Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix

Huizhi Xie
Netflix
100 Winchester Circle
Los Gatos, California, USA
kxie@netflix.com

Juliette Aurisset
Netflix
100 Winchester Circle
Los Gatos, California, USA
jaurisset@netflix.com

ABSTRACT

Controlled experiments are widely regarded as the most scientific way to establish a true causal relationship between product changes and their impact on business metrics. Many technology companies rely on such experiments as their main data-driven decision-making tool. The sensitivity of a controlled experiment refers to its ability to detect differences in business metrics due to product changes. At Netflix, with tens of millions of users, increasing the sensitivity of controlled experiments is critical as failure to detect a small effect, either positive or negative, can have a substantial revenue impact. This paper focuses on methods to increase sensitivity by reducing the sampling variance of business metrics. We define Netflix business metrics and share context around the critical need for improved sensitivity. We review popular variance reduction techniques that are broadly applicable to any type of controlled experiment and metric. We describe an innovative implementation of stratified sampling at Netflix where users are assigned to experiments in real time and discuss some surprising challenges with the implementation. We conduct case studies to compare these variance reduction techniques on a few Netflix datasets. Based on the empirical results, we recommend to use post-assignment variance reduction techniques such as post stratification [7] and CUPED [3] instead of at-assignment variance reduction techniques such as stratified sampling [2] in large-scale controlled experiments.

Keywords

Controlled experiment; Variance reduction; A/B testing; Randomized experiment; Sensitivity

1. BACKGROUND AND MOTIVATION

Controlled experiments are key for data-driven decisions in many technology companies. Running controlled experiments that are not sensitive enough to differences in business metrics caused by product changes can lead to suboptimal decisions with large revenue impact for companies like Netflix.

There are three ways to improve the sensitivity of controlled experiments: increasing sample sizes in the experiments, designing product changes that lead to large differences in business metrics, and reducing the sampling variance of business metrics.

The simplest way to increase sensitivity is to increase sample sizes. While the Netflix user base is very large, this option is not always practical. Many experimental product features affect only a small proportion of the user base, e.g., testing a new kids search experience on Android tablets has a relatively small audience, which limits the sample sizes. Moreover, while Netflix runs over a thousand experiments per year, there is always a desire to increase the pace of innovation by scaling up the number of experiments. With some experiments colliding with one another, available users for each experiment can become scarce. For these reasons, increasing the number of users assigned to experiments is often not feasible.

Two other avenues to improve the sensitivity of controlled experiments are explored in parallel at Netflix. Product managers lead cross-team efforts to focus on bold product changes that can lead to large positive differences in business metrics, while the experimentation team constantly seeks new experimentation methodologies to reduce the sampling variance of our business metrics.

This paper compares a few variance reduction techniques both theoretically and empirically based on a few Netflix datasets and provides guidance to experimenters on the choice of variance reduction techniques. Our primary contributions are threefold. First, we review the theory of three variance reduction techniques: stratified sampling [2], post stratification [7], and CUPED [3] and establish theoretical connections between them. Second, we describe an innovative implementation of stratified sampling at Netflix that addresses the challenges posed by assigning users to experiments in real time. Third, we conduct an empirical evaluation of these variance reduction techniques on a few Netflix datasets and compare their amount of variance reduction relative to simple random sampling.

2. CONTROLLED EXPERIMENTS AT NETFLIX

Netflix has a data-driven decision-making culture. We have learned through years of experimentation that using subjective intuition, even in a collective way, to make product decisions often yields the wrong answer. One way to

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2939733

make product decisions is to hear what users have to say. But what users ask for and what actually works are very different. Running controlled experiments and making product decisions based on business metrics is the best way to bridge this gap. Business metrics should be chosen such that improving them is highly related to increasing the value users get from the Netflix service. See [6] for some examples where actual experiment results do not agree with subjective intuition in the movie/TV recommendation algorithm area and a more complete description of how controlled experiments are used to improve our movie/TV recommendation algorithms.

At Netflix, controlled experiments are leveraged in many different product areas such as movie/TV recommendation algorithms, user interface design, and messaging. Different product experiences in experiments are referred to as *cells*. In each experiment, we are typically interested in comparing various new experience(s), referred to as *test cell(s)*, with the current production experience, referred to as the *control cell*. For example, in a controlled experiment in the movie/TV recommendation algorithm area, the control cell maps to the current production algorithm and the test cell(s) map to new algorithm(s) we want to compare with the production one.

2.1 Test Audience

At Netflix, controlled experiments are run on both new and existing users [6]. New users are assigned to experimental conditions at the time of signup, while existing users can be assigned anytime after their free trial ends. While product decisions rely on results from both cohorts, they are more heavily based on results from new users since they have not been exposed to the Netflix experience before. For existing users, it is difficult to tease apart whether a movement in business metrics is simply due to a change in experience (change effect) [6] or whether it is caused by the new experience itself. One way to remove such change effect is to run experiments longer and observe if the difference in business metrics persists after a long time. But this slows down our pace of innovation. The other reason to favor results from new users is because they are more sensitive to product changes since they start with a free trial during which they are in an evaluation mode. Note that testing on new users is not a common practice in the industry but is very important for Netflix to make product decisions for the reasons just explained.

2.2 Business Metrics

Netflix's monthly subscription business model suggests a framework to define business metrics for controlled experiments. Our revenue comes solely from the monthly subscription fee that current users pay, and current users can cancel the subscription at anytime. Thus we believe maximizing revenue through product changes is closely related with maximizing the value we provide to our users. Revenue is proportional to the number of users that is affected by three processes: the acquisition rate of new users, current user cancellation rate, and the rate at which former users rejoin. The focus of this paper is on product changes that directly impact only current users. Hence the primary business metric of interest is current user cancellation rate or retention rate. However, there are some challenges with just looking at retention rate in product experiments.

