}$$

$$= n\mathbb{E}(\sum_{i=1}^n w_i^2\mathbb{V}ar(\mu_i)) + n\mathbb{V}ar((\sum_{i=1}^n w_i)\mathbb{E}\mu_i) = n\mathbb{E}(\sum_{i=1}^n w_i^2)\mathbb{V}ar(\mu_i) \tag{10}$$

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 $n\sum_{i=1}^{n} w_i^2 \to \frac{\mathbb{E}K_i^2}{(\mathbb{E}K_i)^2}$ a.s. Combine this result with (8) and (10), we've proved (4).

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

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

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. By $N/n \to \mathbb{E}K_i$, and also $\mathbb{E}\big(\sum_{j=1}^{K_i} X_{i,j}^2\big) = \mathbb{E}K_i\mathbb{E}X_{i,j}^2 = \mathbb{E}K_i(\mathbb{E}\mu_i^2 + \mathbb{E}\sigma_i^2)$, (6) follows. □

(4) in Theorem 1 shows the variance of a page level metrics can be decomposed into two parts. The first part $C\mathbb{V}ar\mu_i$ is contributed by user effect (μ_i, σ_i^2). We call this between user variance. The second part $\mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i)$ represents the variance not explained by user effect and we might call this within user variance. When $\mathbb{E}K_i$ large, note that $C = \mathbb{E}K_i^2/(\mathbb{E}K_i)^2 \geq 1$ by Jensen's inequality, the between user variance will dominate. This is intuitively easy to understand since within user variance decrease to 0 as each user has more and more page views.

In empirical data, we do see between user variance contributed a large proportion of the total variance, especially for those experiments ran for a few weeks. This observation motivated us to use page view as randomization unit for some experiments if the key metrics we are interested in are at page view level. A caveat here is that not all experiment can be done with page view level randomization, mainly due to the inconsistent user experience. We leave the discussion later in Section 5 and only focus on theoretical property in the next section.

3. Page View as Randomization Unit

3.1 A Two Layer Randomization Framework

Suppose all page views are randomly divided into different groups. By all page views we mean page views from all users that could show up. Under this framework, it is from the typical marginalization argument that we can treat page view level measurement as i.i.d. i.e., there is no user effect in the analysis because the page views are drawn from all users and no user selection variance is induced in this randomization scheme. Since page view level measurements are i.i.d., statistical analysis for page view level metrics is therefore straightforward.

The case that is of most interest is the following. We first randomly selected n user from all the users that could show up. n is usually only a small percentage of the total number M of users that could show up in the universe. Let us assume M is infinity and hence assume users are drawn independently. All page views from these n users are then randomly split into control and treatment. The goal is to make inference by comparing certain metrics in control and treatment. There are

at least two reasons that we favor this two layer randomization framework over the one layer framework mentioned in the previous paragraph. One is that we want to run more than one experiments simultaneously and for a particular page view randomization experiment we only allow part of the whole traffic and reserve other traffic for user randomization experiment. Another reason is that we might not want all users to experience the page level randomization experience, which could potentially be inconsistent or non-sticky. Section 5 covers this in more detail.

An obvious difference between page view randomization and user randomization is the availability of many user level metrics. If the randomization unit is page view, first of all, page views per unique user does not make any sense anymore. Other metrics that are often used in search experimentation that will be unavailable include sessions per unique user, queries per unique user, page loading time per unique users, etc. However, as we will see in Section 3.2, the strength of using page view as randomization unit is the improvement in variance reduction for page level metrics, hence a boost in statistical power.

