Cluster RCT Estimators Technical Document

Miratrix

2022-11-16

Introduction

This document contains the technical details for the identified estimators for estimating average effects for Blocked, Cluster-Randomized RCTs.

Once an estimand is identified, a researcher must decide how to estimate it, given the observed data, and obtain uncertainty estimates (standard errors) for those estimates. We call the ATE estimators "effect estimators" and the uncertainty estimators "standard error estimators." In the following sections, we systematically go through four general classes of effect estimators, discussing variants of implementing each. The four general classes are linear modeling estimators, multilevel (random effects) estimators, aggregation-and-regression estimators, and design-based estimators.

Because blocking is so prevalent for cluster randomized trials, we often discuss blocked cluster-randomized trials as a first order, leaving simple cluster randomized trials as a special case, even though it increases some of the technical detail and complexity. When there is no blocking, one can simply assume a single block that holds all the clusters, in most cases. This will collapse some estimator variants (variants that aggregate across the districts differently, for example) into a single one. We will note departures from this as we go along.

Notation-wise, we have individuals i (e.g., students) in clusters j (e.g., schools), nested in blocks k (e.g., districts or sites). Our outcome is then Y_{ijk}, and treatment assignment is an indicator Z_{jk}.

Before we look at estimators, let's spend a moment thinking about estimands. In particular, for each individual we have $Y_{ijk}(0)$ and $Y_{ijk}(1)$, with an individual treatment effect of $\tau_{ijk} = Y_{ijk}(1) - Y_{ijk}(0)$. Technically, this is the treatment impact on an individual from treating the entire cluster of said individual, which would include spillover effects and so forth within the cluster. For any individual, we can never see both potential outcomes, and thus we will never know τ_{ijk}, regardless of how we assign treatment.

Similarly, for each cluster we have the average treatment effect in that cluster as

$$\beta_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}(1) - Y_{ijk}(0).$$

where N_{jk} is the size of cluster jk. The β_{jk} are the average treatment effect of the individuals in cluster jk. Similar to the τ_{ijk}, we have no ability to estimate a cluster's β_{jk} because we treat the entire cluster, or not. This is very different from multisite randomized trials, where we get to see some of the individuals in a site treated and some not treated, allowing for (a perhaps noisy) estimate of the site's average treatment impact.

Because of this, a cluster-randomized experiment can be thought of as an individually randomized experiment where the individuals are the clusters themselves. In fact, this view is prevalent in much of the design based causal inference literature. It is also the intuition behind why we might think of a very large (in terms of N) experiment as quite small (in terms of J, the total number of clusters).

Within a district k we have different possible estimands of interest. Similar to Miratrix et al. (2021), we can have the simple average of the β_{jk} (the cluster-average effect) of

$$\beta_k = \sum_j \beta_{jk}/J_k,$$

or the person-weighted average of the β_{jk} (the person-average effect) of

$$\beta_k = \frac{1}{N_k} \sum_j N_{jk}\beta_{jk},$$

where $N_k = \sum_j N_{jk}$ is the total number of students in district k.

The latter is the average effect across all individuals in site k. The former is the average effect of the clusters, regardless of their size, in site k.

We can then weight these β_k across the districts in different ways: a simple average, giving the district-average of the district-level estimands, regardless of district size, a cluster-weighted average, weighing each district by the number of clusters in it, or a person-weighted average, weighing each district by the number of people in it.

We argue it is not typically sensible to, for example, calculate a person-weighted average of cluster-average impacts. The weighting at level 3 should probably be aligned with the weighting at level 2. Furthermore, we generally think it is a rare instance where a simple average across the districts is a desired quantity. That all being said, there are many weighting options one might select. The main ones are:

1. Person-weighted: Average impact for all people.
2. Cluster-weighted: The average of the cluster-average impacts
3. District average of person-weighted: Each district has an average impact on its individual students, and we might want the average of that.
4. District average of cluster-weighted: This would be the simple average of each district's average impact across clusters.

