Stratification and Post-Stratification in Randomized Experiments

Peng Ding, Email: pengdingpku@berkeley.edu

Block what you can and randomize what you cannot. – G. Box

1. Stratification

aka randomized block design

Why? Complete randomization may generate undesirable treatment allocation. Consider a discrete covariate $X_i \in \{1, \ldots, K\}$ with $n_{[k]} = \#\{i : X_i = k\}$ and $\pi_{[k]} = n_{[k]}/n$ for stratum k. A completely randomized experiment assigns n_1 units to the treatment group and n_0 units to the control group, which results in

$$n_{[k]1} = \#\{i : X_i = k, Z_i = 1\}, \quad n_{[k]0} = \#\{i : X_i = k, Z_i = 0\}$$

units in the treatment and control groups within stratum k. With positive probability, $n_{[k]1}$ or $n_{[k]0}$ is zero for some k. Even none of the $n_{[k]1}$'s or $n_{[k]0}$'s are zero, with high probability

$$\frac{n_{[k]1}}{n_1} - \frac{n_{[k]0}}{n_0} \neq 0. \tag{1}$$

That is, the proportions of units from stratum k are different across the treatment and control groups although on average their difference is zero:

$$E\left(\frac{n_{[k]1}}{n_1} - \frac{n_{[k]0}}{n_0}\right) = 0.$$

When $n_{[k]1}/n_1 - n_{[k]0}/n_0$ is large, the treatment and control groups have undesirable covariate imbalance.

How to avoid covariate imbalance? We can fix the $n_{[k]1}$'s or $n_{[k]0}$'s in advance and conduct strat-

ified randomized experiments (SRE), that is, we conduct K independent completely randomized experiments (CRE) within strata of X. The total number of randomizations is

$$\prod_{k=1}^{K} \binom{n_{[k]}}{n_{[k]1}},$$

and each feasible randomization has equal probability. Within stratum k, the proportion of units receiving the treatment is

$$e_{[k]} = \frac{n_{[k]1}}{n_{[k]}},$$

which is also called the propensity score.

For every unit i, we have potential outcomes $Y_i(1)$ and $Y_i(0)$, and individual causal effect $\tau_i = Y_i(1) - Y_i(0)$. For stratum k, we have stratum-specific average causal effect

$$\tau_{[k]} = n_{[k]}^{-1} \sum_{X_i=k} \tau_i.$$

The average causal effect is

$$\tau = n^{-1} \sum_{i=1}^{n} \tau_i = n^{-1} \sum_{k=1}^{K} \sum_{X_i=k} \tau_i = \sum_{k=1}^{K} \pi_{[k]} \tau_{[k]},$$

which is also the weighted average of the stratum-specific average causal effects.

If we are interested in $\tau_{[k]}$, then we can use the methods for the CRE within stratum k. Below I will discuss τ.

2. Fisherian inference

Sharp null hypothesis: $H_{0\mathrm{F}} : Y_i(1) = Y_i(0)$ for all units $i = 1, \ldots, n$.

A small modification in the FRT: we permute the treatment indices within strata of X (conditional permutation test).

Below I give some canonical choices of the test statistic.

Example 1 (Stratified estimator)**.** Motivated by estimating τ, we can use the following stratified

estimator in the FRT:

$$\hat{\tau}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]} \hat{\tau}_{[k]},$$

where

$$\hat{\tau}_{[k]} = n_{[k]1}^{-1} \sum_{i=1}^{n} I(X_i = k, Z_i = 1) Y_i - n_{[k]0}^{-1} \sum_{i=1}^{n} I(X_i = k, Z_i = 0) Y_i$$

is the stratum-specific difference-in-means within stratum k.

Example 2 (Van Elteren (1960)'s statistic)**.** We first compute the Wilcoxon rank sum statistic $W_{[k]}$ within stratum k and then combine them as

$$V = \sum_{k=1}^{K} \frac{W_{[k]}}{n_{[k]} + 1}.$$

Example 3 (Hodges and Lehmann (1962)'s aligned rank statistic)**.** Van Elteren (1960)'s statistic works well with a few large strata but does not work well with many small strata. Hodges and Lehmann (1962) proposed a test statistic that make more comparisons across strata after standardization. In particular, they suggested first standardizing the outcomes as

$$\tilde{Y}_i = Y_i - \bar{Y}_{[k]}$$

with the stratum-specific mean $\bar{Y}_{[k]} = n_{[k]}^{-1} \sum_{X_i = k} Y_i$, then obtaining the ranks $(\tilde{R}_1, \ldots, \tilde{R}_n)$ of the standardized outcomes $(\tilde{Y}_1, \ldots, \tilde{Y}_n)$, and finally constructing the test statistic

$$\tilde{W} = \sum_{i=1}^{n} Z_i \tilde{R}_i.$$

We can simulate the exact distributions of the above test statistics under the SRE. We can also calculate their means and variances and obtain the p-values based on Normal approximations. I do not find detailed discussion of the Kolmogorov–Smirnov statistic for the SRE, but below is my proposal.