First of all, as much as we hope that better product or user experience can increase user retention rate, it can be affected by many other factors that are not directly related to our product changes. Secondly, since our subscription is month by month, users typically choose to cancel their subscription by their next payment period. For new users, it typically takes a whole month to observe retention since they are assigned to the experiments at the beginning of their first payment period. For existing users, the wait time to observe retention varies depending on the number of days it takes from the start of the experiments to the users' next payment period.

Fortunately, we have observed that user engagement metrics are highly correlated with retention but are more sensitive. Moreover, we get to observe such engagement metrics from the start of an experiment. One good example of user engagement metrics is streaming hours. However, the relationship between streaming hours and retention is not linear. What we have learnt from historical data is that getting users that stream few hours per month to stream more has a much larger positive impact on retention than getting users that already stream a lot to stream a bit more. This is because those users with low streaming hours are more likely to be one of those on the fence of cancellation and are more sensitive to product changes.

So we summarize the distribution of streaming hours using I streaming thresholds $T_i,\ i = 1, ..., I$. For a given user, T_i is a binary metric indicating whether the user streamed more than H_i hours in a given time period. H_i's are chosen to minimize the loss of information from summarizing the distribution of streaming hours using these thresholds. Details are not covered in this paper. From a business perspective, these streaming thresholds allow decision makers to gain more insight on which part of the distribution of streaming hours is changed. While we have tried more sophisticated versions of streaming measurement in the past, these thresholds work well because they are easy to understand without much loss of information.

3. REVIEW ON VARIANCE REDUCTION TECHNIQUES

3.1 Terminology and Notation

We define all the users that can potentially be impacted by an experiment as the *population* for the experiment. Suppose there is one or more variable(s) that are correlated with the business metrics. These variables are measurable prior to an experiment and independent of the different experiences in the cells of the experiment. As an example, the signup country of a user is correlated with how likely the user is to retain but does not depend on the experiences tested in the experiment. We refer to these variables as *covariates* and denote them as X. The two sampling schemes considered in this paper are simple random sampling and stratified sampling. In stratified sampling, covariates are used to divide the population into K subpopulations called strata. For example, since Netflix is available in 190 countries, we can divide the population for an experiment in 190 strata based on the signup country covariate.

Now we introduce some terminology and notations used throughout the paper. Note that the following notations

are used for both simple random sampling and stratified sampling.

- Random sample: a subset of users that are representative of the population
- Y: the business metric
- $T_i, i = 1..., I$: the binary streaming thresholds defined in Section 2.2
- $\mu = E(Y)$: the population mean of the business metric
- μ_k: the mean of the business metric for users in the kth stratum
- $\sigma^2 = var(Y)$: the population variance of the business metric
- σ_k^2: the variance of the business metric for users in the kth stratum
- p_k: proportion of the population in the kth stratum
- n_k: number of users from the kth stratum in a cell
- n: number of users in a cell from all the K strata, i.e., $n = \sum_{k=1}^{K} n_k$
- $Y_{11}, ...Y_{1n_1}, ...Y_{K1}, ..., Y_{Kn_K}$: business metrics of a random sample (either based on simple random sampling or stratified sampling) of users from the population where Y_{kj} is the business metric of the jth user from the kth stratum
- *Effect size*: difference between the population mean under the experience in the test cell and that in the control cell

Next we define two estimates of the population mean. The first is the standard simple sample average denoted as $\bar{Y}$. It is defined as

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}. \tag{1}$$

The second is a weighted average denoted as $\hat{Y}_{strat}$. It is defined as

$$\hat{Y}_{strat} = \sum_{k=1}^{K} p_k \bar{Y}_k, \tag{2}$$

where p_k is defined above and $\bar{Y}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj}$ is the average of the business metric for users from the kth stratum. Note that, under stratified sampling, the two estimates in (1) and (2) are the same. More details around why this is true are in Section 3.3. However, these two estimates are not the same under simple random sampling. This is the reason why post stratification leads to variance reduction. See Section 3.4 for the details. The subscript *strat* is used in the weighted average estimate (2) because it comes from stratified sampling. Throughout the paper, we use E_{srs} and E_{strat} to denote the expectation of an estimate under simple random sampling and stratified sampling, respectively. Similarly, we use var_{srs} and var_{strat} to denote the variance of an estimate under simple random sampling and stratified sampling, respectively.

3.2 Overview

Variance reduction is a procedure to increase the precision of the sample estimate of some parameter such as the population mean. The sample estimate is typically based on a random sample of the population. While a well-known procedure in statistics, Monte Carlo simulation [8], and some other areas, its application in controlled experiments is relatively new. Next we review a few popular variance reduction techniques that can be easily applied to controlled experiments.