Covariate adjustment

Cluster randomized experiments are notoriously difficult to power. To improve precision researchers frequently turn to level 1 and level 2 covariates to explain some of the outcome variation and thus reduce standard errors of an average effect estimator. In this document we generally focus on the no-covariate case for clarity, the covariate-adjusted models typically parallel that discussion.

In particular, most of the various estimation strategies, in particular the linear regression models, explicitly allow for covariate adjustment: simply add in the baseline covariates into the regression. Aggregation estimators, which work with cluster averages, can also allow for covariate adjustment; in these cases individual-level covariates are averaged just like the outcomes, and then used in the aggregation estimators as level two covariates.

The design-based estimators generally have covariate-based sister estimators that can be shown to generalize the special no-covariate case. E.g., Shochet et al (2020) and (2021) show how design-based approaches, rooted in randomization, can be expressed as a weighted linear regression, which naturally then allows for covariate adjustment. We will use this work showing equivalence to use weighted regression for our design based estimators.

Uncertainty estimation and framing

Uncertainty estimation, implicitly or explicitly, will depend on the population model used. For example, we could focus entirely on the randomization that was conducted, holding individuals, clusters, and sites as fixed (i.e., finite). This fully finite-sample view is the most well-known for design-based inference; for this general context, see Middleton and Aronow (2015) or Schochet (2015).

Alternatively, we could model the districts as sampled from some infinite population. Within the districts, we could again model the clusters as sampled from some district-specific infinite populations. Or we can take the entire collection of clusters within the district as being fixed, once we sample the district. Finally, within the clusters, the individuals could then be viewed finite, or not.

Instead of infinite superpopulations, finite superpopulations can also be considered; this is the framing taken up by (CITE Abadie etc). Here we would have to employ a finite population correction: if we have most of our districts in our sample, for example, then our uncertainty based on the representativeness of our districts must be low.

Each of these modeling frameworks points towards different kinds of uncertainty we need to account for. The largest factor is whether districts are considered random, or not. If they are, then we need to account for the uncertainty of whether the districts we have in our sample are representative, in terms of their true average treatment impact, of their parent population. If they are not, then we do not need to account for this uncertainty. Also, if we are willing to assume there is no treatment variation by district, then there is no generalization uncertainty to account for.

Generally modeling the clusters as randomly sampled, or the individuals as randomly sampled, will have less impact on uncertainty estimation. This is partly due to the so-called "correlation of potential outcomes problem," which is that we cannot directly identify treatment heterogeneity at the individual (or cluster, given cluster treatment assignment) level, and thus have to assume, in effect a "worst case" uncertainty that corresponds to a constant treatment effect. This is essentially equivalent to assuming an infinite population at the individual or cluster level, as a conservative bound on the standard error.

Acknowledgements

Parts of this document, in particular some notation and mathematical formulation, are taken from the technical supplement to the Hunter et al. paper on PUMP.

Linear Regression Estimators

One of the most common estimation strategies for cluster RCTs in the social sciences is linear regression. For a two-level context (no districts), a natural choice is to fit a linear regression of the outcome onto the treatment indicator (and some individual or cluster-level control variables) and then use cluster robust standard errors (CRVE).

In particular, we might fit the simplest model of

$$Y_{ij} = \alpha + \beta Z_j + \epsilon_{ij}$$

to the student-level data.

Here the individual nature of the data will naturally target a person-average estimand, as larger clusters will have larger weight. In particular, the above will boil down to a simple difference in means estimate of $\bar{Y}_1 - \bar{Y}_0$.

Given that this model is estimating the mean of all treated units minus the mean of all control, we might expect it to be unbiased. Unfortunately, it is not, as discussed in Middleton and Aronow (2015); we give an overview of the argument below, in Section .

For inference, handling the clustering could be considered tricky. First, we cannot include fixed effects at the cluster level, as these would be completely confounded with treatment assignment. (We can have them at the district level, however, as discussed in the following sub-section.) We thus rely on cluster-robust standard errors, which view the clusters as being sampled from a larger super-population: we are thus doing super-population inference. For precision improvement, we could include cluster-level (or individual-level) covariates. Cluster-level covariates predictive of outcome will be particularly beneficial.