3.2 Page Level Metrics

Denote a page view level metric as $\overline{X}_r = \sum_{i=1}^{n} \sum_{j=1}^{K_i{}^{(r)}} X_{i,j}^{(r)}/N_r$ where $r = 1, 2$ stands for control and treatment. In Section 2 , we never considered the variance of both control and treatment together. This is because the control and treatment groups have different users and since randomization is based on user, the metrics of the two groups are naturally independent. As a result the variance of the difference of the metrics is simply the sum of the two variances of the same metric in each group. What make things more complicated here is that under the two layer randomization framework, control and treatment share the same group of n users. It is now the page view, not the user that is randomized into two groups. Due to this very fact, $\overline{X}_1$ and $\overline{X}_2$ are no longer independent.

What we need is an asymptotically unbiased estimator for $\mathbb{V}ar(\overline{X}_1 - \overline{X}_2)$ when the page views are split into control and treatment with fixed weights. Under the same hierarchical model, conditioned on K_i, $K_i^{(r)}$ follows $binomial(K_i, p)$ distribution where p depends on the weights of treatment and control. If we only consider one group, say control. Then the only difference between this framework and that of Section 2 is that now $K_i^{(r)}$ follows from a different distribution(from K_i). But note that all the results in Section 2 does not depend on the distribution of K_i. Therefore all results in Section 2 directly apply on $\overline{X}_1$ (or $\overline{X}_2$). Particularly,we have the following proposition for free.

Proposition 2. *Let* $w_i^{(r)} = K_i^{(r)} / \sum_{i=1}^{n} K_i^{(r)}$ *and* $C_r = \frac{\mathbb{E}(K_i^{(r)})^2}{(\mathbb{E}K_i^{(r)})^2}$. *Then for* $r = 1, 2$

$$\widehat{\sigma_{nr}^2} \to \frac{1}{\mathbb{E}(K_i^{(r)})}(\mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)) \tag{11}$$

$$\widehat{\sigma_{dr}^2} \to C_r \mathbb{V}ar(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i^{(r)}). \tag{12}$$

What Proposition 2 says is exactly that if we apply naive formula or delta method formula to one group, we will get asymptotically unbiased estimator for the right hand side of (11) and (12), respectively.

To analyze $\mathbb{V}ar(\overline{X}_1 - \overline{X}_2)$, by applying conditioned variance formula as in the

proof of Theorem 1,

$$\mathbb{V}ar(\overline{X}_1 - \overline{X}_2) = \mathbb{V}ar\Big(\frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(1)}} X_{i,j}^{(1)}}{N_1} - \frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(2)}} X_{i,j}^{(2)}}{N_2}\Big)$$
$$= \mathbb{V}ar\Big(\mathbb{E}\Big(\frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(1)}} X_{i,j}^{(1)}}{N_1} - \frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(2)}} X_{i,j}^{(2)}}{N_2} | K_i^{(r)}, \mu_i^{(r)}, \sigma_i^{(r)}, i = 1, \ldots, n, r = 1, 2)\Big)\Big)$$
$$+ \mathbb{E}\Big(\mathbb{V}ar\Big(\frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(1)}} X_{i,j}^{(1)}}{N_1^2} - \frac{\sum_{i=1}^n \sum_{j=1}^{K_i^{(2)}} X_{i,j}^{(2)}}{N_2^2} | K_i^{(r)}, \mu_i^{(r)}, \sigma_i^{(r)}, i = 1, \ldots, n, r = 1, 2\Big)\Big)$$
$$= \mathbb{V}ar\Big(\frac{1}{N_1}\sum_{i=1}^n K_i^{(1)}\mu_i - \frac{1}{N_2}\sum_{i=1}^n K_i^{(2)}\mu_i\Big) + \mathbb{E}\Big(\frac{1}{N_1^2}\sum_{i=1}^n K_i^{(1)}\sigma_i^2 + \frac{1}{N_2^2}\sum_{i=1}^n K_i^{(2)}\sigma_i^2\Big) \tag{13}$$