As a starting point, we briefly review the statistical inference in controlled experiments. Suppose we are interested in comparing a test cell and the control cell in an experiment. Denote the business metric in the test cell and the control cell as $Y^{(t)}$ and $Y^{(c)}$, respectively. We start with a pair of hypotheses. The null hypothesis is that $Y^{(t)}$ and $Y^{(c)}$ have the same mean and the alternative is that they do not. The sample size in the experiments at Netflix is at least in thousands. Regular two-sample t-test is thus applied to test the hypotheses. The t-test statistic is defined as follows.

$$\frac{\bar{Y}^{(t)} - \bar{Y}^{(c)}}{\sqrt{var(\bar{Y}^{(t)} - \bar{Y}^{(c)})}}, \tag{3}$$

where $\bar{Y}^{(t)}$ is an unbiased estimate for the population mean in the test cell and $\bar{Y}^{(c)}$ is an unbiased estimate for the population mean in the control cell. Thus $\bar{Y}^{(t)} - \bar{Y}^{(c)}$ is an unbiased estimate for the effect size. In controlled experiments, variance reduction is about reducing $var(\bar{Y}^{(t)} - \bar{Y}^{(c)})$. The sampling in our controlled experiments is without replacement because a user can not be assigned to two cells at the same time and the population is finite. Thus, strictly speaking, the samples in control and test are not independent from each other. But the users assigned to a single experiment are typically a small proportion of the Netflix user base. Hence the dependence is negligible and we have

$$var(\bar{Y}^{(t)} - \bar{Y}^{(c)}) = var(\bar{Y}^{(t)}) + var(\bar{Y}^{(c)}). \tag{4}$$

Equation (4) shows the equivalence between reducing the variance of the mean estimate in a single cell and reducing the variance of the effect size estimate. Therefore we focus the discussion that follows on variance reduction in a single cell. Fundamentally, variance reduction in a single cell of controlled experiments can be achieved by leveraging covariates that are measurable prior to the experiments and are correlated with the business metrics. Covariates can be used at different stages of an experiment. When used at-assignment, the covariates are leveraged during the process of assigning users to the cells, e.g., stratified sampling. When used post-assignment, the covariates are leveraged after the user assignment, e.g., post stratification and CUPED.

3.3 Stratified Sampling

Stratified sampling [2] is probably the most well-known at-assignment variance reduction technique. The basic idea of stratified sampling is to divide the population into strata, sample from each stratum independently, and then combine samples across each stratum to give an overall estimate. In stratified sampling, the sample size from the kth stratum n_k is fixed for given total sample size n and they have the following relationship

$$n_k = np_k, \tag{5}$$

where p_k is defined in 3.1 and $k = 1, ..., K$. In stratified sampling, the weighted average in (2) is typically used to estimate the population mean μ. As mentioned in Section 3.1, under stratified sampling, the two estimates in (1) and (2) are the same shown as follows.

$$\begin{aligned}\sum_{k=1}^{K} p_k \bar{Y}_k &= \sum_{k=1}^{K} p_k \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} \\ &= \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.\end{aligned} \tag{6}$$

The first equation in (6) follows from the definition of $\bar{Y}_k$. The second equation is true because of (5). Now we derive some statistical properties of the estimate in (2) under stratified sampling. We first show the estimate in (2) is unbiased under stratified sampling.

$$E_{strat}(\hat{Y}_{strat}) = \sum_{k=1}^{K} p_k E_{strat}(\bar{Y}_k) = \sum_{k=1}^{K} p_k \mu_k = \mu. \tag{7}$$

Secondly, the variance of the estimate in (2) under stratified sampling is

$$\begin{aligned}var_{strat}(\hat{Y}_{strat}) &= \sum_{k=1}^{K} p_k^2 var_{strat}(\bar{Y}_k) \\ &= \sum_{k=1}^{K} \frac{n_k^2}{n^2} \frac{1}{n_k} \sigma_k^2 \\ &= \frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2.\end{aligned} \tag{8}$$

The first equation in (8) holds because sampling from the K strata is done independently from each other. σ_k^2 and n_k are defined in Section 3.1.

In simple random sampling, the standard simple sample average in (1) is used to estimate the population mean. Under simple random sampling, the estimate in (1) is unbiased shown as follows.

$$\begin{aligned}E_{srs}(\bar{Y}) &= E_{srs}(\frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}) \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} E_{srs}(Y_{kj}) \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \mu \\ &= \frac{1}{n} n\mu \\ &= \mu.\end{aligned} \tag{9}$$

The variance of (1) under simple random sampling is derived as follows.

$$\begin{aligned}var_{srs}(\bar{Y}) &= var_{srs}(\frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}) \\ &= \frac{1}{n^2} \sum_{k=1}^{K} \sum_{j=1}^{n_k} var_{srs}(Y_{kj}) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{1}{n} \sigma^2.\end{aligned} \tag{10}$$

Note that $var_{srs}(Y_{kj}) = \sigma^2$ because Y_{kj} are all random samples under simple random sampling from the distribution of Y. Next we make a connection between (8) and (10). First let Z denote the stratum number of a random observation from the distribution of Y under simple random sampling. Note that Z is a multinomial random variable that takes values $1, ..., K$ and $P(Z = k) = p_k$. Then we have

$$\begin{aligned}var_{srs}(Y) &= E_{srs}(var_{srs}(Y|Z)) + var_{srs}(E_{srs}(Y|Z)) \\ &= E_{srs}(\sum_{k=1}^{K} \sigma_k^2 I(Z = k)) + var_{srs}(\sum_{k=1}^{K} \mu_k I(Z = k)) \\ &= \sum_{k=1}^{K} \sigma_k^2 E_{srs}(I(Z = k)) + E_{srs}(\sum_{k=1}^{K} \mu_k I(Z = k))^2 \\ &\quad - (E_{srs}(\sum_{k=1}^{K} \mu_k I(Z = k)))^2 \\ &= \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} \mu_k^2 p_k - \mu^2 \\ &= \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} p_k (\mu_k - \mu)^2,\end{aligned} \tag{11}$$

where $I(Z = k)$ is an indicator variable with value 1 if $Z = k$ and 0 otherwise. Combing (10) and (11), we have

$$var_{srs}(\bar{Y}) = \frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2 + \frac{1}{n} \sum_{k=1}^{K} p_k (\mu_k - \mu)^2. \tag{12}$$

To summarize the comparison between stratified sampling and simple random sampling, estimates in both sampling techniques are unbiased. But the variance of the estimate in stratified sampling is smaller than that in simple random sampling by $\frac{1}{n} \sum_{k=1}^{K} p_k (\mu_k - \mu)^2$. The intuition for variance reduction based on stratified sampling is that the variance of the estimate based on simple random sampling can be decomposed into within-strata variance and between-strata variance. Stratified sampling achieves variance reduction by removing the between-strata variance. Fundamentally, this is because the mean of the business metric is different across strata. From a sampling point of view, stratified sampling removes the variation of sample size from each stratum for a given total sample size n and thus reduces the variance of the estimate.