Linear regression with district fixed effects

The prior two-level model can naturally be extended to three levels by simply including fixed effects for district. We give some notation for this to connect to other regression estimators discussed below. Define K district dummies $S_{q,ijk}$, $q = 1, \ldots, K$ with $S_{q,ijk} = 1$ if individual ijk is in site k; in other words, $S_{q,ijk} = 1\{q = k\}$. The simplest OLS working model for estimating an impact would then arguably be

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \beta Z_{jk} + \epsilon_{ijk},$$

where the α_q are the K fixed effects for the districts. This model has $K + 1$ coefficients in total. The covariance of the errors would be a block-diagonal matrix, one block per cluster (not district). We exclude an overall intercept to have individual fixed effects for each district and no reference district.

One advantage of the fixed effects model, compared to a model with treatment by district interactions, is that the single impact parameter reduces the degrees of freedom used by the model, which could improve asymptotic inference in smaller studies. Given how few clusters exist in many education experiments, this may indeed be worthwhile.

For simplicity, we can leave the dummy variables implicit, simply writing

$$Y_{ijk} = \alpha_k + \beta Z_{ijk} + \epsilon_{ijk}.$$

Interestingly, for point estimation this model is the same as the fixed effect linear regression model used for a multisite trial, with the districts as sites. We know from work in that area that this means the estimated $\widehat{\beta}$ is nominally "precision-weighted" in that each district will be weighted proportional to $1/N_k p_k(1 - p_k)$ where p_k is the proportion of individuals treated in district k. For a multisite trial, these weights are the precisions of estimating the average treatment effect within the each site, assuming homoskedasticity. For a blocked cluster-randomized trial, these are no longer proportional to precision as they do not capture the uncertainty due to the cluster assignment within district. Departures can be especially severe if the clusters are unequally sized, and/or differently so across districts. Thus the targeted estimand of this estimator does not have a clean interpretation.

Linear regression with treatment by district interactions

We can extend the above model to estimate impacts within each district as so:

$$\begin{aligned} Y_{ijk} &= \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \sum_{q=1}^{K} \beta_k S_{q,ijk} Z_{jk} + \epsilon_{ijk} \\ &= \alpha_k + \beta_k Z_{jk} + \epsilon_{ijk}. \end{aligned}$$

The treatment-by-district interaction terms means each $\hat{\beta}_k$ will be the estimated impact from the simple 2-level model discussed above. In particular, each $\hat{\beta}_k$ is targeting the average impact for students in cluster k. We end up with K estimates, $\hat{\beta}_k$, $k = 1, \ldots, K$, in total, that would then need to be averaged.

Again using cluster-robust standard errors, we would obtain standard errors of $\widehat{SE}_k$ for each $\hat{\beta}_k$, which would be taken from the diagonal of the variance-covariance matrix coming out of the cluster-robust standard error estimation. Packages such as `estimatr` (Blair et al., 2022) will implement this.

To average we need to make explicit choices about how to weight the elements. All estimators will be of the form

$$\hat{\beta} = \sum_{k=1}^{K} w_k \hat{\beta}_k,$$

where the weights w_k (with, generally, $\sum_k w_k = 1$) control how we aggregate the district-specific effect estimates.

If we wanted to tend towards the individual-average effect we would weight as

$$\hat{\beta} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_k,$$

where N_k are the number of individuals in district k.

We could also weight the $\hat{\beta}_k$ equally: this would give us the average person-average impact across the districts. If we are wanting to weight each cluster equally, we would weight each district by the number of clusters it has, as follows:

$$\hat{\beta} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_k, \tag{1}$$

where J_k are the number of clusters in district k.

This last weighting schemes is somewhat suspect: because the $\hat{\beta}_k$ are estimating the average treatment impact for the individuals in site k, the cluster-weighted expression (Equation 1) is going to weight each district by the number of clusters, but each district estimate is the person-average treatment effect. The final implied estimand is not easily considered. If we truly want the average cluster impact, we would need to ensure the $\hat{\beta}_k$ are aligned with our final district weighting step, i.e., are estimating the cluster average impacts. This could be done by adding weights to the regression, or via other methods discussed below.