By using the short hand notation $w_i^{(r)}$, we can simplify $n\mathbb{V}ar(\overline{X}_1 - \overline{X}_2)$ into

$$n\mathbb{V}ar\Big(\sum_{i=1}^n (w_i^{(1)} - w_i^{(2)})\mu_i\Big) + n\mathbb{E}\Big(\sum_{i=1}^n (w_i^{(1)}/N_1 + w_i^{(2)}/N_2)\sigma_i^2\Big). \tag{14}$$

Comparing to (8), we see

$$n\mathbb{E}\Big(\sum_{i=1}^n (w_i^{(1)}/N_1 + w_i^{(2)}/N_2)\sigma_i^2\Big) \to \frac{\mathbb{E}\sigma_i^2}{\mathbb{E}K_i^{(1)}} + \frac{\mathbb{E}\sigma_i^2}{\mathbb{E}K_i^{(2)}} \tag{15}$$

where the last term is because $K_i^{(1)}$ has the same distribution as $K_i^{(2)}$.

By using conditional variance formula for another time and following the exact same argument as in (10) (replace w_i by $(w_i^{(1)} - w_i^{(2)})$), we have

$$n\mathbb{V}ar\Big(\sum_{i=1}^n (w_i^{(1)} - w_i^{(2)})\mu_i\Big) = n\mathbb{E}\big(\sum_{i=1}^n (w_i^{(1)} - w_i^{(2)})^2 \mathbb{V}ar\mu_i\big)$$
$$= \Big(n\mathbb{E}\big(\sum_{i=1}^n (w_i^{(1)})^2\big) + n\mathbb{E}\big(\sum_{i=1}^n (w_i^{(2)})^2\big) - 2n\mathbb{E}\big(\sum_{i=1}^n w_i^{(1)} w_i^{(2)}\big)\Big)\mathbb{V}ar\mu_i. \tag{16}$$

In the proof of Theorem 1, we proved $n\mathbb{E}\big(\sum_{i=1}^n w_i^2\big) \to \frac{\mathbb{E}K_i^2}{(\mathbb{E}K_i)^2} = C$. Same argument can be extended to prove the following:

$$n\mathbb{E}\big(\sum_{i=1}^n (w_i^{(r)})^2\big) \to \frac{\mathbb{E}(K_i^{(r)})^2}{(\mathbb{E}K_i^{(r)})^2} = C_r, r = 1, 2$$
$$n\mathbb{E}\big(\sum_{i=1}^n (w_i^{(1)} w_i^{(2)})\big) \to \frac{\mathbb{E}(K_i^{(1)} K_i^{(2)})}{\mathbb{E}K_i^{(1)}\mathbb{E}K_i^{(2)}} := C_x.$$

Plugging into (16) entails

$$n\mathbb{V}ar\Big(\sum_{i=1}^n (w_i^{(1)} - w_i^{(2)})\mu_i\Big) \to (C_1 + C_2 - 2C_x)\mathbb{V}ar\mu_i. \tag{17}$$

Combining this with (15), we have proved the following.

Proposition 3. *Under the framework of this section, let* $C_r = \frac{\mathbb{E}(K_i^{(r)})^2}{(\mathbb{E}K_i^{(r)})^2}$ *and* $C_x = \frac{\mathbb{E}(K_i^{(1)}K_i^{(2)})}{\mathbb{E}K_i^{(1)}\mathbb{E}K_i^{(2)}}$. *As* $n \to \infty$,

$$n\mathbb{V}ar(\overline{X}_1 - \overline{X}_2) \to (C_1 + C_2 - 2C_x)\mathbb{V}ar\mu_i + \sum_{r=1,2} \frac{\mathbb{E}\sigma_i^2}{\mathbb{E}K_i^{(r)}}. \tag{18}$$

The remaining piece is to figure out what is $C_1+C_2-C_x$. The next result shows $C_1 + C_2 - 2C_x = \sum_{r=1,2} \frac{1}{\mathbb{E}K_i^{(r)}}$.