3.4 Post Stratification

Post stratification is a popular post-assignment variance reduction technique. It assumes simple random sampling but uses the estimate in (2) instead of (1). Note that, when

simple random sampling is used, the estimates in (2) and (1) are different. This is because the sample size n_k from the kth stratum is not necessarily equal to np_k under simple random sampling. Here n_k, n, and p_k are defined in Section 3.1. In fact, $n_1, ..., n_K$ are all random under simple random sampling. The intuition behind post stratification is very simple. The weighted average (2) gives more weights to observations from the strata that are under-represented in the sample. Thus if a sample is badly balanced for some covariate such as signup country, the weighted average estimate automatically corrects for it. We now sketch the derivation of the variance of the estimate in (2) under simple random sampling.

$$\begin{aligned} var_{srs}(\hat{Y}_{strat}) &= E_{srs}(var_{srs}(\hat{Y}_{strat}|n_1, ..., n_K)) \\ &+ var_{srs}(E_{srs}(\hat{Y}_{strat}|n_1, ...n_K)) \\ &= E_{srs}(\sum_{k=1}^{K} p_k^2 var_{srs}(\bar{Y}_k|n_k)) + var_{srs}(\sum_{k=1}^{K} p_k\mu_k) \\ &= E_{srs}(\sum_{k=1}^{K} p_k^2 \frac{1}{n_k}\sigma_k^2) + var_{srs}(\mu) \\ &= \sum_{k=1}^{K} p_k^2\sigma_k^2 E_{srs}(\frac{1}{n_k}) \end{aligned} \tag{13}$$

What is remaining to calculate the variance of the estimate in (2) under simple random sampling is to calculate $E_{srs}(\frac{1}{n_k})$, where $k = 1, ..., K$. Note that, n_k is a Bernoullian random variable with expected value np_k for given sample size n. It can be shown that $E_{srs}(\frac{1}{n_k}) = \frac{1}{np_k} + \frac{1-p_k}{n^2p_k^2} + o(\frac{1}{n^2})$ [9], where $o(\frac{1}{n^2})$ is a residual term that converges to 0 faster than $\frac{1}{n^2}$ as $n \to \infty$. The proof in [9] is based on some complicated factorial expansions because it is for more general cases than the reciprocal of a Bernoullian random variable. In this paper, we provide a simpler proof based on Taylor expansion as follows.

$$\begin{aligned} E_{srs}(\frac{1}{n_k}) &= E_{srs}(\frac{1}{np_k} + (-\frac{1}{n^2p_k^2})(n_k - np_k) \\ &+ \frac{1}{n^3p_k^3}(n_k - np_k)^2) + o(\frac{1}{n^2}) \\ &= \frac{1}{np_k} + \frac{1}{n^3p_k^3}E_{srs}(n_k - np_k)^2 + o(\frac{1}{n^2}) \\ &= \frac{1}{np_k} + \frac{1}{n^3p_k^3}np_k(1-p_k) + o(\frac{1}{n^2}) \\ &= \frac{1}{np_k} + \frac{1}{n^2p_k^2}(1-p_k) + o(\frac{1}{n^2}), \end{aligned} \tag{14}$$

where the first equation is simply a Taylor expansion of $\frac{1}{n_k}$ at $\frac{1}{np_k}$ and the other equations follow from the fact that n_k is Bernoullian variable with mean np_k and variance $np_k(1-p_k)$. Thus, we have

$$var_{srs}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1-p_k)\sigma_k^2 + o(\frac{1}{n^2}). \tag{15}$$

Since p_k's, μ_k's, K, and μ are finite values, we can always find a large enough n such that the following is true.

$$\frac{1}{n^2p_k^2}(1-p_k) + o(\frac{1}{n^2}) \leq \frac{1}{n}\sum_{k=1}^{K} p_k(\mu_k - \mu)^2. \tag{16}$$

When equation (16) is true, we have $var_{srs}(\hat{Y}_{strat}) \leq var_{srs}(\bar{Y})$. This shows that post stratification leads to variance reduction for large enough sample size. Hence for large enough n, the comparison of variance of the estimates based on simple random sampling, stratified sampling, and post stratification can be summarized as follows.

$$\begin{aligned} &var_{strat}(\hat{Y}_{strat}) = var_{srs}(\hat{Y}_{strat}) + O(\frac{1}{n^2}) = var_{srs}(\bar{Y}) + O(\frac{1}{n}), \\ &var_{strat}(\hat{Y}_{strat}) \leq var_{srs}(\hat{Y}_{strat}) \leq var_{srs}(\bar{Y}). \end{aligned} \tag{17}$$

Thus, although it is true that the variance of the estimate based on stratified sampling is the smallest, when n is large, the variance difference between post stratification and stratified sampling will be much smaller than that between simple random sampling and stratified sampling. This means that post stratification achieves similar variance reduction as stratified sampling when the sample size n is large. It is worth pointing out the derivation of the variance in post stratification requires a regularity condition that none of the n_k's is zero [7]. Although the derivation of variance is only of theoretical interest, in practice, we need a mechanism to estimate the mean for those strata that do not have any observation in a cell. One way is to pool or collapse similar strata [7]. This is potentially an issue for post stratification in practice.

