Stratification Benefits in Experimental Design: Evidence from Multiple Contexts

Melany Gualavisi melanygd@amazon.com
Stefan Hut hutstefa@amazon.com
Mahnaz Islam maislam@amazon.com

Abstract This paper examines the effectiveness of stratification in experimental design using evidence from multiple large-scale experiments. We analyze data from experiments ranging from approximately 30,000 to 180,000 units across different business contexts. Our results show that pre-stratification and post-stratification achieve virtually identical precision improvements - largest in smaller samples (10% improvement in standard errors) and moderate but positive in larger samples (5% improvement). In more homogeneous populations, these benefits decrease substantially in small samples (24-63% reduction) but remain relatively stable in large samples (only 1-5% reduction). Post-stratification offers greater flexibility by allowing adjustments for any pre-treatment variable after randomization without risking accidental bias, while pre-stratification optimizes balance for chosen variables (reducing imbalances by 54-92%) but may create imbalance in variables negatively correlated with the stratification variables. The choice between methods ultimately depends on whether the primary goal is to improve precision, maintain implementation flexibility, or achieve visible balance in specific known covariates for the experimental context.

1. Introduction Experimental design faces a fundamental challenge in achieving sufficient precision to detect treatment effects. Statistical theory offers two main approaches to address this: covariate adjustment and stratified randomization. While stratification can provide substantial precision gains, it often requires more complex implementation. This raises a critical question: when do stratification's benefits justify its implementation costs?

Stratification comes in two forms: pre-stratification, which groups similar units before randomization, and post-stratification, which incorporates grouping information after assignment. While both methods achieve equivalent precision improvements, they differ in how they handle balance. Pre-stratification provides visible balance by design—covariate differences are demonstrably smaller before any adjustment making it valuable when stakeholders require transparent allocation. Post-stratification achieves balance through statistical adjustment, offering greater flexibility: it can adjust for any pre-treatment variable ex-post, avoids accidental bias in non-stratified variables, and requires simpler infrastructure. The choice between methods ultimately depends on whether visible balance or implementation flexibility is the priority.

This paper provides empirical evidence on these trade-offs through analyses of multiple large-scale experiments. Using A/A tests and repeated simulations, we compare three approaches: simple randomization with covariate adjustment, pre-stratification, and post-stratification. We examine how stratification's value varies with sample size and population heterogeneity across different experimental contexts. This multi-context approach allows us to establish broader principles about stratification's value that extend beyond any single experimental setting.

2. Theoretical framework Following Miratrix et al. [2013], we present the key variance comparisons that guide our analysis. Let τ be the average treatment effect across K strata, with n total sample size and n_k units in stratum k. Let $\sigma^2(1)$ and $\sigma^2(0)$ denote outcome variances in treatment and control groups, with $\sigma_k^2(1)$ and $\sigma_k^2(0)$ representing their stratum-specific counterparts. Let $\bar{\sigma}^2(1)$ and $\bar{\sigma}^2(0)$ be the between-stratum variances, and $\bar{\gamma}$ and γ_k represent the between-stratum and within-stratum covariances between treatment and control outcomes. Treatment assignment proportions are denoted by β_1 and β_0 overall, β_{1k} and β_{0k} within strata, and p_k for the fixed proportion under pre-stratification.

The efficiency gains of pre-stratification over simple randomization can be decomposed as:

$$\begin{aligned}\mathrm{Var}(\hat{\tau}_{sd}) - \mathrm{Var}(\hat{\tau}_{pre}) = &\left[\frac{1}{n}\Big\{\beta_1\,\bar{\sigma}^2(1) + \beta_0\,\bar{\sigma}^2(0) + 2\,\bar{\gamma}(1,0)\Big\} + \frac{1}{n}\sum_{k=1}^{K}\frac{n_k-1}{n-1}\Big\{\beta_1\,\sigma_k^2(1) + \beta_0\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\Big\}\right] \\ &- \left[\frac{1}{n}\sum_{k=1}^{K}\frac{n_k}{n}\Big\{\tfrac{1-p_k}{p_k}\,\sigma_k^2(1) + \tfrac{p_k}{1-p_k}\,\sigma_k^2(0) + 2\,\gamma_k(1,0)\Big\}\right].\end{aligned} \tag{1}$$

The first term represents potential gains from between-strata variation - when strata effectively separate units with different outcomes. The second term captures possible precision losses from within-strata random imbalances, which increase as strata become smaller.