Proposition 4. *Suppose control has weight* p *and treatment weight* q.

$$C_1 + C_2 - 2C_x = \frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}. \tag{19}$$

Therefore,

$$n\mathbb{V}ar(\overline{X}_1 - \overline{X}_2) \to \Big(\frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}\Big)\Big(\mathbb{V}ar\mu_i + \mathbb{E}\sigma_i^2\Big). \tag{20}$$

Proof of Proposition 4. (20) follows from Proposition 3 and (19). Here we only prove (19). To see this, note that $K_i = K_i^{(1)}+K_i^{(2)}$ and $K_i^{(1)}$ follows $Binomial(K_i, p)$.

$$\begin{aligned}
\mathbb{E}K_i^{(1)} &= p\mathbb{E}K_i \\
\mathbb{E}K_i^{(2)} &= q\mathbb{E}K_i \\
\mathbb{E}\big((K_i^{(1)})^2\big) &= pq\mathbb{E}K_i + p^2\mathbb{E}K_i^2 \\
\mathbb{E}\big((K_i^{(2)})^2\big) &= pq\mathbb{E}K_i + q^2\mathbb{E}K_i^2 \\
\mathbb{E}K_i^{(1)}K_i^{(2)} &= p\mathbb{E}K_i^2 - pq\mathbb{E}K_i - p^2\mathbb{E}K_i^2 = pq\mathbb{E}K_i^2 - pq\mathbb{E}K_i.
\end{aligned}$$

By definition,

$$\begin{aligned}
C_1 + C_2 - 2C_x &= \frac{\mathbb{E}(K_i^{(1)})^2}{(\mathbb{E}K_i^{(1)})^2} + \frac{\mathbb{E}(K_i^{(2)})^2}{(\mathbb{E}K_i^{(2)})^2} - 2\frac{p\mathbb{E}K_i^2 - pq\mathbb{E}K_i - p^2\mathbb{E}K_i^2 = pq\mathbb{E}K_i^2 - pq\mathbb{E}K_i}{\mathbb{E}K_i^{(1)}\mathbb{E}K_i^{(2)}} \\
&= \frac{1}{(\mathbb{E}K_i)^2}\Big(\frac{1}{p^2}\mathbb{E}\big((K_i^{(1)})^2\big) + \frac{1}{q^2}\mathbb{E}\big((K_i^{(2)})^2\big) - \frac{2}{pq}\mathbb{E}K_i^{(1)}K_i^{(2)}\Big) \\
&= (q/p + p/q + 2)\frac{1}{\mathbb{E}K_i} = (1/p + 1/q)\frac{1}{\mathbb{E}K_i}.
\end{aligned}$$

On the other hand,

$$\frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}} = (1/p + 1/q)\frac{1}{\mathbb{E}K_i}.$$

Hence $C_1 + C_2 - 2C_x = \frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}$. □

We can now summarize the result in this section into the following theorem.

Theorem 5. *Under the framework of this section, as* $n \to \infty$

$$n\mathbb{V}ar(\overline{X}_1 - \overline{X}_2) \to \Big(\frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}\Big)\Big(\mathbb{V}ar\mu_i + \mathbb{E}\sigma_i^2\Big).$$

Moreover $\widehat{\sigma_{n1}^2} + \widehat{\sigma_{n2}^2}$ *is an asymptotically unbiased estimator for* $n\mathbb{V}ar(\overline{X}_1 - \overline{X}_2)$.

In the following of this paper, we denote $\widehat{\sigma_{n1}^2}+\widehat{\sigma_{n2}^2}$ as Formula P, where P stands for "randomization by page view".