Regardless of weighting, the standard error for the overall $\hat{\beta}$ can be calculated using the weights used for averaging. Let $w = (w_1, \ldots, w_K)$ be the weights given to the β_k. Then the standard error for the overall average would be

$$SE(\hat{\beta}) = \left(w' \widehat{\Sigma} w \right)^{1/2},$$

with $\widehat{\Sigma} = diag(\widehat{SE}_1^2, ..., \widehat{SE}_K^2)$. Because the randomization happens within district independently, the standard errors for the individual impacts are also independent, in principle. One could also simply subset the relevant rows and columns of the $\widehat{\Sigma}$ variance-covariance matrix that correspond to the $\hat{\beta}_k$ from the original regression package and use that instead of the forced diagonal.

Hierarchical linear models (aka random effects models)

Multilevel modeling explicitly models the hierarchical nature of the data and allows for estimation of the degree of variation in outcomes at each level (and variation in impacts in the three-level case) as well as the overall ATE. Multilevel modeling is also probably the most common tool used to analyze data from CRT experiments in education contexts. As a lens on typical practice, we use the models identified in the popular PowerUp! power calculator (Dong & Maynard, 2013). This tool explicitly notes different couplings of design and modeling choices, and provides specific uncertainty formulas for several contexts.

Cluster RCT, no blocking by district

For the two-level case, PowerUp! suggests a random intercept model, with a single coefficient for average treatment effect. This assumes a constant treatment effect for all schools, and we can also include school covariates.

The simplest model for estimating impacts is:

$$\begin{aligned} Y_{ij} &= \theta_{0,j} + \epsilon_{ij} \\ \theta_{0,j} &= \alpha + \beta Z_j + u_{0,j} \end{aligned}$$

with distributions:

$$\begin{aligned} u_{0,jk} &\sim N\left(0, \tau_0^2\right) \\ \epsilon_{ij} &\sim N\left(0, \sigma^2\right). \end{aligned}$$

The reduced form is

$$Y_{ij} = \alpha + \beta Z_j + u_{0,j} + \epsilon_{ij}$$

The standard error of the treatment effect estimate is given by:

$$SE(\hat{\beta}) = \sqrt{\frac{\mathrm{ICC}_2(1-R_2^2)}{\bar{T}(1-\bar{T})J} + \frac{(1-\mathrm{ICC}_2)(1-R_1^2)}{\bar{T}(1-\bar{T})J\bar{n}}}.$$

The degrees of freedom are

$$\mathrm{df}_m = J - g_1 - 2.$$

In R, fitting this multilevel model would be a `lmer()` formula along the lines of:

```
Yobs ~ 1 + Z + (1 | S.id)
```

Whether the estimated β will correspond to a person-weighted or cluster-weighted estimand is unclear. On one hand, we have cluster random effects, and these could pick up any treatment heterogeneity across clusters, implying a cluster-average impact. On the other hand, similar to the adaptive nature of FIRC (Miratrix et al., 2021), we might find the larger clusters drive estimation, leading to more of a person-weighted estimand.

The modeling options multiply once we allow for three levels, as we then can potentially model cross-site means and impact variation using either fixed or random effects at level 3. PowerUp! identifies two estimators for this context, which they call fixed and random effects. The random effect model allows the treatment coefficient to vary across sites (implying a superpopulation of sites). We discuss these two, along with a simpler model, in the following.

District effects with no treatment variation

The simplest extension to the two-level MLM is to simply add fixed effects for district. This assumes a constant treatment effect across all districts and all schools within district. We extend the above to

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + \epsilon_{ijk} \\ \theta_{0,jk} &= \alpha_k + \beta Z_{jk} + u_{0,jk} \end{aligned}$$

with distributions:

$$\begin{aligned} u_{0,jk} &\sim N(0, \tau_0^2) \\ \epsilon_{ijk} &\sim N(0, \sigma_m^2)\,. \end{aligned}$$

The reduced form is

$$Y_{ijk} = \alpha_k + \beta Z_{jk} + u_{0,jk} + \epsilon_{ijk}$$

This model is analogous to the fixed effect regression with cluster robust standard errors. The key difference is we are modeling the error structure with a cluster random effect, rather than using robust standard errors.

