CI = STRATIFICATION & RERANDOMIZATION

In this chapter, we discuss stratification and rerandomization, two strategies that improve covariate balance in RCTs.

I - STRATIFIED DESIGNS

- For a set of units $i = 1, .., n$, observe (X_i, Y_i, W_i) where
  - ↘ $X_i \in \{1, .., K\}$ = discrete covariate. We say that i belongs to the k-th stratum ⟺ $X_i = k$
  - ↘ $W_i \in \{0, 1\}$ = treatment allocation. In this section, we cosider two random allocations: completely randomized, and stratified.
  - ↘ $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$
  - ↘ $(Y_i(0), Y_i(1))$ = Potential Outcomes.

We are interested in estimating the ATE $= E(Y_i(1) - Y_i(0))$ (under an infinite population model for the P.O.)

- Put $\mu_{jk} := E(Y_i(j) \mid X_i = k)$

  $\sigma^2_{jk} := var(Y_i(j) \mid X_i = k)$

  The conditional means & variances may differ across strata (unless the covariate X_i is independent of the PO)

  ↳ A stratified design exploits this to reduce the variance of estimators of the ATE.

- Suppose you have collected n_k units with $X_i = k$, and put

$\hat{\pi}_k = \frac{n_k}{n}$. We consider two randomization procedures:

[CRE] In a CRE, draw a fixed number n_t of units at random from the total (sampled) population of n units. These units are assigned to the treatment group, and the remaining $n_c := n - n_t$ to the control group.

$$\mathbb{P}(\underline{W} \mid \underline{Y}(0), \underline{Y}(1), \underline{X}) = \frac{1}{\binom{n}{n_t}} \quad \forall \underline{W} \text{ s.t. } \sum_{i=1}^{n} W_i = n_t$$

↑ vector notation = $\underline{W} = (W_1, .., W_n)^t$

We write $\underline{W} \sim [CRE]$.

A natural choice for the estimation of the ATE is the difference in means estimator

$$\hat{\Delta} = \frac{1}{n_t} \sum_{i=1}^{n} W_i Y_i - \frac{1}{n_c} \sum_{i=1}^{n} (1 - W_i) Y_i .$$

[ST] In a [CRE], unfortunate draws may lead to severe unbalance of the covariates between the treatment and control groups. Taking for example $X_i \in \{\text{man}, \text{woman}\}$ $(K = 2)$, a [CRE] draw may place all men in the treatment group, and all women in the control group. Stratified designs do not allow this to happen by randomizing units stratum per stratum. Specifically, let

- n_{1k} = # of treated units in the k-th stratum
- $n_{0k} := n_k - n_{1k}$

Put $p_k = \frac{n_{1k}}{n_k}$ = proportion of treated units in the k-th stratum.

The assignment probability is

$$\mathbb{P}(\underline{W} \mid \underline{Y}(0), \underline{Y}(1), \underline{X}) = \prod_{k=1}^{K} \frac{1}{\binom{n_k}{n_{1k}}}$$

$$\forall \underline{W} \text{ s.t. } \sum_{X_i=k} W_i = n_{1k}$$

We write $[W] \sim [ST]$

The total number of treated units is $n_t = \sum_{k=1}^{K} n_{1k}$

$$= \sum_k p_k n_k \, .$$

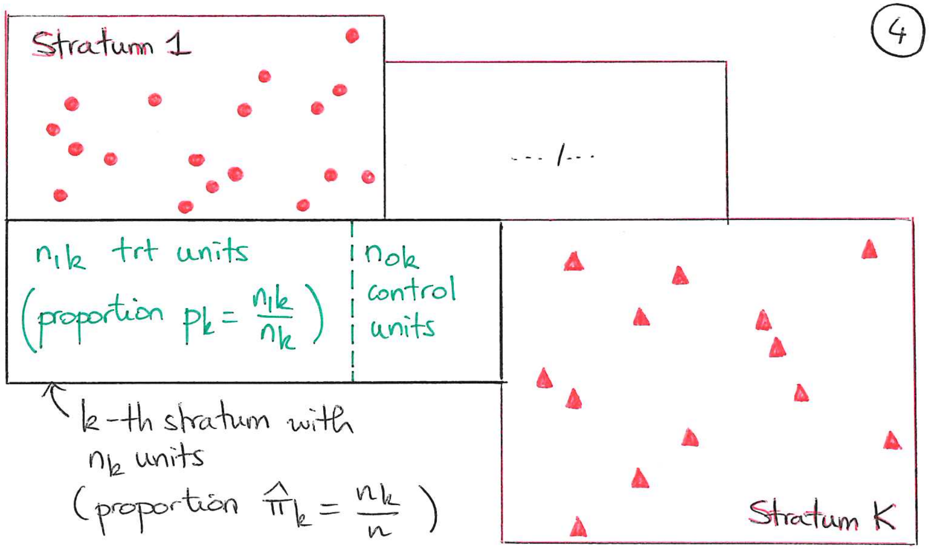
Under a [ST] randomisation scheme, we consider the aggregated difference-in-mean estimator:

$$\hat{\Delta}_{agg} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\Delta}(k)$$

$$\hat{\Delta}(k) = \frac{1}{n_{1k}} \sum_{\substack{W_i=1 \\ X_i=k}} Y_i - \frac{1}{n_{0k}} \sum_{\substack{W_i=0 \\ X_i=k}} Y_i$$

In the remainder of this section, we compare the bias and variance of $\hat{\Delta}$ and $\hat{\Delta}_{agg}$ under the two randomization schemes.

x Remark: Conditionally on the strata sizes, the ATE can be equivalently expressed $ATE = \sum_{k=1}^{K} \hat{\pi}_k \, \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = k)$.



k-th stratum with n_k units (proportion $\hat{\pi}_k = \frac{n_k}{n}$)

& $\mu_{jk} = \mathbb{E}(Y_i(j) \mid X_i = k)$

$\sigma^2_{jk} = \mathrm{Var}(Y_i(j) \mid X_i = k)$

- In addition,

$$\mu_j := \mathbb{E}\, Y_i(j) = \mathbb{E}_X\, \mathbb{E}(Y_i(j) \mid X) = \sum_k \hat{\pi}_k \mu_{jk}$$

We assume implicitely here that the n individuals represent the whole population of interest, so that $\mathbb{P}(X_i = k) = \frac{n_k}{n} = \hat{\pi}_k$

[Interlude] The assumption above is a consequence of considering non-random $n_1, \dots, n_K$. We may be interested instead in assuming a superpopulation model $\mathbb{P}(X_i = k) = \pi_k$; and selecting n individuals at random from that superpopulation. In this case,

$(n_1, \cdots, n_K) \sim$ Multinomial $(\pi_1, \dots, \pi_K)$ are random variables, and $\frac{n_k}{n} = \hat{\pi}_k \xrightarrow{a.s} \pi_k$ as $n \to \infty$. Choosing $\{n_k\}$ as fixed quantities removes some noise from the estimators & we begin our investigations under this assumption. The ATE becomes

$$ATE = \sum_{k=1}^{K} \hat{\pi}_k (\mu_{1k} - \mu_{0k})$$

. later on, we consider random $(n_1, .., n_K)$ and quantify the impact on the variance of our estimators.

- $\sigma_j^2 = Var\ Y_i(j)$

$$= \underbrace{\mathbb{E}_X \underbrace{var(Y_i(j) \mid X_i)}_{var(Y_i(j) \mid X_i = k) = \sigma_{jk}^2}}_{= \sum_{k=1}^{K} \hat{\pi}_k \sigma_{jk}^2} + \underbrace{Var_X \underbrace{\mathbb{E}(Y_i(j) \mid X_i)}_{\mathbb{E}(Y_i(j) \mid X_i = k) = \mu_{jk}}}_{= \sum_{k=1}^{K} \hat{\pi}_k (\mu_{jk} - \mu_j)^2}$$

thus $\boxed{\sigma_j^2 = \sum_{k=1}^{K} \hat{\pi}_k \sigma_{jk}^2 + \sum_{k=1}^{K} \hat{\pi}_k (\mu_{jk} - \mu_j)^2}$

Unconditional variance of the P.O. $Y_i(j)$

- The table on the next page summarizes the properties of the two estimators of the ATE in terms of their means (& bias) and variances.