4. Simulation and Empirical Results

We use page click rate (PCR) as example to verify Formula P and show how it performs. We also compare the variance of PCR from a randomization by user experiment to that from a page view randomization experiment to empirically see the variance reduction from page view randomization. Like click through rate (CTR), page click rate is a page based metric focusing on users' click engagement on a page. The difference between PCR and CTR is that the page view level measurement of CTR is the total number of clicks on the page for a page view, while for PCR it is a binary number indicating whether a page view generates any click. For a fixed n, we first simulate p_i, $i = 1, \ldots, n$, the click through rate for a user from a $Beta(0.1, 0.5)$ distribution(see Figure 1 to get a sense of the shape of the distribution). We then simulate the total number of page view K_i from some distribution, which we can vary, and then use binomial distribution to split K_i into $K_i^{(1)}$ and $K_i^{(2)}$. For user i, we then simulate $\sum_{j=1}^{K_i^{(r)}} X_{i,j}^{(r)}$ from $Binomial(p_i)$. In each simulation run, we record $\overline{X}_1 - \overline{X}_2$, as well as $\widehat{\sigma_n^2}$, $\widehat{C_1}$,$\widehat{C_2}$, $\widehat{C_x}$, $\widehat{\mathbb{E}K_i^{(r)}}$. [1] We repeat this step for 1000 times. After the 1000 simulation run, we can estimate $\mathbb{V}ar(\overline{X}_1 - \overline{X}_2)$ from the sample variances of the 1000 realizations of $\overline{X}_1 - \overline{X}_2$, which we denote by $\widehat{\sigma_{sim}^2}$ for the normalized variances, which is n times the sample variance of $\overline{X}_1 - \overline{X}_2$. We also use bootstrap simulation(100 subsamples) to get an estimate of the standard deviation of the normalized variance estimator $\widehat{\sigma_{sim}^2}$. On the other hand, for each of these 1000 simulation run, we can apply Formula P to estimate the normalized variance. We then compare the distribution of these 1000 estimates from Formula P to the 95% confidence interval $(\widehat{\sigma_{sim}^2} - 1.96SD(\widehat{\sigma_{sim}^2}), \widehat{\sigma_{sim}^2} + 1.96SD(\widehat{\sigma_{sim}^2}))$. In all the simulation, we fixed $n = 100,000$ and $p = q = 0.5$.

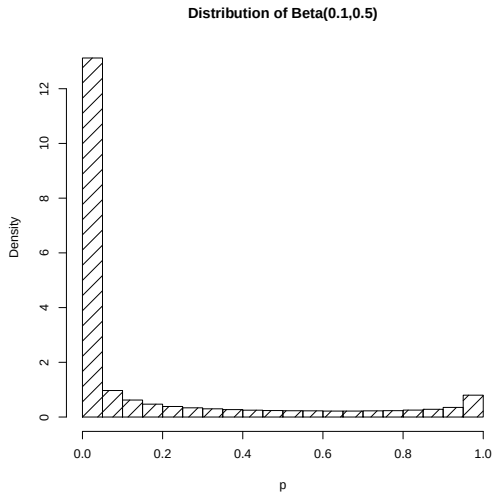

Figure 1: The histogram of p_i for $n = 100,000$ users from a $Beta(0.1, 0.5)$ distribution.

[1] For $\widehat{\sigma_n^2}$, $\widehat{C_1}$, $\widehat{C_2}$, $\widehat{\mathbb{E}K_i^{(r)}}$, we can actually calculate from both control and treatment and then take the average to get a more accurate estimate.

4.1 Performance of Formula P

We first use $Poisson(6)$ to generate K_i. The plot on the left in Figure 2 shows that $C_1+C_2-2C_x$ is indeed close to $1/\mathbb{E}K_i^{(1)}+1/\mathbb{E}K_i^{(2)}$. The ratio of the two is normally distributed and concentrated around 1. The plot on the right shows all the 1000 estimates from Formula P are within the bootstrapped 95% confidence interval.