3.5 CUPED

Another variance reduction technique is based on control variates. It has been used in Monte Carlo simulation [5]. One can think of the control variates here as covariates defined in Section 3.1. The control variates technique was applied to controlled experiments as a variance reduction technique in [3]. The authors name the technique CUPED (controlled experiments utilizing pre-experiment data) in their paper because the control variates in their paper are based on pre-experiment data. CUPED is also a post-assignment variance reduction technique because it is based on simple random sampling. Next we briefly review how CUPED works. Suppose the pre-experiment data X is a one-dimensional control variate. In CUPED, instead of looking at the business metric Y, we look at a new metric defined as

$$Y_{CUPED} = Y - \theta X, \tag{18}$$

where θ is some parameter that needs to be defined. Next we discuss how to choose θ to complete the definition of the new metric. For the variance of Y_{CUPED} under simple random sampling, we have

$$var_{srs}(Y_{CUPED}) = var_{srs}(Y) + \theta^2 var_{srs}(X) - 2\theta cov_{srs}(X, Y), \tag{19}$$

where $cov_{srs}(X, Y)$ is the covariance between X and Y under simple random sampling. Using simple calculus, we can show that $var_{srs}(Y_{CUPED})$ is minimized by choosing θ equal to $cov_{srs}(X, Y)/var_{srs}(X)$, where the minimal value is

$$var_{srs}(Y_{CUPED})_{min} = var_{srs}(Y)(1 - \rho^2), \tag{20}$$

where $\rho = corr_{srs}(X, Y)$ is the Pearson correlation between X and Y under simple random sampling. The intuition be-

hind variance reduction using control variates is that the total variance of the business metric Y can be decomposed into two parts: the part that is caused by the variance of the control variate X, and the part that is explained by other unknown variables. By looking at the corrected metric Y_{CUPED}, we have removed the variance caused by X and thus the variance is reduced. So far, the discussion is all in the context of a single cell. It is clear that $E_{srs}(Y_{CUPED})$ is different from $E_{srs}(Y)$. In controlled experiments, we are typically interested in the difference between the means of the business metric in a test cell and the control cell. Hence the authors suggest using the same θ for different cells so that the difference between the means of the new metric Y_{CUPED} is the same as that of the original business metric Y. In practice, θ can be estimated based on pre-experiment data once we know X. Based on (20), X should be chosen to maximize the magnitude of $corr_{srs}(X, Y)$. In practice, the authors suggest using the same metric Y in the pre-experiment period because the same metric over different time periods typically correlate well. In our analysis, we take the authors' two suggestions: using the same θ across cells, and using the same business metric prior to experiment as the control variate.

There is also an interesting connection between CUPED and stratified sampling. When X is categorical, it can be mathematically shown that CUPED and stratified sampling (X is used to define strata in stratified sampling) is equivalent. For the detailed proof, please see [3].

4. NETFLIX'S IMPLEMENTATION OF STRATIFIED SAMPLING

We have learned through years of research that many factors not related to the product correlate with our business metrics. For example, the signup country of users correlates with retention. The most impactful factors are leveraged as covariates in stratified sampling to help reduce the sampling variance of business metrics. More covariates are leveraged for existing members since we know more about them. Results in Section 5 show that this extra information for existing users leads to significantly more variance reduction.

Prior to running an experiment, a target sample size is determined, and triggers for assigning users to the experiment are defined. For new users, the trigger is signing up for Netflix. For existing users, an example of trigger for a product change on the Netflix kids webpage could be a user visit to that page. The triggering rule and target sample size together decide the length of the recruitment period, which can last weeks. In this context, assigning users to cells happens in real-time when the trigger condition is satisfied.

Implementing stratified sampling in a real-time assignment scenario is rarely discussed and poses the challenge of having equal representation of the covariates in the test and control cells throughout the whole recruitment period. To address this issue, we rely on a queue system composed of one queue per experiment e and stratum s. Each queue consists of 100-slot segments. Prior to user assignment, the sampling rate for each cell in the experiment is specified. Sampling rate for a cell is defined as the share of users in the experiment that receive the experience in the cell. The cell sampling rate ranges between 0% and 100% in an increment of 1%. For each segment of 100 slots, the slots are mapped to cells such that the share of slots assigned to a cell exactly matches the sampling rate for that cell. Note that the increment cannot be more granular than 1% due to the 100-slot segment design.

Here is a simple example to illustrate the assignment in a single segment. Suppose we want to run an experiment with two cells and allocate 50% of the users to each of the two cells. We first get a sequence of integers between 1 and 100 as seen in Figure 1 (a) and then reshuffle this sequence as in Figure 1 (b). Finally we map integers 1-50 to Cell 1 and 51-100 to Cell 2 as in Figure 1 (c). As mentioned above, a queue consists of many 100-slot segments. The cell assignment in each 100-slot segment is done independently from each other within a single queue, and independently across queues. When a new user eligible for the experiment signs up, we first decide the strata that he falls into based on his covariate information and then assign him to the corresponding queue for his strata. He will then take the next available slot in the queue and gets assigned to the cell for the slot. A simple example of new user assignment with two strata and two cells is shown in Figure 2.

The implementation of stratified sampling based on our queue system does not always achieve perfect balance of strata across cells. This can diminish the amount of variance reduction based on stratified sampling. There are two factors that contribute to the imbalance.