The difference between post- and pre-stratification variances is:

$$var(\hat{\tau}_{ps}) - var(\hat{\tau}_{pre}) = \frac{1}{n}\sum_{k}\frac{n_k}{n}\{(\beta_{1k} - \frac{1-p_k}{p_k})\sigma_k^2(1) + (\beta_{0k} - \frac{p_k}{1-p_k})\sigma_k^2(0)\} \tag{2}$$

This difference reflects that post-stratification inherits random treatment imbalances within strata, captured by the deviation of realized assignment proportions (β_{1k}, β_{0k}) from their fixed design targets $(p_k, 1-p_k)$. These random imbalances inflate variance in finite samples but diminish at rate $1/n$ as sample size grows.

This framework generates three testable predictions: (1) Both pre- and post-stratification improve precision when strata capture systematic outcome variation (i.e., when stratifying variables are predictive of outcomes). (2) The efficiency gap between post- and pre-stratification, driven by random treatment imbalances within strata, diminishes at rate $1/n$ as sample size increases. (3) For pre-stratification, precision gains shrink as the number or granularity of strata increases, since very small strata inflate sampling variance and may offset the benefits of balance, see Miratrix et al. [2013].

3. Methodological Approach Our methodology compares three approaches: simple randomization, pre-stratification, and post-stratification. We construct strata using three key characteristics in each experimental context, yielding between 48 and 64 strata. To examine heterogeneity effects, we create "less heterogeneous" versions of each dataset by reducing extreme unit sizes while maintaining the overall structure.

We evaluate methods through A/A tests with $N = 100$ iterations of random treatment assignment, generating distributions of treatment effects $\{\hat{\tau}_0^n\}_{n=1}^N$ and standard errors $\{\hat{\sigma}^n\}_{n=1}^N$. For precision evaluation, we use two metrics: direct precision measured as $\text{Precision}_n = \frac{1}{(\hat{\sigma}^n)^2}$ and precision ratio = $(1/SE$ with simple randomization$)/(1/SE$ with stratification$)$. For balance assessment, we examine treatment-control differences $\Delta_k^X = \bar{X}k^T - \bar{X}k^C$ through mean absolute differences $E[|\Delta_k|]$, and balance improvement ratio = $\frac{E[|\Delta k,\text{simple}^X|]}{E[|\Delta k,\text{strat}^X|]}$.

4. Main Findings Our analysis reveals consistent patterns about stratification's effectiveness across experimental contexts. Using one experimental setting as illustration, we demonstrate three key findings while noting similar patterns hold in our other experiments.

First, both pre- and post-stratification achieve virtually identical precision improvements over simple randomization (Figure 1). These gains are largest in smaller samples (10% improvement in standard errors) and moderate but positive in larger samples (5% improvement).

Second, pre-stratification provides important balance benefits through its controlled treatment assignment (Figure 2). Pre-stratification reduces covariate imbalances by 54-92%, with these improvements extending beyond variables directly used in stratification.

Third, sample size reveals distinct patterns for precision and balance (Figure 3). For precision, gains diminish from 10%-11.5x in small samples (2k-30k) to 5%-4.0x in large samples (115k 187k), with pre- and post-stratification achieving nearly identical performance at scale. For balance, pre-stratification maintains stable benefits across all sample sizes, consistently reducing covariate imbalances by 54-92%.