[ST] Mean $\hat{\Delta}_{agg}$	$\mathbb{E}\hat{\Delta}_{agg} = \sum_k \hat{\pi}_k (\mu_{1k} - \mu_{0k})$ (zero bias)
[ST] Var $\hat{\Delta}_{agg}$	$n\ var\ \hat{\Delta}_{agg} = \sum_k \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k} \right)$
[CRE] Mean $\hat{\Delta}$	$\mathbb{E}\hat{\Delta} = \mu_1 - \mu_0 = \sum_k \hat{\pi}_k (\mu_{1k} - \mu_{0k})$ (zero bias)
[CRE] Var $\hat{\Delta}$	$n\ var\ \hat{\Delta} = \frac{n}{n_t} \left(\sum_k \hat{\pi}_k \sigma_{1k}^2 + \sum_k \hat{\pi}_k (\mu_{1k} - \mu_1)^2 \right)$ $+ \frac{n}{n_c} \left(\sum_k \hat{\pi}_k \sigma_{0k}^2 + \sum_k \hat{\pi}_k (\mu_{0k} - \mu_0)^2 \right)$

[See Appendix A]

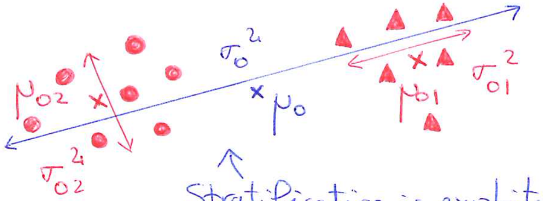
- Remarks

(i) Both $\hat{\Delta}$ and $\hat{\Delta}_{agg}$ are unbiased for the ATE

(ii) When $p_k = p \ \forall k$ (all strata have the same assignment probability), $n_t = pn$ and

$$\begin{cases} n\ var\ \hat{\Delta}_{agg} = \sum_k \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p} + \frac{\sigma_{0k}^2}{1-p} \right) \\ n\ var\ \hat{\Delta} = \sum_k \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p} + \frac{\sigma_{0k}^2}{1-p} \right) \\ \qquad + \sum_k \frac{\hat{\pi}_k}{p} (\mu_{1k} - \mu_1)^2 + \sum_k \frac{\hat{\pi}_k}{1-p} (\mu_{0k} - \mu_0)^2 \end{cases}$$

$\Rightarrow$ var $\hat{\Delta}$ > var $\hat{\Delta}_{agg}$ unless all means $\mu_{1k} = \mu_1$, $\mu_{0k} = \mu_0$ are equal across all strata. In this case, the P.O. are uncorrelated with X and stratification does not reduce the variance compared to a simple [CRE] procedure. On the other hand, the more heterogeneity there is across strata and the more variance reduction we get.



Stratification is exploiting this type of structure in the data

In practice, the strata are constructed from treatment-insensitive vector of covariates. For example, $X_i \in \{$man, woman$\}$ and we randomize individuals within the m and f populations. With a constant $p_k = p$ across the two strata, this ensures that men and women are well balanced across the treatment and control groups. The [CRE] may suffer from some unlikely imbalance for a specific random draw.

With continuous covariate, we may start by discretizing them first, or consider rerandomization (see later)

x Remark: We may consider a [ST] randomization scheme, but analyse the experiment omitting stratum information, pooling all the data together and using a difference in means estimator. Doing so yields the following expressions for the mean and variance:

[ST] Mean $\hat{\Delta}$	$\mathbb{E}\hat{\Delta} = \sum_k \hat{\pi}_k \left(\frac{n p_k}{n_t} \mu_{1k} - \frac{n(1-p_k)}{n_c} \mu_{0k} \right)$
[ST] Var $\hat{\Delta}$	$n \operatorname{var} \hat{\Delta} = \sum_k \hat{\pi}_k \left(\left(\frac{n}{n_t}\right)^2 p_k \sigma_{1k}^2 + \left(\frac{n}{n_c}\right)(1-p_k)\sigma_{0k}^2 \right)$

[Appendix B]

$\searrow$ the difference in means estimator computed under a [ST] randomization scheme is biased in general, unless $p_k = p \ \forall k$, or $\mu_{1k} = \mu_1$ and $\mu_{0k} = \mu_0 \ \forall k$

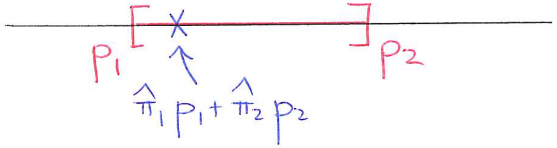
$\searrow$ however, the variance of $\hat{\Delta}$ may be smaller than that of $\hat{\Delta}_{agg}$ $\longrightarrow$ bias / variance tradeoff.

Recall that $n \operatorname{var} \hat{\Delta}_{agg} = \sum_k \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k} \right)$

Comparing this expression with $n \operatorname{var} \hat{\Delta}$ indicates that one must evaluate the respective contributions of $\left(\frac{n}{n_t}\right)^2 p_k$ and $\frac{1}{p_k}$ in front of σ_{1k}^2 (and similarly for σ_{0k}^2).

Note that $\left(\frac{n}{n_t}\right)^2 p_k > \frac{1}{p_k} \Leftrightarrow \frac{p_k}{\sum_\ell p_\ell \hat{\pi}_\ell} > 1$

Take $K=2$, $p_1 < p_2$. Since $\hat{\pi}_1 + \hat{\pi}_2 = 1$, $p_1\hat{\pi}_1 + p_2\hat{\pi}_2$ lies in the interval $[p_1, p_2]$



and necessarily one $\frac{p_k}{\sum_\ell p_\ell \hat{\pi}_\ell}$ is larger than 1, and the other one smaller than 1.

↳ Select $\hat{\pi}_1$ small so that $\hat{\pi}_1 p_1 + \hat{\pi}_2 p_2$ lies close to p_2.

Then

$$\begin{cases} \frac{p_2}{\sum p_\ell \hat{\pi}_\ell} \text{ close to 1, slighty larger than 1} \\ \frac{p_1}{\sum p_\ell \hat{\pi}_\ell} \text{ close to } 0 \end{cases}$$

To find cases where $\hat{\Delta}$ has smaller variance than $\hat{\Delta}_{agg}$, we need to select

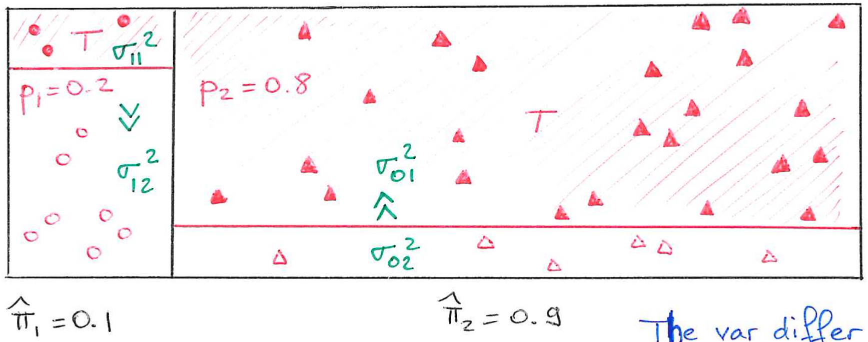
σ_{11}^2 large when $\left(\frac{n}{n_t}\right)^2 p_1 < \frac{1}{p_1} \Leftrightarrow \frac{p_1}{\sum p_\ell \hat{\pi}_\ell} < 1$

& σ_{12}^2 small when $\left(\frac{n}{n_t}\right)^2 p_2 > \frac{1}{p_2} \Leftrightarrow \frac{p_2}{\sum_\ell p_\ell \hat{\pi}_\ell} > 1$

& the other way around for σ_{01}^2 and σ_{02}^2.