Firstly, we only guarantee perfect balance within each segment of 100 slots. Thus the total sample size of a stratum across cells needs to be a multiple of 100 to achieve perfect balance. For example, if there are 100,090 users in a stratum prior to cell assignment, then the queue system guarantees balance for the first 100k users but not the last 90 users. For the last 90 users, the actual sampling rate in each cell may not exactly match the intended sampling rate. And thus, after cell assignment, the percent of users from each stratum may be different across cells. The rationale for having a segment size of 100 is mainly for the convenience of specifying sampling rate per cell (increment of 1%). Potentially we can decrease the segment size to achieve better balance but it also makes the sampling rate specification less granular, e.g., if we change the segment size to 50, then the sampling rate needs to be in increment of 2%. The impact of this imbalance depends on the sample size in each stratum.

The second factor preventing us from achieving perfect balance is that we usually have to use many machines to conduct the sampling because of both high volume of sampling requests and occasional failures of the machines. With M machines, there will be M queues for experiment e and stratum s. When a user eligible for an experiment signs up, he is first randomly assigned to a machine, and then assigned to a queue on the machine based on his covariate information, and finally he takes the next available slot in the queue and gets assigned to the cell for the slot. It is intuitive that with multiple machines, it is more difficult to achieve the strata balance across cells. For example, if the sample size of a stratum for the whole experiment is 100k, it would achieve perfect balance with a single machine but not necessarily with multiple machines because the number of users from the stratum on each single machine is not necessarily a multiple of 100. The likelihood to achieve perfect balance decreases as the number of machines increases. The impact of the number of machines on the variance reduction amount based on stratified sampling is quantified in the next section.

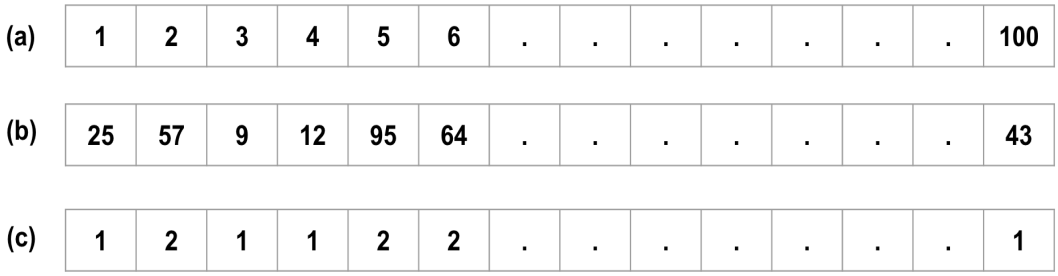

Figure 1: Illustration of Cell Assignment in One Segment: (a) generate a sequence of integers between 1 and 100, (b) random shuffling of the sequence of integers, (c) mapping of integers to cell

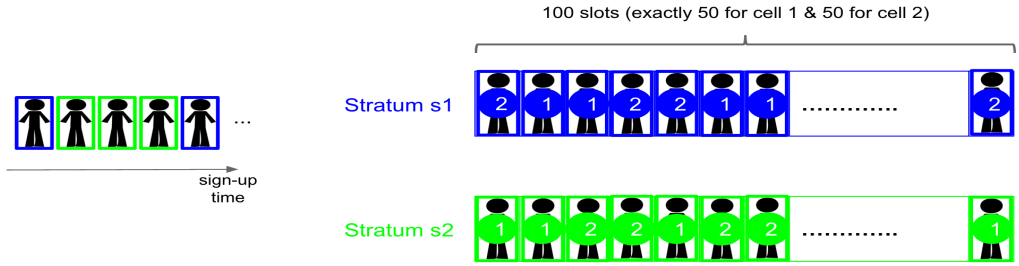

Figure 2: Illustration of stratified sampling with one machine for new users

5. EMPIRICAL EVALUATION

5.1 Evaluation Methodology

In this section, we compare the amount of variance reduction achieved by stratified sampling, post stratification and CUPED on a few datasets from Netflix. The business metrics considered are customer retention and seven out of the I streaming thresholds defined in Section 2.2. Simple random sampling is used as the baseline to estimate the amount of variance reduction achieved by each technique. The comparison is done on both new and existing users. For each user type, we collect covariates and business metrics for a cohort of users. We define these users as the population and repeatedly simulate A/A experiments, which are controlled experiments with two cells and zero effect size. In the case of stratified sampling, an A/A experiment is simulated by splitting users into two cells based on the implementation described in Section 4. For post stratification and CUPED, an A/A experiment is simulated by splitting users into two cells based on simple random sampling. After the user assignment to two cells in a single A/A experiment is determined, for each business metric, we compute an estimate of the effect size. For the simple random sampling baseline, this estimate is the difference of the simple sample averages in (1). For stratified sampling and post stratification, this estimate is the difference of the weighted averages in (2). For CUPED, this estimate is the difference of the averages for the corrected metric in (18). For each business metric and variance reduction technique, 100k A/A experiments are simulated independently from each other on the same cohort of users, yielding 100k estimates of the effect size. The sample variance of these estimates is then compared with the theoretical variance based on simple random sampling to quantify the amount of variance reduction for each metric and technique combination. For stratified sampling, we also estimated the variance reduction percentage pretending there is only one machine to get a sense of the additional variance introduced by the use of multiple machines. For new users, we do not have pre-experiment data for streaming and retention that can be used for CUPED. Thus, eight regression models (one per metric) were built on a different set of users from those used to simulate our A/A experiments. The predictors in the regression models are the same set of covariates used to define strata in stratified sampling. This ensures fair comparison between CUPED and stratified sampling. The predicted mean values of the metrics from the models are then used as pre-experiment data. For existing users, the same metric is used as the pre-experiment data for the streaming thresholds. We did not apply CUPED on retention for existing users since the amount of variance reduction for retention is very small based on the other techniques and we do not expect CUPED to make a significant difference. We report on the variance reduction point estimates along with error bars based on Bootstrap [4] as a measure of the uncertainty of the results based on a finite (although 100k is already pretty large) number of simulations.