One can also use random effects at the district level instead of fixed. The only difference is putting a distribution of $\alpha_k \sim N(\alpha, \sigma_\alpha^2)$. If there is no treatment by covariate interaction, the random intercept version will be virtually the same as the fixed effect version. For further discussion of this, see the discussion of the Random Intercept, Constant Coefficient (RICC) model as compared to the multi-site fixed effect model in Miratrix et al. (2021).

Fixed blocks interacted with treatment, random effects for clusters

The next way to move to three levels from two with multilevel modeling is to interact the fixed effects with treatment, allowing for cross-district impact heterogeneity. This model would then be analogous to the interacted fixed effect linear model in Section .

The fixed effect model for estimating impacts on outcome m is given by:

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + \epsilon_{ijk} \\ \theta_{0,jk} &= \alpha_k + \beta_k Z_{jk} + u_{0,jk} \end{aligned}$$

with distributions:

$$\begin{aligned} u_{0,jk} &\sim N(0, \tau_0^2) \\ \epsilon_{ijk} &\sim N(0, \sigma_m^2)\,. \end{aligned}$$

The α_k are estimated as coefficients to dummy variables for district membership. This is equivalent to an infinite prior placed on their distribution.

The reduced form is

$$Y_{ijk} = \alpha_k + \beta_k Z_{jk} + u_{0,jk} + \epsilon_{ijk}$$

The standard error of the treatment effect estimate of $\hat{\beta} = \frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_k$, assuming equal sized districts and clusters, is given by:

$$SE(\hat{\beta}) = \sqrt{\frac{\mathrm{ICC}_2(1-R_2^2)}{\bar{T}(1-\bar{T})JK} + \frac{(1-\mathrm{ICC}_2-\mathrm{ICC}_3)(1-R_1^2)}{\bar{T}(1-\bar{T})JK\bar{n}}}.$$

The degrees of freedom are

$$\mathrm{df}_m = K(J-2) - g_2.$$

Under different weighing schemes, the standard error formula will be scaled a bit differently—the weights would give an effective sample size depending on the weights.

The interacted model technically assumes no variation of impacts within schools, and no variation within the district level (so the school-level impacts within district are all considered the same). If there is variation, this model will implicitly average across the variation; the uncertainty could be impacted depending on the amount of heterogeneity, although if the variation of the heterogeneity is around 0.20^2 in effect size units, the impact of this is generally negligible.

The R model is a `lmer` call with formula

```
Yobs ~ 0 + Z * D.id - Z + (1 | S.id)
```

The overall treatment effect is then the average of the treatment by district interaction terms. All the questions about weighting and averaging from Section apply. Here, what the estimates $\hat{\beta}_k$ are estimating is less clear, however, due to the random cluster intercepts.

Full random effects model

In the fully random effects model we have random intercepts and random impact effects for districts, random intercepts for schools, and assumed constant effects for schools within a district.

The model for estimating impacts is given by:

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + r_{ijk} \\ \theta_{0,jk} &= \alpha_k + u_{0,jk} \\ \alpha_k &= \alpha + \beta_k Z_k + w_{0,k} \\ \beta_k &= \beta + w_{1,k} \end{aligned}$$

with reduced form:

$$Y_{ijk} = \alpha + \left(\beta + w_{1,k}\right) Z_{jk} + w_{0,k} + u_{0,jk} + r_{ijk}$$

and distributions:

$$\begin{aligned} u_{0,jk} &\sim N\left(0, \tau_0^2\right) \\ \begin{pmatrix} w_{0,k} \\ w_{1,k} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \eta_0^2 & \kappa_m^w \eta_0 \eta_1 \\ \kappa_m^w \eta_1 \eta_0 & \eta_1^2 \end{pmatrix}\right) \\ r_{ijk} &\sim N\left(0, \sigma_m^2\right). \end{aligned}$$

The standard error of the treatment effect estimate is given by:

$$SE(\hat{\beta}) = \sqrt{\frac{\mathrm{ICC}_3 \omega_3}{K} + \frac{\mathrm{ICC}_2(1 - R_2^2)}{\bar{T}(1-\bar{T})JK} + \frac{(1 - \mathrm{ICC}_2 - \mathrm{ICC}_3)(1 - R_1^2)}{\bar{T}(1-\bar{T})JK\bar{n}}}.$$

The degrees of freedom are often taken as $\mathrm{df}_m = K - 1$.

