# Value of Stratification in Cluster-Randomized Experiments

**Stefan Hut** Amazon.com
**Blake Mason** Amazon.com
**Mahnaz Islam** Amazon.com
**Lledo Esquerra** Amazon.com

## CODE 2023: Extended Abstract

There are many experimental settings that may suffer from cross-unit (customers, seller, advertiser, etc.) spillovers, for instance through network effects. Such effects introduce bias and prevent the experimenter from drawing trustworthy insights on the data. One approach to dealing with such spillovers is to group units into clusters and randomize treatment status at the cluster level. Examples of clusters are groups of advertisers or sellers, or geographic clusters (ZIP codes, Designated Marketing Areas (DMAs)) that group customers. While clustering helps to reduce bias in the form of spill-over effects, on its own, it can present other challenges. Clusters of advertisers, sellers, or customers are often highly skewed in size and other characteristics (Chung and Lu (2006)). Naively randomizing treatment or control allocations at the cluster level can lead to unchecked experiment imbalance if, for example, the largest clusters are all allocated to either treatment or control.

In this paper we assess the value of using a stratified randomization approach in the context of cluster-randomized experiments and demonstrate how doing so can improve statistical power while also reducing imbalance in the treatment allocation compared to traditional randomization schemes. In a stratified experiment, clusters of individual units (e.g., a single advertiser) are stratified (i.e., grouped or blocked) according to the values of a set of pre-experiment covariates. Clusters are then randomized within each of the strata, with the goal of helping to mitigate pre-period covariate imbalances between treated and control clusters. By randomizing at the strata level, the allocations for any two different strata are pairwise independent. Ultimately, this improves statistical power in the post-experiment analysis.

We begin by demonstrating the theoretical precision gains, measured through the change in the standard error of the treatment effect estimate, from stratification. Consider two estimators of the population treatment effect: (1) a simple difference estimator, (2) a stratified estimator. For the latter, the mean difference is computed within each stratum, after which we construct the overall mean difference estimator as a weighted average of the strata-level estimates, where the weights represent the relative sample size in each stratum. Following, Miratrix and Yu (2013), formally, the difference in variance between these two estimators is given by:

$$var(\hat{\tau}_{sd}) - var(\hat{\tau}_{strat}) = \frac{1}{n}\left\{\bar{\sigma}^2(1) + \bar{\sigma}^2(0) + 2\bar{\gamma}\right\} - \frac{1}{n^2}\sum_k \frac{n - n_k}{n-1}\left\{\sigma_k^2(1) + \sigma_k^2(0) + 2\gamma_k\right\} \tag{1}$$

where $\sigma^2(l)$ are the variances of the observed outcomes for each group $l = \{0, 1\}$, $\gamma$ represents the covariance of the estimates, and $\gamma_k$ denotes the covariance for strata $k$.

The equation has two parts. The first term is the between-strata variation (denoted by the upper bar). It measures how much the mean potential outcomes vary across strata and captures how well the stratification variable separates out different units, on average. The larger the separation, the more there is to gain from stratification. The second term is the within-strata variation and captures a penalty from stratification which is due to a loss in efficiency from estimating the treatment effect within each stratum. If the between-strata variation is larger than the cost paid, then the difference in variance is positive and it is good to stratify. In particular, if the first term is non-zero, then it will dominate for sufficiently large $n$. In addition, the equation tells us that stratifying on variables that relate to the outcome is likely to result in larger between-strata variation and thus a larger reduction in variance compared to a simple difference estimator. Note lastly, that having more strata is not necessarily better. As $K$ increases we have more terms in the sum and hence a greater potential penalty from stratification.