5.2 Results

The resulting variance reduction estimates are presented in Figures 3 and 4, separately for new and existing users for each of the eight business metrics (retention and seven streaming thresholds). The results show that identifying covariates that are highly correlated with the business metrics is key for the success of any of these variance reduction techniques. The empirical results also align well with the theory in Section 3. Indeed, ignoring challenges posed by practical implementation, stratified sampling, post stratification and

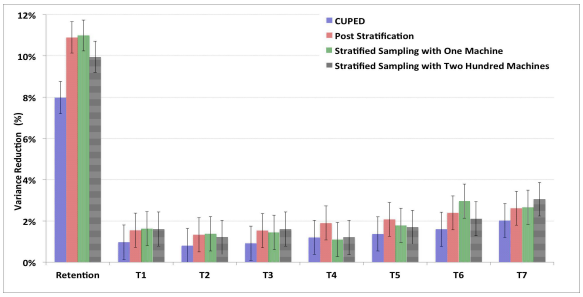

Figure 3: New users. Variance reduction results of stratified sampling, CUPED, and post stratification compared to simple random sampling

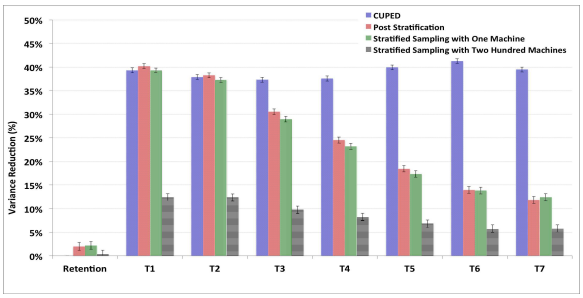

Figure 4: Existing users. Variance reduction results of stratified sampling, CUPED, and post stratification compared to simple random sampling

CUPED achieve similar variance reduction amount when leveraging the same covariates. However, in practice, the variance reduction achieved by stratified sampling can be severely impacted by the 100-slot design and the need to use multiple machines as described in Section 4. This is not the case for post stratification and CUPED which are post-assignment techniques. See the following subsections for more detailed findings.

5.2.1 *Influence of Covariates*

For new users, the amount of variance reduction achieved is very low regardless of the metric or the variance reduction technique used. This is due to the lack of covariates highly correlated with the business metrics for these users at the time of cell assignment. Indeed, the Pearson correlation between the covariates and business metrics ranges from 0.2 to 0.4 for new users.

For streaming thresholds, the variance reduction for existing users can be up to 40% because we included pre-experiment streaming activity as a stratification dimension or in the post-assignment correction. Note that for existing users, the lowest streaming threshold T_1 prior to the experiment is the only streaming threshold used as covariate in stratified sampling and post stratification. Thus it is expected that the amount of variance reduction becomes lesser as the streaming threshold moves further away from T_1 and the correlation between this covariate and these streaming thresholds becomes weaker.

For retention, while the amount of variance reduction is small for both new and existing users, it is higher for new users. The reason is because the covariates used to define strata for new users are more correlated with retention than for existing users. Also new users get a free trial in the first month and there is more user-level variation of retention than for existing users who have already passed the free trial period and whose retention metric becomes less sensitive.

5.2.2 *Post Stratification Observations*

For all the metrics and both user types, post stratification is comparable to stratified sampling with one machine. This is consistent with the theoretical understanding that post stratification achieves similar variance reduction as stratified sampling when the sample size is large. The sample size in the dataset used for evaluation is at the scale of hundreds of thousands.

5.2.3 *Stratified Sampling Observations*

As discussed in Section 3.3, from a sampling point of view, stratified sampling achieves variance reduction compared to simple random sampling because it removes the variation of the sample size from each stratum once the total sample size n is given. In Section 4, we described how the practical implementation of stratified sampling cannot completely remove this variation because of the use of a 100-slot queue system and the need to conduct sampling on multiple machines. The empirical evaluation shows that for existing users, variance reduction based on stratified sampling with multiple machines is less than half of that with one machine. This impact is not as clear for new users for whom the number of machines used for sampling is only one fifth of that for existing users. Also the increase in the number of machines tend to have a larger impact on existing users that have smaller strata sizes.

To provide some intuition around the impact of multiple machines, we run an evaluation procedure similar to the one described in Section 5.1 on a cohort of existing users. In this evaluation, we show the impact of number of machines on the sampling variance of the sample sizes of each stratum. Since there is more than one stratum, we define a weighted average of standard deviation metric as follows

$$\sum_{k=1}^{K} p_k \sigma_k, \tag{21}$$

where p_k is the proportion of users from stratum k in the population and σ_k is the standard deviation of the sample size from stratum k in a random sample for fixed total sample size n. The number of machines used for stratified sampling is varied from one to two hundred. For a given number of machines m, we simulate 100k A/A experiments to estimate σ_k. Each A/A experiment first randomly assigns users to machines and then for each machine splits users into two cells based on stratified sampling as described in Section 5.1. The estimates of σ_k are then plugged into (21) to get the estimate of the weighted average of standard deviation metric. The results are shown in Figure 5. Note that the variation of the weighted standard deviation metric monotonically increases as the number of machines increases. The error bars are calculated using the Bootstrap technique [4]. There is no error bar for simple random sampling because it is the theoretical value. So, as the number of machines increases, the variation of the sample size from each stratum increases, which translates to higher variance of the final es-