Taking $\begin{cases} p_1 = 0.2 = 1 - p_2 \\ \hat{\pi}_1 = 0.1 = 1 - \hat{\pi}_2 \\ 10 = \sigma_{11}^2 > \sigma_{12}^2 = 1 \\ 2 = \sigma_{01}^2 < \sigma_{02}^2 = 8 \end{cases}$ does the job :

$$\begin{cases} n \text{ var } \hat{\Delta} = 25.9 \\ n \text{ var } \hat{\Delta}_{Agg} = 42.3 \end{cases}$$



The var differ on page 6 since the ω are $\sim$[ST] or [CRE]

* Remark = If all $p_k = p \ \forall k$,

$$\hat{\Delta}_{Agg} = \sum_{k=1}^{K} \hat{\pi}_k \left(\frac{1}{n_{1k}} \sum_{\substack{\omega_i=1 \\ X_i=k}} Y_i - \frac{1}{n_{0k}} \sum_{\substack{\omega_i=0 \\ X_i=k}} Y_i \right)$$

$$= \sum_{k=1}^{K} \frac{1}{n} \left(\underbrace{\frac{n_k}{n_{1k}}}_{= \frac{1}{p}} \sum_{\substack{\omega_i=1 \\ X_i=k}} Y_i - \underbrace{\frac{n_k}{n_{0k}}}_{= \frac{1}{1-p}} \sum_{\substack{\omega_i=0 \\ X_i=k}} Y_i \right)$$

$$= \underbrace{\frac{1}{np}}_{= \frac{1}{n_t}} \sum_{k=1}^{K} \sum_{\substack{\omega_i=1 \\ X_i=k}} Y_i - \underbrace{\frac{1}{n(1-p)}}_{= \frac{1}{n_c}} \sum_{k=1}^{K} \sum_{\substack{\omega_i=0 \\ X_i=k}} Y_i = \hat{\Delta}$$

x Remark : Mean and variances can be computed unconditionally on $(n_1, \dots, n_K)$, assuming a multinomial distribution $\text{Mult}(\pi_1, \dots, \pi_K)$ for $(n_1, \dots, n_K)$ [Since all X_i are assumed independent with $\mathbb{P}(X_i = k) = \pi_k$]. The ATE is $\sum_{k=1}^{K} \pi_k(\mu_{1k} - \mu_{0k})$.

[ST] Mean $\hat{\Delta}_{agg}$	$\mathbb{E}\hat{\Delta}_{agg} = \sum_{k=1}^{K} \pi_k(\mu_{1k} - \mu_{0k})$ (unbiased)
[ST] Var $\hat{\Delta}_{agg}$	$n\text{var}\,\hat{\Delta}_{agg} = \sum_k \pi_k\left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k}\right)$ $+ \sum_k \pi_k(\mu_{1k} - \mu_{0k})^2$ $- \left[\sum_k \pi_k(\mu_{1k} - \mu_{0k})\right]^2$
[CRE] Mean $\hat{\Delta}$	$\mathbb{E}\hat{\Delta} = \sum_{k=1}^{K} \pi_k(\mu_{1k} - \mu_{0k})$ (unbiased)
[CRE] Var $\hat{\Delta}$ [(n_t, n_c) remain fixed]	$n\text{var}\,\hat{\Delta} = \left(\frac{n}{n_t}\right)\sum_{k=1}^{K} \pi_k\left[\sigma_{1k}^2 + (\mu_{1k} - \mu_1)^2\right]$ $+ \left(\frac{n}{n_c}\right)\sum_{k=1}^{K} \pi_k\left[\sigma_{0k}^2 + (\mu_{0k} - \mu_0)^2\right]$

[See Appendix C]

↘ Both $\hat{\Delta}_{agg}$ and $\hat{\Delta}$ remain unbiased

↘ Comparing the variance terms with those on page 6, we see that taking into account the variability in the n_k's increases the variance of the estimators (as expected). The green terms p.11 represent the additional terms.

↘ The expression of var $\hat{\Delta}_{agg}$ under [ST] was also derived on p.3 in CI: UNCONFOUNDEDNESS.

Similar calculations can be derived for $\hat{\Delta}$ under [ST]. For example, one can show that

$$\mathbb{E}\hat{\Delta} = \sum_k \pi_k\left(\frac{p_k}{\sum p_\ell \pi_\ell}\mu_{1k} - \frac{1-p_k}{\sum_\ell (1-p_\ell)\pi_\ell}\mu_{0k}\right) + O(n^{-1}$$

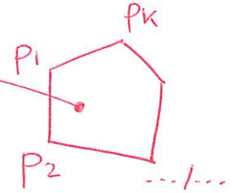
Under $\underline{W} \sim$ [ST]

[See Appendix D]

[n_t, n_c are randon here, unlike under $\underline{W} \sim$ [CRE]]

Note that the asymptotic term differs from the ATE, since all coefficients $\frac{p_k}{\sum p_\ell \pi_\ell}$ ldiffer from the value 1; unless all p_k are equal.

↘ To see this geometrically, note that $\sum p_\ell \pi_\ell$ belongs to the interior of the convex hull of $\{p_1, \dots, p_K\}$ for $0 < \pi_\ell < 1$ $\forall \ell$



x **Remark:** Finite population approach.

The mean and variance calculations can be derived similarly assuming fixed potential outcomes $\{Y_i(0), Y_i(1)\}$. One can show that

$W \sim [ST]$

$$n \operatorname{var} \hat{\Delta}_{agg} = \sum_{k=1}^{K} \hat{\pi}_k \left(\frac{s_{1k}^2}{p_k} + \frac{s_{0k}^2}{1-p_k} - s_{\Delta k}^2 \right)$$

where
- $\hat{\pi}_k = \frac{n_k}{n}$; n_k fixed (non-random) ; $p_k = \frac{n_{1k}}{n_k}$
- $s_{jk}^2 := \frac{1}{n_k - 1} \sum_{X_i = k} \left(Y_i(j) - \bar{Y}_k(j) \right)^2$

  $\bar{Y}_k(j) := \frac{1}{n_k} \sum_{X_i = k} Y_i(j)$
- $s_{\Delta k}^2 := \frac{1}{n_k - 1} \sum_{X_i = k} \left(Y_i(1) - Y_i(0) - \bar{\Delta}_k \right)^2$

  $\bar{\Delta}_k := \frac{1}{n_k} \sum_{X_i = k} \left(Y_i(1) - Y_i(0) \right)$

See also p.7 in <u>CI : RANDOMIZED CONTROL TRIALS</u> for the expression of the variance of the simple difference in means in the finite population approach. ■

x **Remark:**

$(n_1, .., n_K)$ fixed = sample n_k units from category k, where the K categories are known in advance.

$(n_1, .., n_K)$ random = sample n units & then create strata based on covariates. The # of units in each stratum is not known in advance.

II - RERANDOMIZATION

Li, Ding & Rubin (2018) : "Although complete randomization ensures covariate balance on average, the chance of observing significant differences between treatment and control covariate distributions increases with many covariates. Rerandomization discards randomizations that do not satisfy a predetermined covariate balance criterion, generally resulting in better covariate balance and more precise estimates of causal effects."

In this section, we present asymptotic results for the difference in means estimator of the ATE under a rerandomization scheme using a Mahalanobis criterion: For some covariates X_i, compute $\hat{\Delta}_X := \frac{1}{n_t} \sum_{W_i=1} X_i - \frac{1}{n_c} \sum_{W_i=0} X_i$ and accept the assignment vector $(W_1, .., W_n)$ if the normalized covariate mean difference between the treatment and control groups $\frac{\hat{\Delta}_X^2}{\operatorname{var} \hat{\Delta}_X}$ is less than some predefined threshold ε.

Specifically, we provide LT (Limit Theorems) for $\boxed{\sqrt{n}(\hat{\Delta} - ATE) \,\Big|\, \frac{\hat{\Delta}_X^2}{\operatorname{var} \hat{\Delta}_X} \leq \varepsilon}$

↘ Assuming a finite & infinite population model

↘ Based on the work of Morgan & Rubin (2012, 2015), Li, Ding & Rubin (2018)

↘ Extend to stratified designs Wang, Wang, Liu (2021).