We use this framework to estimate the precision gains from stratification empirically by comparing precision in the post-experiment analysis from the stratified randomization approach to a simple randomization. In the latter case, we randomize each cluster to treatment or control without first grouping the clusters into strata. In practice, the estimated

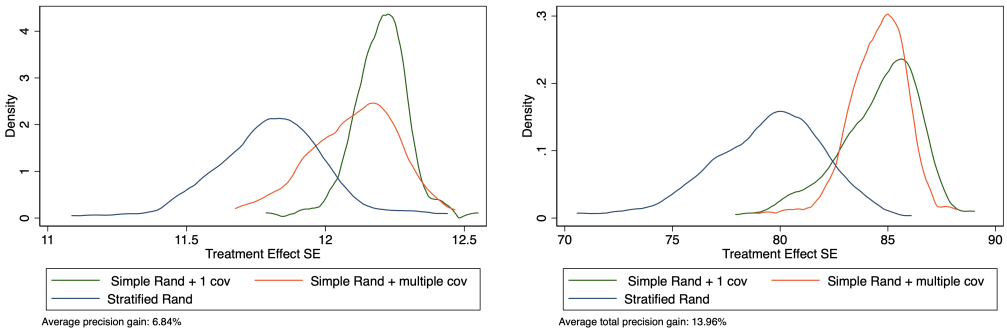


Figure 1: Distribution of estimated $SE(\hat{\beta})$ across 100 simulations - Two Advertiser Metrics

treatment effect in a given experiment is a random variable due to the randomness in both the random allocations to treatment and control and the randomness in the observations conditioned on these allocations. To deal with this, we perform randomization many times to estimate the average standard error. We conduct an A/A test where we randomly split our data into treatment and control groups though there is no actual treatment. Because of this, the true treatment effect is known to be zero. Simulating this 1000 times, we capture the variance reduction from stratification vs. simple randomization.

In each simulation round, we analyze the experiment (A/A test) using the re-randomized treatment allocation. Let $Z_i$ be a vector of covariates $X_i$, and strata fixed effects $1(Strata_i = s)$ for each stratum $s \in [1, S]$:

$$Z_i = [X_i^\top, 1(Strata_i = 1), \cdots, 1(Strata_i = S)]^\top$$

Let $\dot{Z}_i$ denote the demeaned covariate vector. We run the following regression:

$$y_i = \theta_0 + \beta \cdot T_i + \theta_1^\top \cdot \dot{Z}_i + \theta_2^\top \cdot \dot{Z}_i \cdot T_i + \epsilon_i \tag{2}$$

The outcome variable $y_i$ denotes a metric of interest for a given unit $i$ in the experiment. We cluster standard errors at the cluster level in order to account for the cluster-randomization. For the simple randomization approach we omit strata fixed effects from the analysis. In each simulation round we capture the Standard Error (SE) of the treatment effect, $\hat{\beta}$ both for simple and stratified randomization. We plot the distribution of standard errors and compute the average precision gain $(precision = 1/SE^2)$ across the 100 simulations. We plot variances from these 100 simulations for three approaches: (1) simple randomization and including only one covariate (the pre-experiment version of the outcome metric; (2) simple randomization and including multiple covariates (the pre-experiment outcome metric and several additional pre-experiment advertiser metrics); (3) stratified randomization. The comparison across these tells us how much variance reduction we obtain through stratification versus including a richer set of covariates in a linear regression.

We apply this to data from a completed cluster-randomized experiment run on advertisers at Amazon. We stratified advertiser clusters using three stratification variables: advertiser count per cluster, and two advertising revenue related metrics. For advertiser count, we stratify into four blocks: (1) 0-50th percentile, (2) 50th-95th percentile, (3) 95th-99th percentile, (4) 99th+ percentile. The goal is mainly to ensure that outlier clusters in terms of advertiser count (in blocks 3 and 4) are equally distributed between treatment and control. For the other two metrics, we stratify into four additional blocks by quantile. Altogether, we have $4^3 = 64$ blocks within which we randomize.

Figure 1 shows the distribution of treatment effect standard errors (SEs) across the 100 simulations for the three approaches. We see clear gains from stratification: the distribution of SEs is shifted to the left of the SEs from either simple randomization approach. On average, we observe precision gains of about 7-14%, depending on the metric.