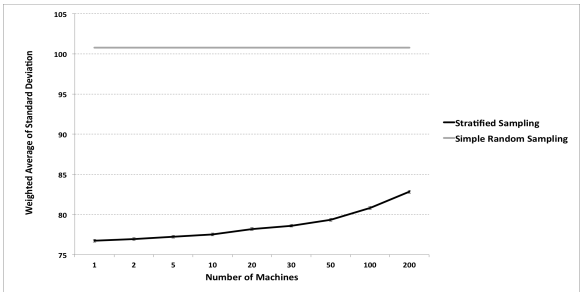

Figure 5: Impact of the number of machines used in stratified sampling on strata sample size variation

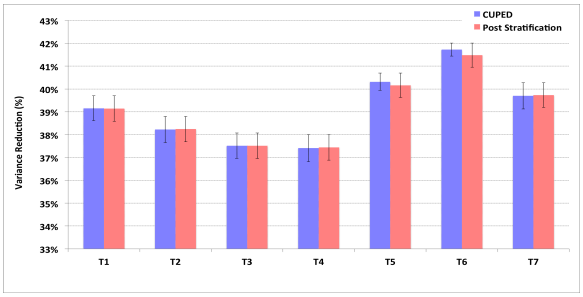

Figure 6: Existing users. Variance reduction of CUPED and post stratification using the same covariates compared to simple random sampling

timate based on stratified sampling and diminished variance reduction based on stratified sampling.

5.2.4 CUPED Observations

CUPED performs slightly worse than stratified sampling and post stratification for new users because the pre-experiment data in CUPED is essentially a one-dimensional summary of the covariates used to define strata and there is some loss of information in this summary.

For existing users, we have the flexibility of correcting streaming thresholds with the same metric prior to the experiment. Hence the amount of variance reduction is consistently around 40% for all the streaming thresholds. Finally, it is worth emphasizing that the difference in variance reduction between CUPED and the other techniques for high streaming thresholds on existing users in Figure 4 is due to the difference of covariates used in the techniques, not the techniques themselves. The comparison settings were set as such because the initial objective of this case study was to compare CUPED with what was used in production at Netflix. We also did a fair comparison between CUPED and post stratification by using exactly the same covariates for each of the streaming threshold metrics on existing users. The results in Figure 6 clearly show that these two techniques perform comparably when using the same covariates. This is expected due to the following two reasons.

- Post stratification achieves similar variance reduction as stratified sampling when the sample size is large, which is true here.
- Stratified sampling is equivalent to CUPED when the covariate is categorical, which is also true here since the covariates are binary streaming thresholds. For more on this point, please see the appendix of [3].

Comparisons with stratified sampling based on the same covariates are omitted since we expect stratified sampling to perform the same as post stratification given the large sample size in the data.

6. CONCLUSIONS

For companies that use controlled experiments to make product decisions, it is critical to run highly sensitive controlled experiments in order to not miss product changes that can have a substantial impact on user experience and revenue. In this paper, we focused on improving the sensitivity of controlled experiments by reducing sampling variance of the business metrics. We compared a few variance reduction techniques, both at-assignment (stratified sampling) and post-assignment (post stratification & CUPED). We showed that theoretically, these techniques achieve similar variance reduction when the sample size is large, which is typically the case in online controlled experiments. We applied them to a few Netflix datasets and our empirical results aligned with theory when the same set of covariates is used for all the techniques. However, in practice, stratified sampling performs worse than post-assignment techniques such as post stratification and CUPED because real-time experimental assignment requires a queue system and the use of multiple machines. Moreover, post-assignment techniques are cheaper to implement, and very flexible in choosing the covariates for post-assignment correction. It is thus recommended to apply post-assignment variance reduction techniques when running large-scale controlled experiments. Our results on new users emphasize that identifying covariates that are highly correlated with the business metrics is key for the success of variance reduction techniques. In cases where such covariates cannot be easily found, other methods to improve the sensitivity of controlled experiments should be explored. At Netflix we continuously research new user engagement metrics that are a better tradeoff between sensitivity and correlation with retention to help make better production decisions. In the context of TV/movie recommendation and search algorithms, we leverage offline experiments as a pre-selection mechanism to reduce the number of test cells in a single experiment run online. With fewer test cells , larger sample sizes can be used for each cell and sensitivity is thus improved. See [6] for a more complete description of the offline experiments method as well as its challenges. Another example is leveraging interleaving-based experiments [1] to remove the between-user variance associated with our traditional controlled experiment design. Finally, more sophisticated experimental designs such as fractional factorial designs [10] can be used to reduce the number of test cells when multiple changes are tested in one single experiment.

7. ACKNOWLEDGEMENTS

We are grateful to Bryan Gumm and Carlos Gomez-Uribe for designing the queue system for stratified sampling at Netflix. We would like to give many thanks to Carlos Gomez-Uribe and Harald Steck for their thoughtful comments on earlier drafts of this paper.

8. REFERENCES

[1] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. on Information Systems*, 30(1), February 2012.

[2] W. G. Cochran. *Sampling Techniques*. Wiley, 1977.

[3] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM 13 Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 123–132, 2013.

[4] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[5] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[6] C. A. Gomez-Uribe and N. Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. on Management Information Systems*, 6(4), December 2015.

[7] D. Holt and T. Smith. Post stratification. *J.R. Statist. Soc. A*, 142(1):33–46, 1979.

[8] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2010.

[9] F. F. Stephan. The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. *The Annals of Mathematical Statistics*, 16(1):50–61, March 1945.

[10] C. J. Wu and M. S. Hamada. *Experiments: Planning, Analysis, and Optimization*. Wiley, 2009.