We proceed as follows:

(i) Provide CLT in the finite & infinite population models for $\hat{\Delta}$

(ii) Provide limit Theorems for the difference in means estimator under rerandomization with Mahalanobis distance

(iii) Provide limit Theorems under a stratified design with rerandomization.

II.1. Limit Theorems for $\hat{\Delta}$

- Set-up: For a set of units $i = 1, \dots, n$ we observe (X_i, Y_i, W_i) where
  - $X_i \in \mathbb{R}$ = covariate
  - $W_i \in \{0,1\}$; $\sum_{i=1}^{n} W_i = n_t$

    $(W_1, \dots, W_n) \sim$ Completely Randomized
  - $Y_i = W_i Y_i(1) + (1-W_i) Y_i(0)$

We consider both a finite & infinite population approach:

[finite] $\text{ATE} = \Delta^n = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$

[∞] $\text{ATE} = \Delta^\infty = \mathbb{E}(Y_i(1) - Y_i(0))$

$$\hat{\Delta} = \frac{1}{n_t} \sum_{i=1}^{n} W_i Y_i - \frac{1}{n_c} \sum_{i=1}^{n} (1-W_i) Y_i$$

We discard covariate information X_i in this section

- Goal:

[finite] Establish $\sqrt{n}\,(\hat{\Delta} - \Delta^n) \xrightarrow{d}$?

[∞] Establish $\sqrt{n}\,(\hat{\Delta} - \Delta^\infty) \xrightarrow{d}$?

- [finite] Put
  - $S_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(j) - \bar{Y}(j))^2$
  - $S_{01} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(1) - \bar{Y}(1))(Y_i(0) - \bar{Y}(0))$
  - $S_\Delta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0) - \Delta^n)^2$

One can show that $\hat{\Delta}$ has mean Δ^n and variance

$$\text{var}\,\hat{\Delta} = \frac{S_0^2}{n_c} + \frac{S_1^2}{n_t} - \frac{S_\Delta^2}{n} \quad \text{(see also p.13)}$$

$$= \frac{S_0^2}{n_c} + \frac{S_1^2}{n_t} - \frac{1}{n}\left(S_0^2 + S_1^2 - 2S_{01}\right)$$

since $S_\Delta^2 = \frac{1}{n-1} \sum (Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0)))^2$

$= \frac{1}{n-1} \sum (Y_i(1) - \bar{Y}(1) - (Y_i(0) - \bar{Y}(0)))^2$

$= S_0^2 + S_1^2 - 2S_{01}$

$$= \left(\frac{1}{n_c} - \frac{1}{n}\right) S_0^2 + \left(\frac{1}{n_t} - \frac{1}{n}\right) S_1^2 + 2\,\frac{S_{01}}{n}$$

put $\hat{\rho}_{01} = \frac{S_{01}}{S_0 S_1}$ = sample correlation coefficient

$$= \frac{n_t}{n\, n_c} S_0^2 + \frac{n_c}{n\, n_t} S_1^2 + \frac{2}{n} \hat{\rho}_{01} S_0 S_1$$

Assuming $\frac{n_t}{n} \to p \quad \frac{n_c}{n} \to 1-p$ as $n \to \infty$,

& that $S_0^2, S_1^2, \hat{\rho}_{01} \to \sigma_0^2, \sigma_1^2, \rho_{01}$,

$$\boxed{n \,\text{var}\,\hat{\Delta} \longrightarrow V_\Delta(p) = \frac{p}{1-p}\sigma_0^2 + \frac{1-p}{p}\sigma_1^2 + 2\rho_{01}\sigma_0\sigma_1 \quad (*)}$$

- [∞] All derivations can be made conditionally on a design $\underline{\omega} = (\omega_1, \dots, \omega_n)$. (17)

$$\operatorname{var}(\hat{\Delta} \mid \underline{\omega}) = \operatorname{var}\Big(\sum_{i=1}^{n} \underbrace{\Big\{ \frac{\omega_i Y_i(1)}{n_t} - \frac{(1-\omega_i) Y_i(0)}{n_c} \Big\}}_{=:\, T_i} \mid \underline{\omega} \Big)$$

$$= \sum_{i=1}^{n} \operatorname{var}(T_i \mid \underline{\omega}) + \sum_{j \neq i} \cancelto{0}{\operatorname{cov}(T_i, T_j \mid \underline{\omega})}$$

independent

$$= \sum_{i=1}^{n} \Big\{ \frac{\omega_i^2}{n_t^2} \sigma_1^2 + \frac{(1-\omega_i)^2}{n_c^2} \sigma_0^2 - \frac{2}{n_c n_t} \cancelto{0}{\omega_i (1-\omega_i)} \operatorname{cov}(Y(0), Y(1))$$

$$= \frac{\sigma_1^2}{n_t} + \frac{\sigma_0^2}{n_c}.$$

↓

Note that unconditionally on $\underline{\omega}$,

$$\operatorname{var} \hat{\Delta} = \underbrace{\operatorname{var} \underbrace{\mathbb{E}(\hat{\Delta} \mid \underline{\omega})}_{= \mu_1 - \mu_0}}_{= 0} + \mathbb{E} \operatorname{var}(\hat{\Delta} \mid \underline{\omega}) = \operatorname{var}(\hat{\Delta} \mid \underline{\omega})$$

(**)

$$\boxed{n \operatorname{var} \hat{\Delta} \longrightarrow V_\Delta^{\infty}(p) := \frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p}.}$$

∎

Theorem:

[finite] $\sqrt{n}(\hat{\Delta} - \Delta^n) \xrightarrow{d} \mathcal{N}(0, V_\Delta(p))$

[∞] $\sqrt{n}(\hat{\Delta} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V_\Delta^{\infty}(p))$

where $V_\Delta(p)$, $V_\Delta^{\infty}(p)$ are defined in (*) and (**) p. 16/17

× Remarks: (i) The results p. 17 complement the discussion page 12 in CI: UNCONFOUNDEDNESS & justify the normal approximation made by Neyman. (18)

(ii) When the P.O. are uncorrelated, $\rho_{01} = 0$.

Putting $\sigma_1^2 := \lambda \sigma_0^2$; $\lambda > 0$,

$$\begin{cases} V_\Delta^{\circ}(p) = \Big(\frac{p}{1-p} + \frac{1-p}{p} \lambda \Big) \sigma_0^2 \\ V_\Delta^{\infty}(p) = \Big(\frac{1}{1-p} + \frac{1}{p} \lambda \Big) \sigma_0^2 \end{cases}$$

We immediately see that $V_\Delta^{\circ}(p) < V_\Delta^{\infty}(p)$ $\forall p \in (0,1)$.

(iii) When the P.O. are perfectly correlated, $\rho_{01} = 1$ [sharp null or constant effect across units]

$$\begin{cases} \overline{V}_\Delta(p) = \Big(\frac{p}{1-p} + \frac{1-p}{p} \lambda + 2\sqrt{\lambda} \Big) \sigma_0^2 \\ V_\Delta^{\infty}(p) = \Big(\frac{1}{1-p} + \frac{1}{p} \lambda \Big) \sigma_0^2 \end{cases}$$

One can easily see that $\overline{V}_\Delta(p) \leq V_\Delta^{\infty}(p)$ with equality iff $\lambda = 1 \iff \sigma_0^2 = \sigma_1^2$.

Summary:

- If the P.O. are uncorrelated $\quad V_\Delta(p) < V_\Delta^{\infty}(p) \quad \forall p \in (0,1)$
- If the P.O. are perfectly linearly correlated $\quad V_\Delta(p) \leq V_\Delta^{\infty}(p) \quad \forall p \in (0,1)$ with equality iff $\sigma_0^2 = \sigma_1^2$

II.2. Rerandomization in CRE

For $X_i \in \mathbb{R}$ (univariate), let

$$\hat{\Delta}_X = \frac{1}{n_t}\sum_{i=1}^{n} W_i X_i - \frac{1}{n_c}\sum_{i=1}^{n}(1-W_i)X_i$$

denote the covariate mean difference between the treatment and control group, given the allocation $\underline{W} = (W_1, \dots, W_n)$.