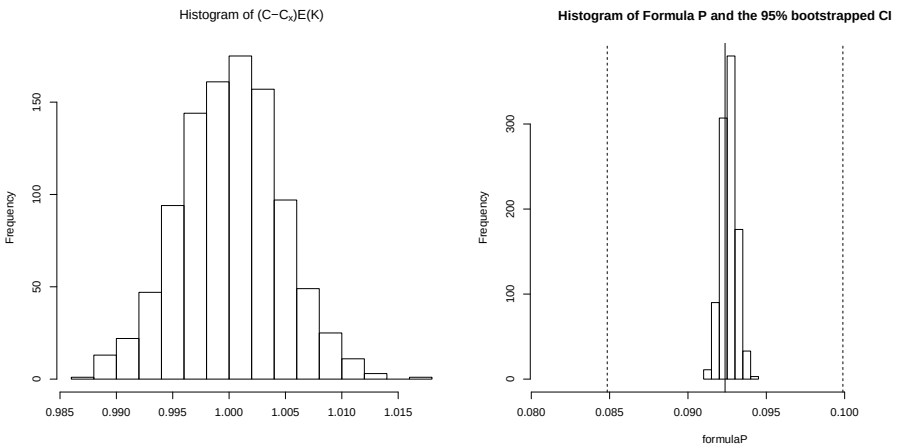

Figure 2: K_i from Poisson(6) distribution. Left: Histogram of $\widehat{(C1 - C_x)\mathbb{E}K_i^{(1)}}$. Right: Histogram of the 1000 estimates from Formula P and the 95% confidence interval form bootstrap. The two dashed lines are lower and upper bound of the confidence interval and the solid line is the sample variance of 1000 realization of $\overline{X}_1 - \overline{X}_2$ multiplied by n

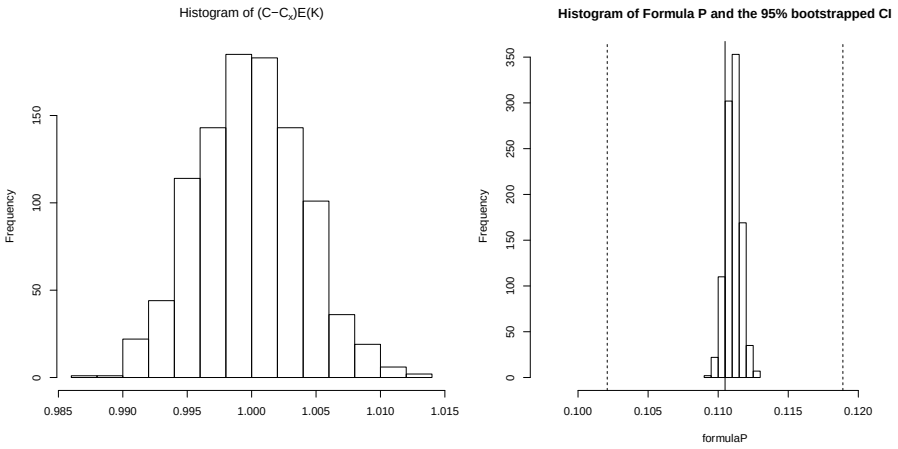

Figure 3: $K_i = 5$. Left: Histogram of $\widehat{(C_1 - C_x)\mathbb{E}K_i^{(1)}}$. Right: Histogram of the 1000 estimates from Formula P and the 95% confidence interval form bootstrap. The two dashed lines are lower and upper bound of the confidence interval and the solid line is the sample variance of 1000 realization of $\overline{X}_1 - \overline{X}_2$ multiplied by n

In Figure 3, we fixed $K_i = 5$, $i = 1, \ldots, n$. The simulation shows similar performance of Formula P.