We further compare the effect of increasing the number of strata we use on the distribution of standard errors of the treatment effect in Figure 2. Shown in blue, we ran the standard stratified randomizer 1000 times with different random seeds and 80 strata. For each of the 1000 runs, we computed an estimated treatment effect and the standard error of this estimate. Next, shown in orange, we re-ran stratification 100 times (due to computational constraints) using approximately 10 times as many strata (roughly 800). The results shown in Figure 2 demonstrate a 23.5% reduction in the standard error on average or equivalently a 29% precision increase. This demonstrates that precision gains from

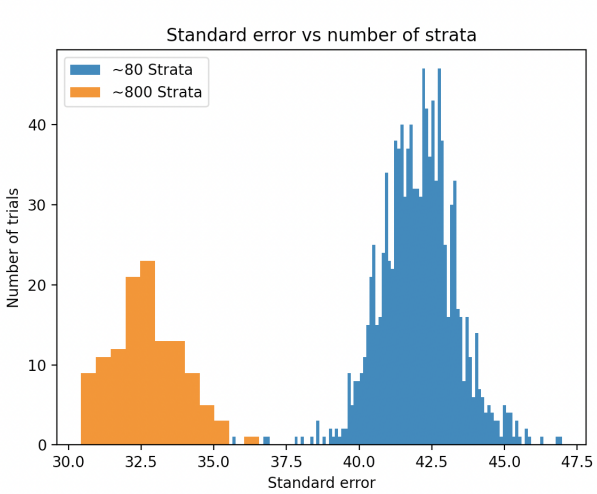


Figure 2: Distribution of estimated $SE(\hat{\beta})$ across 1000 simulations - Advertiser Metric

stratification are highly dependent on the number of strata chosen. In follow-up research, we will investigate at what point we reach diminishing returns to increasing number of strata in the context of online A/B experimentation, which is suggested by theory.

Taken together, the paper shows that stratification can lead to substantial precision gains in the context of cluster-randomized experiments. In particular, we find a 7-14% gain in precision from stratifying compared to methods that do not leverage stratification, and an additional precision gain of up to 30% from adjusting the number of strata used for stratification. Such gains in precision can enable experimentation in areas that typically have low statistical power, including cluster-randomized experiments or smaller online experiments. In addition, power gains will allow experimenters to run their experiments for a shorter amount of time. To contextualize these precision gains, a 10% reduction in variance is sufficient to reduce experiment duration from 4 weeks to 3 at Amazon. The framework presented in the paper can be used to test precision gains from stratification across a wide range of experimental use cases, including non-clustered experiments on customers, sellers, or other experimental units. In addition, it can be used to compare precision gains from different stratification approaches, including, as demonstrated, different number of strata.

## References

Chung, F. and Lu, L. (2006). *Complex graphs and networks*. American Mathematical Society, NSA.

Clauset, A., Newman, M., and Moore, C. (2004). *Finding community structure in very large networks*. American Physical Society.

Dhillon, I. (2006). *Co-clustering documents and words using bipartite spectral graph partitioning*. Sixth International Conference on Data Mining (ICDM'06).

Miratrix, Luke W., J. S. S. and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75.2:369–396.

Xie, H. and Aurisset., J. (1016). *Improving the sensitivity of online controlled experiments: Case studies at netflix*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

## Appendix A: Theoretical Precision Gains

We demonstrate the theoretical precision gains from stratification. We assume for simplicity that there is only one treatment arm and ignore covariate adjustment in the analysis. We define two estimators of the population treatment effect: (1) a simple difference estimator, (2) a stratified estimator. The derivations below borrow heavily from Miratrix and Yu (2013).

### A1: Simple difference estimator

The simple difference in means estimator is given by:

$$\hat{\tau}_{sd} = \sum_{i=1}^{n} \frac{T_i}{W(1)} y_i(1) - \sum_{i=1}^{n} \frac{(1-T_i)}{W(0)} y_i(0)$$

where $W(1) = \sum_i T_i$ is the total number of treated units and $W(1) + W(0) = n$.

### Variance of estimator

The variance of this estimator is:

$$var(\hat{\tau}_{sd}) = \frac{1}{n}\left\{\sigma^2(1) + \sigma^2(0) + 2\gamma\right\}$$

where $\sigma^2(l)$ are the variances of the observed outcomes for each group $l$ and $\gamma$ represents the covariance.