We consider a rerandomization criterion based on Mahalanobis distance $\dfrac{\hat{\Delta}_X^{\,2}}{\operatorname{var}\hat{\Delta}_X}$

the variance of $\hat{\Delta}_X$ depends on the chosen framework = finite or ∞ pop.

Extends to multivariate $X_i \in \mathbb{R}^p$ using $\hat{\Delta}_X^{\,t}\,\Sigma_X^{-1}\,\hat{\Delta}_X$; Σ_X = covariance matrix of X_i

For some $\varepsilon > 0$, put $\mathcal{M}_\varepsilon := \left\{ \dfrac{\hat{\Delta}_X^{\,2}}{\operatorname{var}\hat{\Delta}_X} \leq \varepsilon \right\}$.

To establish the asymptotic distributions of

$$\left(\begin{array}{l} \sqrt{n}\,(\hat{\Delta} - \Delta^n) \mid \mathcal{M}_\varepsilon \\ \sqrt{n}\,(\hat{\Delta} - \Delta^\infty) \mid \mathcal{M}_\varepsilon \end{array}\right. ,$$

we first need the joint asymptotic distribution of $\sqrt{n}\begin{pmatrix}\hat{\Delta} - \Delta^n \\ \hat{\Delta}_X\end{pmatrix}$ (finite) and $\sqrt{n}\begin{pmatrix}\hat{\Delta} - \Delta^\infty \\ \hat{\Delta}_X\end{pmatrix}$ (∞), generalizing the results of section II.1.

Note that in both cases $\mathbb{E}\,\hat{\Delta}_X = 0$ holds.

x Notation =

[finite]

non-random

- $S_{xj} := \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(Y_i(j) - \bar{Y}(j)\right)\left(X_i - \bar{X}\right) \xrightarrow[\text{value}]{\text{some}} c_{xj}$
- $S_X^2 := \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 \longrightarrow \sigma_X^2$
- $S_j^2 \to \sigma_j^2$, $\hat{\rho}_{01} \to \rho_0$ } see p.16

[∞]

- $\begin{pmatrix} X \\ Y(0) \\ Y(1) \end{pmatrix}$ has covariance matrix $\begin{pmatrix} \sigma_X^2 & c_{X0} & c_{X1} \\ * & \sigma_0^2 & c_{01} \\ * & * & \sigma_1^2 \end{pmatrix}$

Theorem = Suppose that $\dfrac{n_t}{n} \to p$ as $n \to \infty$. Then

[finite] $\quad \sqrt{n}\begin{pmatrix}\hat{\Delta} - \Delta^n \\ \hat{\Delta}_X\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \left(\begin{array}{c|c} V_\Delta & V_{\Delta X} \\ \hline * & V_X \end{array}\right)\right)$

[∞] $\quad \sqrt{n}\begin{pmatrix}\hat{\Delta} - \Delta^\infty \\ \hat{\Delta}_X\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \left(\begin{array}{c|c} V_\Delta^\infty & V_{\Delta X} \\ \hline * & V_X \end{array}\right)\right)$

where → V_Δ and V_Δ^∞ are defined in (*) and (**) p.16/17

→ $V_{\Delta X} = \dfrac{c_{X0}}{1-p} + \dfrac{c_{X1}}{p}$, $\quad V_X = \dfrac{\sigma_X^2}{p(1-p)}$ } are common to the two frameworks.

proof = The finite population CLT is used in Li, Ding & Rubin (2018), and proved in Li & Ding (2017) under their regularity condition (18).

In the [∞] framework,

- $$\operatorname{cov}(\hat\Delta, \hat\Delta_X) = \operatorname{cov}\left(\sum_i \left\{ \frac{w_i Y_i(1)}{n_t} - \frac{(1-w_i) Y_i(0)}{n_c} \right\}, \sum_i \left\{ \frac{w_i X_i}{n_t} - \frac{(1-w_i) X_i}{n_c} \right\} \right)$$

$$= \sum_{i,j} \operatorname{Cov}\left(\frac{w_i Y_i(1)}{n_t} - \frac{(1-w_i) Y_i(0)}{n_c}, \frac{w_j X_j}{n_t} - \frac{(1-w_j) X_j}{n_c} \right)$$

$$= \sum_i \left\{ \frac{w_i}{n_t^2} \operatorname{cov}(Y_i(1), X_i) + \frac{1-w_i}{n_c^2} \operatorname{Cov}(Y_i(0), X_i) \right\}$$

$$= \frac{c_{X1}}{n_t} + \frac{c_{X0}}{n_c}.$$

- $$\operatorname{var} \hat\Delta_X = \left(\frac{1}{n_c} + \frac{1}{n_t} \right) \sigma_X^2 = \frac{n}{n_c n_t} \sigma_X^2$$ ∎

Theorem Li, Ding & Rubin (2018) for [finite]:

[finite] $\operatorname{var} \sqrt{n} (\hat\Delta - \Delta^n) | \mathcal{M}_\varepsilon \longrightarrow V_\Delta(p)\left(1 - (1 - v_{1,\varepsilon}) R^2\right)$

[∞] $\operatorname{var} \sqrt{n} (\hat\Delta - \Delta^\infty) | \mathcal{M}_\varepsilon \longrightarrow V_\Delta^\infty(p)\left(1 - (1 - v_{1,\varepsilon}) R_\infty^2\right)$

where

$$R^2 = \frac{V_{\Delta X} V_X^{-1} V_{X\Delta}}{V_\Delta} \quad \& \quad R_\infty^2 = \frac{V_{\Delta X} V_X^{-1} V_{X\Delta}}{V_\Delta^\infty}$$

$$v_{p,\varepsilon} = \frac{\mathbb{P}(\chi^2_{p+2} \leq \varepsilon)}{\mathbb{P}(\chi^2_p \leq \varepsilon)} \quad ; \quad X \in \mathbb{R}^p, \ \varepsilon = \text{threshold}$$

↑ The results of Li, Ding & Rubin (2018) extend naturally to the infinite population framework with V_Δ replaced by V_Δ^∞.

Note that when $X \in \mathbb{R}$ (univariate)

$$R^2 = \frac{\left(\frac{c_{X0}}{1-p} + \frac{c_{X1}}{p} \right)^2}{\frac{\sigma_X^2}{p(1-p)} \left(\frac{p}{1-p} \sigma_0^2 + \frac{1-p}{p} \sigma_1^2 + 2\rho_{01} \sigma_1 \sigma_0 \right)}$$

$$R_\infty^2 = \frac{\left(\frac{c_{X0}}{1-p} + \frac{c_{X1}}{p} \right)^2}{\frac{\sigma_X^2}{p(1-p)} \left(\frac{\sigma_0^2}{1-p} + \frac{\sigma_1^2}{p} \right)}$$

$f(p)$ = common component of R^2 and R_∞^2.