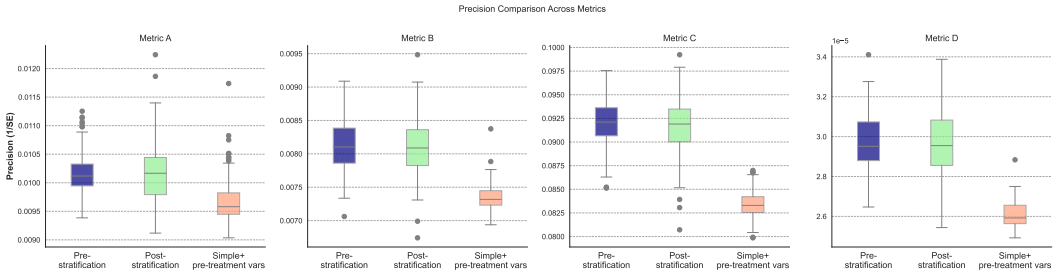

Figure 1: Precision summary for all metrics and all methods

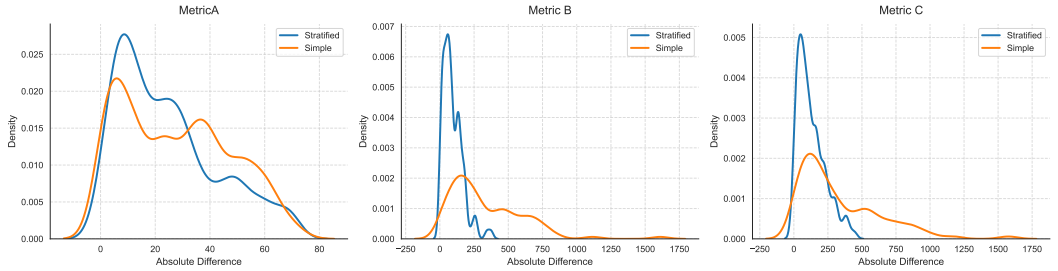

Figure 2: Balance gains from stratification

Additional analyses reveal that benefits decrease by 24-63% in more homogeneous populations, particularly in smaller samples, while larger samples prove more resilient to heterogeneity reduction (only 1-5% reduction) (Figure 4). Regarding strata complexity, while more strata can theoretically capture greater variation, our results suggest that moderate strata numbers (16-64) can capture most benefits while maintaining implementation feasibility.

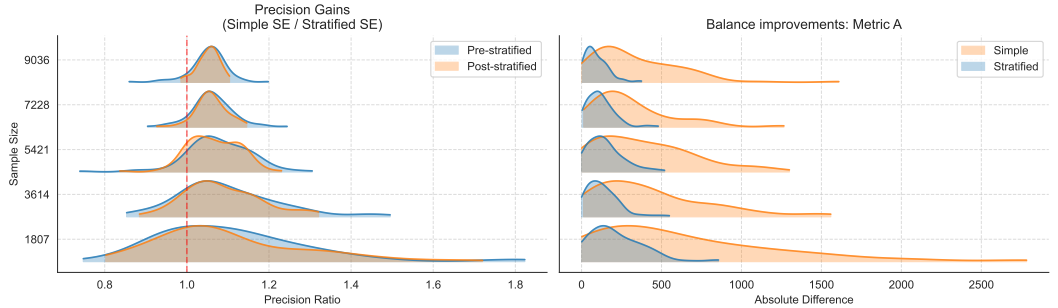

Figure 3: Precision gains in less heterogeneous samples

5. Conclusions This paper demonstrates how stratification methods can improve experimental design through both precision and balance gains. Our analysis reveals that pre- and post-stratification achieve virtually identical precision improvements, particularly valuable in smaller samples. However, their implementation trade-offs differ substantially. Post-stratification offers greater flexibility by allowing adjustments for any pre-treatment variable after randomization. Pre-stratification, while requiring more complex implementation, ensures better balance across treatment groups - a crucial advantage when stakeholder trust requires visible balance in treatment assignment.

These findings suggest that the choice between methods should depend on three key factors: sample size, population heterogeneity, and the importance of demonstrable balance. While precision gains diminish with sample size and homogeneity, balance improvements remain stable. Understanding these trade-offs can help experimenters make informed decisions about when and how to implement stratification, considering both statistical benefits and practical constraints.

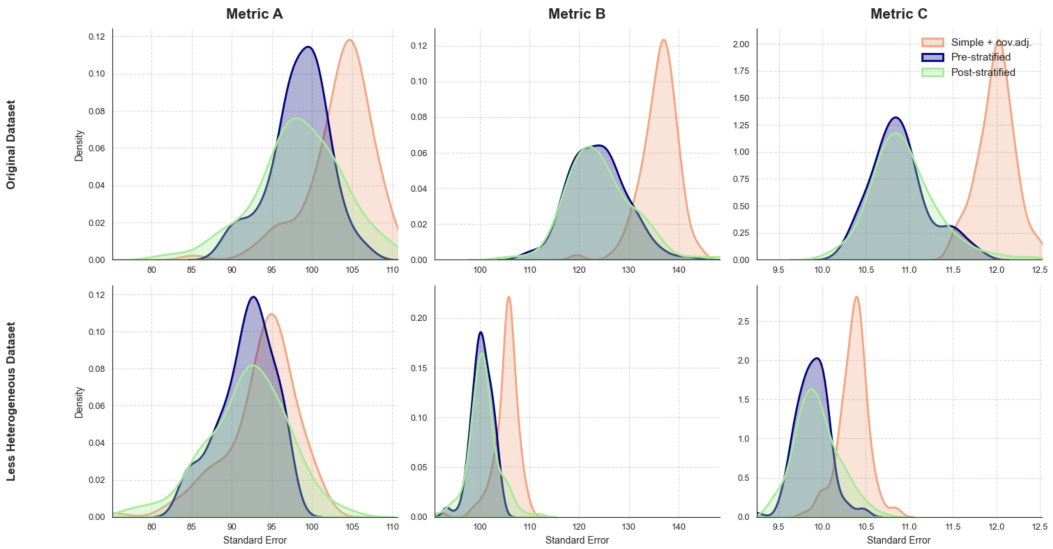

Figure 4: Precision and balance gains across sample sizes

References

L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(2):369–396, 2013.