Example 4 (Kolmogorov–Smirnov statistic)**.** We compute $D_{[k]}$, the maximum difference between the empirical distributions of the outcomes under treatment and control within stratum k. The

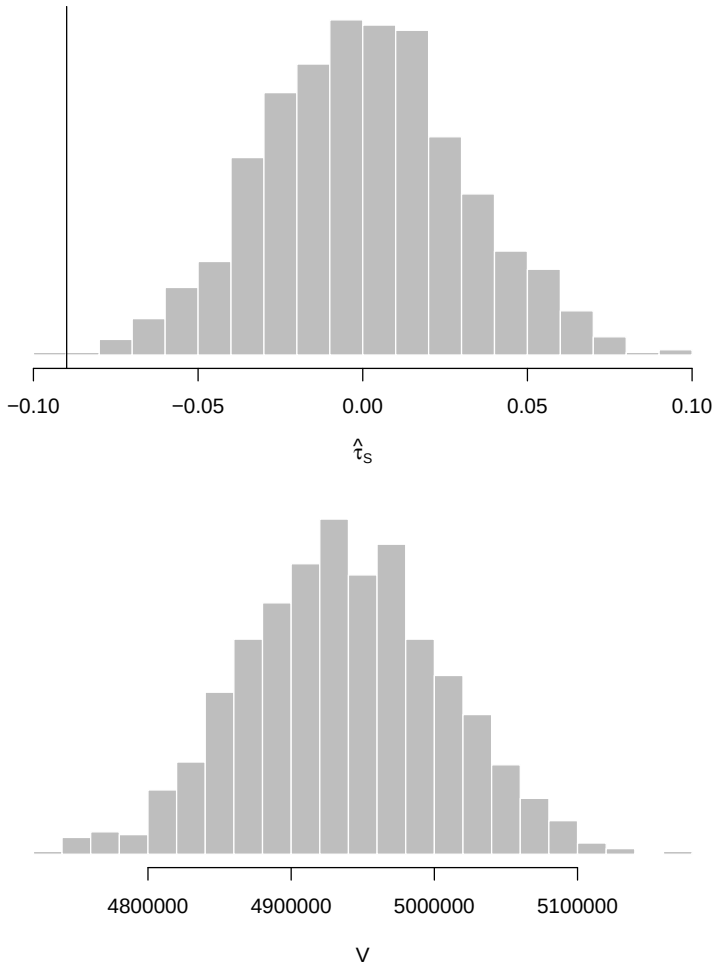

Figure 1: The randomization distributions of $\hat{\tau}_{\mathrm{S}}$ and V, with p-values 0.001 and 0.000

final test statistic is

$$D_{\mathrm{S}} = \sum_{k=1}^{K} \sqrt{\frac{n_{[k]1} n_{[k]0}}{n_{[k]}}} D_{[k]}$$

The Penn Bonus experiment example, in Figure 1.

3. Neymanian inference

3.1. Point and interval estimation

Within stratum k, the difference-in-means $\hat{\tau}_{[k]}$ is unbiased for $\tau_{[k]}$ with variance

$$\mathrm{var}(\hat{\tau}_{[k]}) = \frac{S^2_{[k]}(1)}{n_{[k]1}} + \frac{S^2_{[k]}(0)}{n_{[k]0}} - \frac{S^2_{[k]}(\tau)}{n_{[k]}}.$$

Therefore, the stratified estimator $\hat{\tau}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]}\hat{\tau}_{[k]}$ is unbiased for $\tau = \sum_{k=1}^{K} \pi_{[k]}\tau_{[k]}$ with variance

$$\mathrm{var}(\hat{\tau}_{\mathrm{S}}) = \sum_{k=1}^{K} \pi_{[k]}^2 \mathrm{var}(\hat{\tau}_{[k]}).$$

If $n_{[k]1} \geq 2$ and $n_{[k]0} \geq 2$, then we can obtain the sample variances $\hat{S}^2_{[k]}(1)$ and $\hat{S}^2_{[k]}(0)$ of the outcomes within stratum k and construct a conservative variance estimator

$$\hat{V}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]}^2 \left(\frac{\hat{S}^2_{[k]}(1)}{n_{[k]1}} + \frac{\hat{S}^2_{[k]}(0)}{n_{[k]0}} \right).$$

Wald-type confidence interval based on CLT: $\hat{\tau}_{\mathrm{S}} \pm 1.96\sqrt{\hat{V}_{\mathrm{S}}}$

CLT holds under two regimes: a few large strata and many small strata. Examples: ($K = 5, n_{[k]} = 80$) and ($K = 50, n_{[k]} = 8$). See Figure 2.