Corollary =

[finite] Asymptotic variance is $V_\Delta(p) - (1 - v_{1,\varepsilon}) f(p)$

[∞] —— " —— is $V_\Delta^\infty(p) - \underbrace{(1 - v_{1,\varepsilon}) f(p)}_{}$

Same variance reduction in absolute value in the two frameworks.

x Remark: The higher R^2 and the more the variance reduction, where R^2 corresponds to the squared correlation coefficient between $\hat\Delta$ and $\hat\Delta_X$ i.e. the proportion of variance of $\hat\Delta$ explained by $\hat\Delta_X$. ∎

In general, Li, Ding & Rubin (2018) showed that the limit distribution of $\sqrt{n}(\hat\Delta - \Delta^n) | \mathcal{M}_\varepsilon$ of two independent random variables : standard normal + truncated RV :

Theorem (Theorem 1 in Li, Ding & Rubin (2018))

[finite] $\sqrt{n}\,(\hat{\Delta} - \Delta^{n})\,|\,\mathcal{M}_{\varepsilon} \xrightarrow{d} \sqrt{V_\Delta}\left(\sqrt{1-R^2}\,\mathcal{N}(0,1) + \sqrt{R^2}\,L_{p,\varepsilon}\right)$

where • $L_{p,\varepsilon} \sim \chi_{p,\varepsilon}\, S \sqrt{\beta_p}$

— Truncated RV & has variance $v_{p,\varepsilon}$ appearing p. 21 —

↘ $\chi^2_{p,\varepsilon} \sim \chi^2_p \,|\, \chi^2_p \leq \varepsilon$

↘ $S = \pm 1$ w.p. 1/2

↘ $\beta_p \sim \text{Beta}\left(\frac{1}{2}, \frac{p-1}{2}\right)$

[∞] $\sqrt{n}\,(\hat{\Delta} - \Delta^{\infty})\,|\,\mathcal{M}_{\varepsilon} \xrightarrow{d} \sqrt{V^{\infty}_\Delta}\left(\sqrt{1-R^2_\infty}\,\mathcal{N}(0,1) + \sqrt{R^2_\infty}\,L_{p,\varepsilon}\right)$

↖The authors treat the finite population case explicitely only.

When the covariates are uncorrelated with the P.O., the R^2 coefficient vanish and we recover the usual CLT (with no reduction in variance).

When $R^2 \neq 0$, the asymptotic distribution of the difference in mean estimator is not normal. When computing confidence intervals based on the asymptotic distribution, this must be taken into account.

Let • z_α = α-th quantile of $\mathcal{N}(0,1)$

• $v_\alpha(R^2, p, \varepsilon)$ = α-th quantile of $\sqrt{1-R^2}\,\mathcal{N}(0,1) + \sqrt{R^2}\,L_{p,\varepsilon}$

// $v_\alpha(R^2)$ for simplicity

Under $\mathcal{M}_{\varepsilon}$, the $(1-\alpha)$ confidence interval for $\sqrt{n}\,(\hat{\Delta} - \Delta^{n})$ is $\left[v_{\frac{\alpha}{2}}(R^2)\sqrt{V_\Delta}\,,\; v_{1-\frac{\alpha}{2}}(R^2)\sqrt{V_\Delta}\right]$

Under [CRE], the corresponding interval is

$$\left[z_{\frac{\alpha}{2}}\sqrt{V_\Delta}\,,\; z_{1-\frac{\alpha}{2}}\sqrt{V_\Delta}\right]$$

↗ Li, Ding & Rubin (2018) showed that the [CRE] interval is never less in length than under $\mathcal{M}_{\varepsilon}$.

⇒ Not taking re-randomization at the time of the analysis leads to conservative confidence bounds for $\hat{\Delta}$.

× Remark: When the effect is additive, R^2 is the squared multiple correlation between X and Y(0). As noted in Li, Ding & Rubin (2018), "when ε is close to zero, [...] the asymptotic sampling variance $V_\Delta(p)\left(1-(1-v_{p,\varepsilon})R^2\right)$ reduces to $V_\Delta(p)\,(1-R^2)$, which is identical to the asymptotic sampling variance of the regression-adjusted estimator under the CRE (see Lin (2013)). Therefore, rerandomization accomplishes covariate adjustments in the design stage, whereas regression accomplishes covariate adjustment in the analysis stage" ■

II.3 Rerandomization with stratified designs

Stratified sampling was introduced in Section I. A natural way to extend rerandomization to stratified designs is to rerandomize stratum per stratum until a criterion is reached in all strata. Define

[finite] stratum k [∞]

- $\Delta_k^n = \frac{1}{n_k} \sum_{X_i = k} \{Y_i(1) - Y_i(0)\}$; $\Delta_k^\infty = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = k)$ $= \mu_{1k} - \mu_{0k}$

$\Delta^n = \sum_{k=1}^K \hat{\pi}_k \Delta_k^n$; $\Delta^\infty = \sum_{k=1}^K \hat{\pi}_k \Delta_k^\infty$

where $\hat{\pi}_k = \frac{n_k}{n}$; n_k = # observations in stratum k

- $\hat{\Delta} = \sum_{k=1}^K \hat{\pi}_k \hat{\Delta}(k)$ with $\hat{\Delta}(k) = \frac{1}{n_{1k}} \sum_{S_i = k} \omega_i Y_i - \frac{1}{n_{0k}} \sum_{S_i = k} (1-\omega_i) Y_i$

= DM estimator in k-th stratum

$\hat{\Delta}_X = \sum_{k=1}^K \hat{\pi}_k \hat{\Delta}_X(k)$ with $\hat{\Delta}_X(k) = \frac{1}{n_{1k}} \sum_{S_i = k} \omega_i X_i - \frac{1}{n_{0k}} \sum_{S_i = k} (1-\omega_i) X_i$

$S_i = k \Leftrightarrow$ unit i belong to stratum k

X_i = covariate on which balance is required.

(slight change of notation from section I)

- $p_k = \frac{n_{1k}}{n_k}$

- $V_{X,k} = \frac{\sigma_{X,k}^2}{p_k(1-p_k)}$; $V_{\Delta X,k} = \frac{c_{X0,k}}{1-p_k} + \frac{c_{X1,k}}{p_k}$

- $V_{\Delta,k}^\infty = \frac{\sigma_{1,k}^2}{p_k} + \frac{\sigma_{0,k}^2}{1-p_k}$ [∞]

- $V_{\Delta,k} = \frac{p_k}{1-p_k} \sigma_{0,k}^2 + \frac{1-p_k}{p_k} \sigma_{1,k}^2 + 2 \rho_{01,k} \sigma_{0,k} \sigma_{1,k}$ [finite]

- $R_k^2 = \frac{V_{\Delta X,k} V_{X,k}^{-1} V_{X\Delta,k}}{V_{\Delta,k}}$; $R_{\infty,k}^2 = \frac{V_{\Delta X,k} V_{X,k}^{-1} V_{X\Delta,k}}{V_{\Delta,k}^\infty}$

We consider a rerandomization procedure where each stratum is rerandomized independently, based on the stratum-specific Mahalanobis distances. For a stratum k, specify a threshold ε_k, and compute $\frac{\hat{\Delta}_X^2(k)}{\text{var}\, \hat{\Delta}_X(k)}$.

Accept the stratum k assignment vector if this quantity is less than ε_k. In the multivariate case, compute $\hat{\Delta}_X^t(k) \,\text{cov}^{-1} \hat{\Delta}_X(k)\, \hat{\Delta}_X(k)$. For $\varepsilon = (\varepsilon_1, \dots, \varepsilon_K)$, put

$$\mathcal{M}_\varepsilon = \left\{ (\omega_1, \dots, \omega_n) \mid \frac{\hat{\Delta}_X^2(k)}{\text{var}\, \hat{\Delta}_X(k)} \leq \varepsilon_k \;\; \forall k \right\}$$

Stratum-specific Mahalanobis distance

Wang, Wang, Liu (2021) only

Then

[finite] $\sqrt{n} \begin{pmatrix} \hat{\Delta} - \Delta^n \\ \hat{\Delta}_X \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V)$

where $V = \sum_{k=1}^K \hat{\pi}_k \left(\begin{array}{c|c} V_{\Delta,k} & V_{\Delta X,k} \\ \hline V_{X\Delta,k} & V_{X,k} \end{array} \right)$

[∞] $\sqrt{n} \begin{pmatrix} \hat{\Delta} - \Delta^\infty \\ \hat{\Delta}_X \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V^\infty)$

where $V^\infty = \sum_{k=1}^K \hat{\pi}_k \left(\begin{array}{c|c} V_{\Delta,k}^\infty & V_{\Delta X,k} \\ \hline V_{X\Delta,k} & V_{X,k} \end{array} \right)$

(compare with p. 20)

Under some general conditions, Wang, Wang, Liu (2021) established for the finite population framework that

$$\sqrt{n}\,(\hat{\Delta} - \Delta^n) \,|\, \mathcal{M}_\varepsilon \sim \sqrt{\sum_{k=1}^{K} \hat{\pi}_k V_{\Delta,k}(1-R_k^2)}\; \mathcal{N}(0,1) + \sqrt{\sum_{k=1}^{K} \hat{\pi}_k V_{\Delta,k} R_k^2}\; L^k_{p,\varepsilon_k} ,$$

K fixed, $n_k \to \infty$ $\forall k$

where L^k_{p,ε_k} are independent RVs $\sim L_{p,\varepsilon_k}$.