### A2: Stratified estimator of sample average treatment effect

The basic idea of stratification is to divide the population into strata, and randomize each strata independently. In particular, if $p_k$ is the proportion of the population in the $k^{th}$ stratum, then let:

$$n_k = np_k$$

be the number of units within each stratum. Let the assignment split $W$ be the total number of treated units in each stratum:

$$W = (W_1(1), \ldots, W_k(1))$$

Stratification ensures that $W$ is constant because we randomize within strata, ensuring that a pre-specified number of units are treated in each. We use a simple difference in means estimator within each stratum $k$:

$$\hat{\tau}_k = \sum_{i=1}^{n_k} \frac{T_i}{W_k(1)} y_i(1) - \sum_{i=1}^{n_k} \frac{(1-T_i)}{W_k(0)} y_i(0)$$

And the overall estimator is a weighted average of the strata level estimates:

$$\hat{\tau}_{strat} = \sum_k \frac{n_k}{n} \hat{\tau}_k$$

### Variance of estimator

We express the variance of the estimator with respect to the sample's (unknown) means, variances and covariance of potential outcomes divided into between-strata variation and within-stratum variation.

The within-stratum variances and covariances are, for $k = 1, \ldots, K$:

$\sigma_k^2(l) = \frac{1}{n_k - 1} \sum_{n_k} [y_i(l) - \bar{y}_k(l)]^2$ for $l = 0, 1$

and

$\gamma_k = \frac{1}{n_k - 1} \sum_{n_k} [y_i(1) - \bar{y}_k(1)][y_i(0) - \bar{y}_k(0)]$

where $\bar{y}_k(l)$ denotes the mean of $y_i(l)$ for all units in stratum $k$.

The between-stratum variance and covariances are the weighted variances and covariance of the strata means:

$\bar{\sigma}^2(l) = \frac{1}{1-n}\sum_{k=1}^{K} n_k\{\bar{y}_k(1) - \bar{y}(1)\}^2$ for $l = 0, 1$

and

$\bar{\gamma} = \frac{1}{1-n}\sum_{k=1}^{K} n_k\{\bar{y}_k(1) - \bar{y}(1)\}\{\bar{y}_k(0) - \bar{y}(0)\}$

Given this notation, the variance of the stratified estimator is:

$$var(\hat{\tau}_{strat}) = \frac{1}{n}\sum_{k}\frac{n_k}{n}\left\{\sigma_k^2(1) + \sigma_k^2(0) + 2\gamma_k\right\}$$

Note that we can also re-write the variance of the simple difference estimator in terms of within- and between strata level parameters:

$$var(\hat{\tau}_{sd}) = \frac{1}{n}\left\{\bar{\sigma}^2(1) + \bar{\sigma}^2 + 2\bar{\gamma}\right\} + \frac{1}{n}\sum_{k}\frac{n_k - 1}{n-1}\left\{\sigma_k^2(1) + \sigma_k^2(0) + 2\gamma_k\right\}$$

**A3: Comparing variances**

Taking the difference between the above variances:

$$var(\hat{\tau}_{sd}) - var(\hat{\tau}_{strat}) = \left[\frac{1}{n}\left\{\bar{\sigma}^2(1) + \bar{\sigma}^2(0) + 2\bar{\gamma}\right\}\right] - \left[\frac{1}{n}\sum_{k}\left\{\left(\frac{n_k}{n} - \frac{n_k - 1}{n-1}\right)\sigma_k^2(1) + \left(\frac{n_k}{n} - \frac{n_k - 1}{n-1}\right)\sigma_k^2(0) + 2\left(\frac{n_k}{n} - \frac{n_k - 1}{n-1}\gamma_k\right)\right\}\right]$$

Which simplifies to:

$$var(\hat{\tau}_{sd}) - var(\hat{\tau}_{strat}) = \frac{1}{n}\left\{\bar{\sigma}^2(1) + \bar{\sigma}^2(0) + 2\bar{\gamma}\right\} - \frac{1}{n^2}\sum_{k}\frac{n - n_k}{n-1}\left\{\sigma_k^2(1) + \sigma_k^2(0) + 2\gamma_k\right\}$$