4.2 Variance Reduction

In this simulation study, we compare the variance of PCR from a user randomized experiment with a page view randomized experiment. p_i follows from the same Beta distribution as in the previous section. We simulate K_i in two steps. First, simulate number of sessions of user i from a Poisson(2) distribution, then for each session, simulate number of page views from Poisson(3). K_i for user i is the sum of all page views from all sessions. To simulate a user randomization experiment, n users are then randomized into control and treatment group. To simulate a page view randomized experiment, for each user, K_i page views are randomized into control and treatment groups. For each cases, we run 1000 simulations just as we did in the previous section. Figure 4 shows the results of user randomization and page view randomization together. We can see that the variances when we randomize by user is around 0.64 while the variance when randomization is on page view is about 0.092. Therefore, there is a variance reduction at a factor of 7!

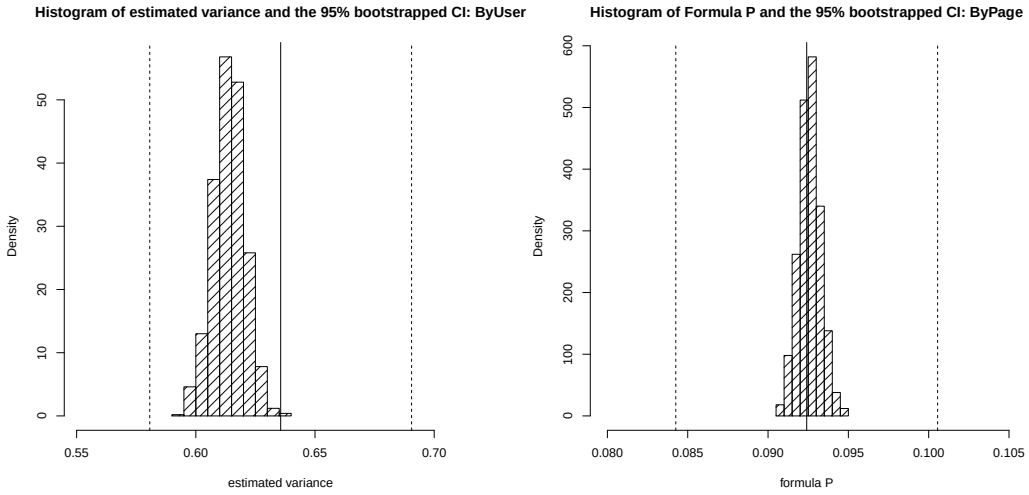

Figure 4: Randomization by user vs. randomization by page view.

5. Conclusion

In this paper, we have provided a way to analyze page view level metrics from an experiment with page view as the randomization unit, under a two layer randomization framework. In particular, we presented a formula for variance estimation and compare the variance from a page view randomized experiment to corresponding user randomized experiment and showed that page view randomized experiment leads to smaller variance for the same page view level metric.

Aside from theoretical property, in practice the more important topic is how do we choose randomization unit between user and page view. The most apparent distinction between the two is whether the user experience is consistent. If randomization is applied on page view, then by design a same user will receive both control and treatment experience. This non-sticky experience might be problematic and even a bad experiment design if the change between treatment and control experience is so large that swapping between them will cause huge user confusion. There are also many cases that one expects certain treatment effect that will require a

user to receive a consistent experience for a period of time for the effect to show up. Whenever consistent user experience is a must, randomization unit should be no finer than user level. Also, we cannot track any user level metrics if randomize by page view. For example, many loyalty metrics such as number of visits per user is at user level. Page view level randomization, on the other hand, still provide an intriguing alternative to user level randomization. As we have shown in this paper, it has an advantage in terms of statistical power on page level metrics due to the variance reduction. There are also cases when we want initial data collection for certain features that we do not want user to consistently experience. As an example, one might want to intentionally slowing down the page loading to study how the page abandon rate will change.

There are a few other alternatives besides user and page. One of them is to randomize by visits. This will guarantee that same user will have consistent experience within each visit, while still gains some variance reduction over pure user randomization by exposing the same user to both treatment and control on different visits. Another choice appeared in the literature was user-day randomization; see Tang, Agarwal, O'Brien & Meyer (2010). We mark the theoretical and empirical investigation of these different randomization units as future work.

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* .