The asymptotic variance of $\sqrt{n}\,(\hat{\Delta} - \Delta^n)$ under $\mathcal{M}_\varepsilon$ is $\sum_{k=1}^{K} \hat{\pi}_k V_{\Delta,k} \{ 1 - (1 - v_{p,\varepsilon_k}) R_k^2 \}$.

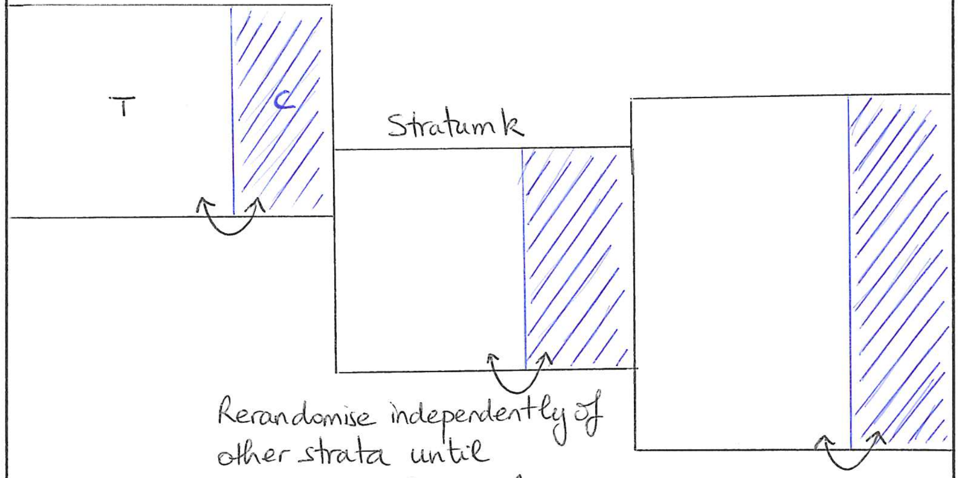
x Remark: These expressions should be compared with the asymptotic distribution of the aggregated difference in means estimator under [ST]:

$$\sqrt{n}\,(\hat{\Delta} - \Delta^n) \xrightarrow{d} \mathcal{N}\Big(0, \sum_{k=1}^{K} \hat{\pi}_k V_{\Delta,k}\Big) .$$ ■

Alternatively, we may consider an overall Mahalanobis distance $\hat{\Delta}_X^t \,\mathrm{cov}^{-1}\hat{\Delta}_X \;\; \hat{\Delta}_X$ and accept the full allocation $(w_1, \dots, w_n)$ if this quantity is less than some value ε.

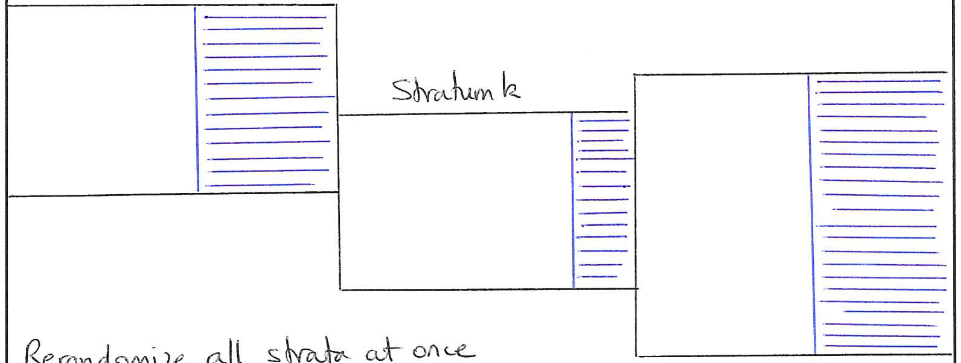
↖ Here, strata are all rerandomized at one, and all accepted at the same time, which differs from the previous rerandomization scheme.

[Scheme 1] Stratum-Specific Distance



Rerandomise independently of other strata until $\hat{\Delta}_X^t(k)\; \mathrm{cov}^{-1}\hat{\Delta}_X(k)\; \hat{\Delta}_X(k) \leq \varepsilon_k$

[Scheme 2] Overall Distance



Rerandomize all strata at once until $\hat{\Delta}_X^t \;\mathrm{cov}^{-1}\hat{\Delta}(k)\; \hat{\Delta}_X \leq \varepsilon$.

Wang, Wang, Liu (2021) showed that for identical thresholds $\varepsilon_1 = \dots = \varepsilon_K = \varepsilon$ in [Scheme 1], the asymptotic variance of $\sqrt{n}\,(\hat{\Delta} - \Delta^n)$ is smaller than for [Scheme 2].

Specifically,

Scheme 1

$$\sum_{k=1}^{K} \hat{\pi}_k V_{\Delta,k} \left\{ 1 - (1 - \upsilon_{p,\varepsilon}) R_k^2 \right\}$$

Scheme 2

$$\leqslant V_\Delta \left\{ 1 - (1 - \upsilon_{p,\varepsilon}) \right\} R^2$$

where $\left(\begin{array}{c|c} V_\Delta & V_{\Delta X} \\ \hline V_{X\Delta} & V_X \end{array} \right) = \sum_{k=1}^{K} \hat{\pi}_k \left(\begin{array}{c|c} V_{\Delta,k} & V_{\Delta X,k} \\ \hline V_{X\Delta,k} & V_{X,k} \end{array} \right)$

and $R^2 = \dfrac{V_{\Delta X} V_X^{-1} V_{X\Delta}}{V_\Delta}$

[Appendix A]

x Consider first $\hat{\Delta}_{agg}$ with $W \sim [ST]$.

Mean $\mathbb{E}\hat{\Delta}_{agg} = \sum_k \hat{\pi}_k \mathbb{E}\hat{\Delta}(k)$; $\mathbb{E}\hat{\Delta}(k) = \mu_{1k} - \mu_{0k}$

Variance $\text{var}\,\hat{\Delta}_{agg} = \sum_k \hat{\pi}_k^2 \,\text{var}\,\hat{\Delta}(k)$

with $\text{var}\,\hat{\Delta}(k) = \dfrac{\sigma_{1k}^2}{n_{1k}} + \dfrac{\sigma_{0k}^2}{n_{0k}}$

$$\text{var}\,\hat{\Delta}_{agg} = \sum_k \frac{n_k}{n} \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{n_{1k}} + \frac{\sigma_{0k}^2}{n_{0k}} \right)$$
$$= \frac{1}{n} \sum_k \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k} \right) \qquad \blacksquare$$

x We turn our attention to $\hat{\Delta} = \frac{1}{n_t} \sum W_i Y_i - \frac{1}{n_c} \sum (1-W_i) Y_i$

Mean $W \sim [CRE]$

$$\mathbb{E}\hat{\Delta} = \frac{n}{n_t} \underbrace{\mathbb{E}(WY)}_{\substack{\shortparallel \\ \mathbb{E}W\,\mathbb{E}Y(1) \text{ since } W \perp Y(1) \\ = \frac{n_t}{n}\mu_1}} - \frac{n}{n_c} \mathbb{E}[(1-W)Y] = \mu_1 - \mu_0$$

Variance: $\text{var}\,\hat{\Delta} = \dfrac{\text{var}\,Y(1)}{n_t} + \dfrac{\text{var}\,Y(0)}{n_c}$

with $\sigma_j^2 = \text{var}\,Y(j) = \sum_k \hat{\pi}_k \sigma_{jk}^2 + \sum_k \hat{\pi}_k (\mu_{jk} - \mu_j)^2$

$\blacksquare$

[Appendix B]

Consider $\hat{\Delta} = \frac{1}{n_t} \sum_{i=1}^{n} w_i Y_i - \frac{1}{n_c} \sum_{i=1}^{n} (1-w_i) Y_i$,

where the $w_i \sim [ST]$.