3.2. Comparing the SRE and the CRE

We assume that $e_{[k]} = e$ for all k and all strata are large. In this case, $\hat{\tau} = \hat{\tau}_{\mathrm{S}}$. We compare the sampling variances.

We decompose the finite population variances as

$$\begin{aligned}
S^2(1) &= (n-1)^{-1} \sum_{i=1}^{n} \{Y_i(1) - \bar{Y}(1)\}^2 \\
&= (n-1)^{-1} \sum_{k=1}^{K} \sum_{X_i=k} \{Y_i(1) - \bar{Y}_{[k]}(1) + \bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 \\
&= (n-1)^{-1} \sum_{k=1}^{K} \sum_{X_i=k} \left[\{Y_i(1) - \bar{Y}_{[k]}(1)\}^2 + \{\bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 \right] \\
&= \sum_{k=1}^{K} \left[\frac{n_{[k]} - 1}{n-1} S^2_{[k]}(1) + \frac{n_{[k]}}{n-1} \{\bar{Y}_{[k]}(1) - \bar{Y}(1)\}^2 \right],
\end{aligned}$$

and similarly,

$$\begin{aligned}
S^2(0) &= \sum_{k=1}^{K} \left[\frac{n_{[k]} - 1}{n-1} S^2_{[k]}(0) + \frac{n_{[k]}}{n-1} \{\bar{Y}_{[k]}(0) - \bar{Y}(0)\}^2 \right], \\
S^2(\tau) &= \sum_{k=1}^{K} \left[\frac{n_{[k]} - 1}{n-1} S^2_{[k]}(\tau) + \frac{n_{[k]}}{n-1} \{\tau_{[k]} - \tau\}^2 \right].
\end{aligned}$$

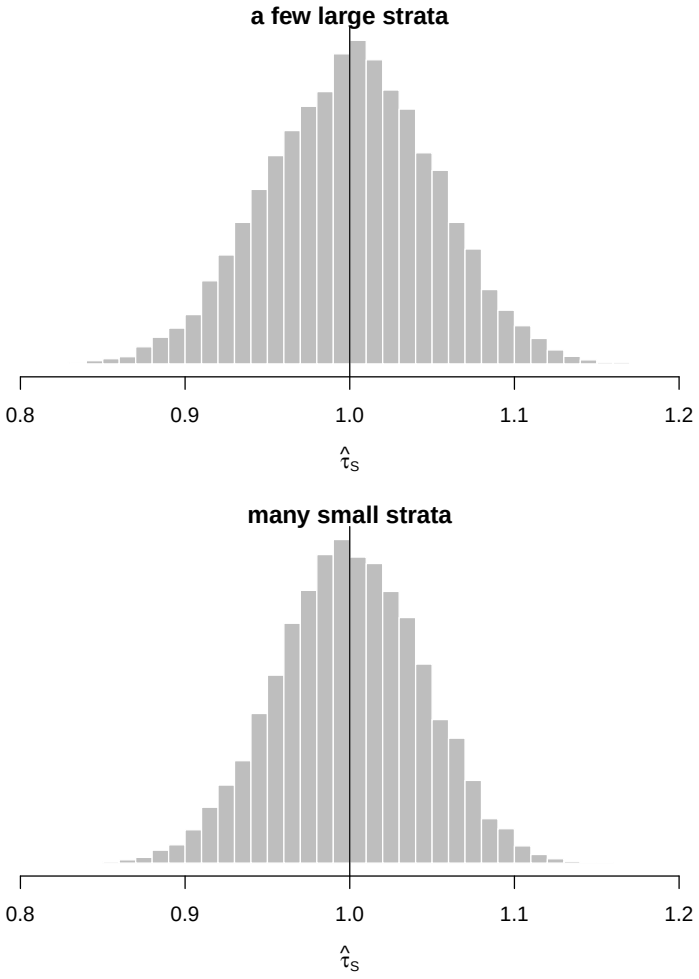

Figure 2: Normal approximations under two regimes

Therefore, the variance of the difference-in-means estimator under complete randomization is