Here, $\{Y_i(0), Y_i(1)\} \perp w_i \mid X_i = k$, $k=1,\dots,K$,
but not unconditionally.

- <u>Mean</u> $\mathbb{E}\hat{\Delta} = \frac{n}{n_t} \mathbb{E} w Y(1) - \frac{n}{n_c} \mathbb{E}(1-w) Y(0)$

with $\mathbb{E} w Y(1) = \mathbb{E}_X \underbrace{\mathbb{E}\, w Y(1) \mid X} = \sum_k \hat{\pi}_k p_k \mu_{1k}$.

since $\mathbb{E}\, w Y(1) \mid X=k$

$$= \mathbb{E}(w \mid X=k)\; \mathbb{E}(Y(1) \mid X=k)$$

$$= \frac{n_{1k}}{n_k} \mu_{1k} = p_k \mu_{1k}$$

& similarly for $\mathbb{E}(1-w)Y(0) = \sum_k \hat{\pi}_k (1-p_k) \mu_{0k}$

$$\Rightarrow \mathbb{E}\hat{\Delta} = \sum_k \hat{\pi}_k \left(\frac{n p_k}{n_t} \mu_{1k} - \frac{n(1-p_k)}{n_c} \mu_{0k} \right)$$

- <u>Variance</u> Put $Z_k := \sum_{X_i=k} \left\{ \frac{w_i Y_i}{n_t} - \frac{(1-w_i) Y_i}{n_c} \right\}$

$\nwarrow n_k$ terms

$Z_1, \dots, Z_K$ are independent, and $\hat{\Delta} = \sum_{k=1}^{K} Z_k$.

$$\text{var}\, Z_k = \frac{n_{1k}}{n_t^2} \sigma_{1k}^2 + \frac{n_{0k}}{n_c^2} \sigma_{0k}^2$$

$$\Rightarrow \text{var}\,\hat{\Delta} = \sum_k \left(\frac{n_{1k}}{n_t^2} \sigma_{1k}^2 + \frac{n_{0k}}{n_c^2} \sigma_{0k}^2 \right) \quad \blacksquare$$

$\frac{n_{0k}}{n_c^2} \rightarrow = (1-p_k) \hat{\pi}_k \left(\frac{n}{n_c}\right)^2 \frac{1}{n}$

$\frac{n_{1k}}{n_t^2} \hookrightarrow = \frac{n_{1k}}{n_k} \frac{n_k}{n} n \frac{1}{n_t^2} = p_k \hat{\pi}_k \left(\frac{n}{n_t}\right)^2 \frac{1}{n}$

[Appendix C]

- $\text{var}\, \hat{\Delta}_{agg} = \mathbb{E}\, \text{var}\left(\hat{\Delta}_{agg} \mid \{n_{1k}, n_{0k}\}\right) + \text{var}\, \mathbb{E}\left(\hat{\Delta}_{agg} \mid \{n_{1k}, n_{0k}\}\right)$

$\swarrow$ p. 6 $\qquad\qquad$ $\swarrow$ p. 6

$$= \mathbb{E}\left(\frac{1}{n} \sum_{k=1}^{K} \hat{\pi}_k \left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k} \right) \right)$$

$$= \frac{1}{n} \sum_{k=1}^{K} \pi_k \left(\frac{\sigma_{1k}^2}{p_k} + \frac{\sigma_{0k}^2}{1-p_k} \right)$$

$$= \text{var}\left(\sum_{k=1}^{K} \hat{\pi}_k (\mu_{1k} - \mu_{0k}) \right)$$

$$= \frac{1}{n^2} \sum_{k,\ell} (\mu_{1k} - \mu_{0k})(\mu_{1\ell} - \mu_{0\ell}) \times \text{Cov}(n_k, n_\ell)$$

where $\text{Cov}(n_k, n_\ell) = -n \pi_k \pi_\ell$

$((n_1, \dots, n_K) \sim \text{Mult}(\pi_1, \dots, \pi_K))$

& $\text{var}\, n_k = n \pi_k (1-\pi_k)$.

$$= -\frac{1}{n} \sum_{k \neq \ell} (\mu_{1k} - \mu_{0k})(\mu_{1\ell} - \mu_{0\ell}) \pi_k \pi_\ell + \frac{1}{n} \sum_k (\mu_{1k} - \mu_{0k})^2 \pi_k (1-\pi_k)$$

$$= -\frac{1}{n} \sum_{k,\ell} (\mu_{1k} - \mu_{0k})(\mu_{1\ell} - \mu_{0\ell}) \pi_k \pi_\ell + \frac{1}{n} \sum_k \pi_k (\mu_{1k} - \mu_{0k})^2$$

$$= -\frac{1}{n} \left[\sum_k (\mu_{1k} - \mu_{0k}) \pi_k \right]^2 + \frac{1}{n} \sum_k \pi_k (\mu_{1k} - \mu_{0k})^2.$$

$= \frac{1}{n} \text{var}\, \Delta(X)$, where $\Delta(k) := \mathbb{E}(Y(1) - Y(0) \mid X=k)$. $\blacksquare$

- $\text{var}\,\hat{\Delta}$ is derived like in [Appendix A] with $\hat{\pi}_k$ replaced by the population proportions π_k $\downarrow$

$$= \frac{1}{n_t} \left(\sum_k \pi_k \left(\sigma_{1k}^2 + (\mu_{1k} - \mu_1)^2 \right) \right) + \frac{1}{n_c} \left(\sum_k \pi_k \left(\sigma_{0k}^2 + (\mu_{0k} - \mu_0)^2 \right) \right)$$

$\blacksquare$

[Appendix D]

- $$\mathbb{E}\hat{\Delta} = \mathbb{E}\,\mathbb{E}\,\hat{\Delta} \mid \{n_k\}$$

↘ p.8

$$= \sum_{k=1}^{K} \left[\mathbb{E}\left(\hat{\pi}_k \frac{n p_k}{\underset{\uparrow}{n_t}}\right) \mu_{1k} - \mathbb{E}\left(\hat{\pi}_k \frac{n(1-p_k)}{n_c}\right) \mu_{0k} \right]$$

Here n_t is random since $n_t = \sum_{k=1}^{K} p_k n_k$

$$\mathbb{E}\, n_t = n\left(\sum_{k=1}^{K} p_k \pi_k\right)$$

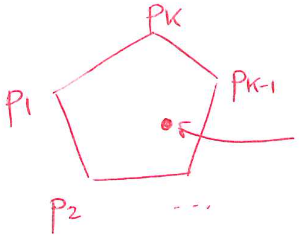
$$\mathbb{E}\left(\overset{\overset{\hat{\pi}_k}{\shortparallel}}{\frac{n_k}{\cancel{n}}} \cancel{n} \frac{p_k}{n_t} \right) = p_k \mathbb{E}\left(\frac{n_k}{n_t}\right)$$

$$= p_k \left(\frac{\mathbb{E}\, n_k}{\mathbb{E}\, n_t} + O(n^{-1}) \right)$$

$$= \frac{p_k \pi_k}{\sum_{\ell=1}^{K} p_\ell \pi_\ell} + O(n^{-1}) .$$

$$\Rightarrow \mathbb{E}\hat{\Delta} = \sum_{k=1}^{K} \pi_k \left(\underset{\updownarrow}{\frac{p_k}{\sum_\ell p_\ell \pi_\ell}} \mu_{1k} - \underset{\updownarrow}{\frac{1-p_k}{\sum_\ell (1-p_\ell)\pi_\ell}} \mu_{0k} \right)$$

Compare these coefficients with the value 1; the reference point for unbiasedness.



$\Sigma p_\ell \pi_\ell$ lies in the interior of the convex hull of $\{p_1, \cdot, p_K\}$ $\Rightarrow$ all coefs $\frac{p_k}{\Sigma p_\ell \pi_\ell} \neq 1$ unless all p_k are equal.