$$\begin{aligned}\mathrm{var}_{\mathrm{CRE}}(\hat{\tau}) &\approx \sum_{k=1}^{K}\left[\frac{\pi_{[k]}}{n_1}S^2_{[k]}(1)+\frac{\pi_{[k]}}{n_0}S^2_{[k]}(0)-\frac{\pi_{[k]}}{n}S^2_{[k]}(\tau)\right] \\ &\quad+\sum_{k=1}^{K}\left[\frac{\pi_{[k]}}{n_1}\{\bar{Y}_{[k]}(1)-\bar{Y}(1)\}^2+\frac{\pi_{[k]}}{n_0}\{\bar{Y}_{[k]}(0)-\bar{Y}(0)\}^2-\frac{\pi_{[k]}}{n}\{\tau_{[k]}-\tau\}^2\right].\end{aligned}$$

We can then rewrite the variance of $\hat{\tau}_{\mathrm{S}}$ under the SRE as

$$\begin{aligned}\mathrm{var}_{\mathrm{SRE}}(\hat{\tau}_{\mathrm{S}}) &= \sum_{k=1}^{K}\pi^2_{[k]}\left[\frac{S^2_{[k]}(1)}{n_{[k]1}}+\frac{S^2_{[k]}(0)}{n_{[k]0}}-\frac{S^2_{[k]}(\tau)}{n_{[k]}}\right] \\ &= \sum_{k=1}^{K}\left[\frac{\pi_{[k]}}{n_1}S^2_{[k]}(1)+\frac{\pi_{[k]}}{n_0}S^2_{[k]}(0)-\frac{\pi_{[k]}}{n}S^2_{[k]}(\tau)\right],\end{aligned}$$

because $\pi_{[k]}/n_{[k]1}=1/(ne), \pi_{[k]}/n_{[k]0}=1/\{n(1-e)\}$, and $\pi_{[k]}/n_{[k]}=1/n$. Approximately, the difference between $\mathrm{var}_{\mathrm{CRE}}(\hat{\tau})$ and $\mathrm{var}_{\mathrm{SRE}}(\hat{\tau}_{\mathrm{S}})$ is

$$\sum_{k=1}^{K}\left[\frac{\pi_{[k]}}{n_1}\{\bar{Y}_{[k]}(1)-\bar{Y}(1)\}^2+\frac{\pi_{[k]}}{n_0}\{\bar{Y}_{[k]}(0)-\bar{Y}(0)\}^2-\frac{\pi_{[k]}}{n}(\tau_{[k]}-\tau)^2\right]. \tag{2}$$

We can easily show that the term in (2) is non-negative, and therefore, stratification helps to improve the asymptotic estimation efficiency for τ if the covariate is predictive to the outcome.

Remarks:

- The above comparison is based on the sampling variance, and we can also compare the estimated variance.
- In general, $\hat{\tau}_{\mathrm{S}} \neq \hat{\tau}$. When $e_{[k]}=e$ for all k, they are identical.
- Increasing K improves efficiency, but this argument depends on the large strata assumption. So we have a tradeoff in practice. We cannot arbitrarily increase K.
- We will discuss a special case $n_{[k]1}=n_{[k]0}=1$ later, which is called the matched pair experiment.

4. Post-stratification

In a CRE with a discrete covariate X, the numbers of units receiving the treatment and control are random within stratum k. In a SRE, these numbers are fixed. But if we conduct conditional inference given $\{n_{[k]1}, n_{[k]0}\}_{k=1}^{K}$, then a CRE becomes a SRE. So we can analyze a CRE with a discrete covariate X in the same way as in a SRE.

The FRT becomes a conditional FRT (Hennessy et al. 2016). The Neymanian analysis becomes post-stratification:

$$\hat{\tau}_{\text{PS}} = \sum_{k=1}^{K} \pi_{[k]} \hat{\tau}_{[k]},$$

which has an identical form as $\hat{\tau}_{\text{S}}$.

Remarks:

- Compared to $\hat{\tau}$, post-stratification improves efficiency in many cases (Miratrix et al. 2013).
- We cannot go too extreme because with a larger K it is more likely that some $n_{[k]1}$ or $n_{[k]0}$ become zero and we must modify the definition of $\hat{\tau}_{\text{PS}}$.
- Stratification uses X in the design stage and post-stratification uses X in the analysis stage. I view them as duals. Asymptotically, their difference is small with large strata (Miratrix et al. 2013).

References

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., and Sarkar, P. (2016). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4:61–80.

Hodges, J. and Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, 33:482–497.

Miratrix, L. W., Sekhon, J. 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)*, 75:369–396.

Van Elteren, P. (1960). On the combination of independent two-sample tests of wilcoxon. *Bulletin of the Institute of International Statistics*, 37:351–361.