The R model formula for the above is

```
Yobs ~ 1 + Z + (1 | S.id) + (1 + Z | D.id)
```

Discussion

The fixed effect model would put no distribution on the w, but would still have a random effect at level 2 for the intercept of each cluster within the district; this implies a finite sample of districts. There are other modeling options within this framing, such as an extension of the FIRC model (Bloom et al., 2017) with only a distribution on $w_{1,k}$, but leaving the $w_{0,k}$ as fixed effects. For the multisite comparison to this choice, see Miratrix et al. (2021).

Aggregation Estimators

Another regression based approach is to first aggregate the data within each cluster and then fit a linear model to the aggregated data. Begin with the average cluster outcome ($\bar{Y}_{jk}$), defined in terms of the individual potential outcomes as

$$\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ij} = Z_{jk}\bar{Y}_{jk}(1) + (1 - Z_{jk})\bar{Y}_{jk}(0).$$

Here Z_{jk} is an 0/1 treatment indicator for cluster j, and we observe either the average of the cluster's treatment or control potential outcomes.

We then regress the aggregated outcomes onto their treatment indicators. At this point, we have reduced our blocked cluster-randomized trial to a multisite trial where the cluster average outcomes are our new outcome and the districts are the sites. Everything from Miratrix et al. (2021) then applies in terms of different estimators we could in principle use, including design-based estimators and multilevel modeling.

For example, we could fit a linear model to the aggregates such as

$$\bar{Y}_{jk} = \alpha_k + \beta Z_{jk} + \epsilon_{jk}$$

This is simply a regression with district-level fixed effects, analogous to the individual-level regression case.

We could weight this regression in either of two ways: weight each observation equally, or weight each observation by N_{jk}. In the former, we are treating clusters equally, and targeting a cluster-average effect. In the latter, we are targeting a person-average effect.

Alternatively, one could include district by treatment interactions, allowing different districts to have different average ATEs, and then these district ATEs can be aggregated to estimate the overall ATE, following the individual-level regression with treatment by district interaction case. In particular, we could weight the district-specific impact estimates equally, by number of clusters in the district, or by total individuals in the district. We would want this weighting to correspond to the weights placed on the individual clusters in the regression itself.

Differing cluster sizes will likely induce heteroskedasticity in a linear regression model: the average of the N_{jk} individuals in cluster jk will generally be more uncertain in smaller clusters than larger due to differing sample sizes. In this case, we can use heteroskedastic robust standard errors to account for this.

If we wanted to treat the districts themselves as being drawn from a super-population, we could instead use cluster-robust standard errors at the district level. This would then inflate our uncertainty if there was evidence of impact variation across district.

Design-based estimators

Design-based estimators, although not as common as linear regression or multi-level modeling, are usually directly and explicitly tied to an estimand of interest (Schochet, 2015). Design-based estimators focus on the random assignment mechanisms and the sampling mechanism as the sources of uncertainty, rather than specifying linear models with random residual

components. They are generally considered to rely on weaker assumptions than other approaches; e.g., homoskedasticity assumptions, or assumptions of a random normal residual, are typically not needed. Some good overview papers on design-based inference for clustered randomized trials are Schochet (2013); Schochet et al. (2021) and Middleton and Aronow (2015).

Design-based estimators often work with the cluster averages (sometimes adjusted with covariates) and are thus related to the aggregation estimators. Indeed, in some cases the point estimate of these and the aggregated linear regressions will be the same. It is typically only the methods for calculating the standard error (and the standard error estimates themselves) that will differ. We first focus on a single district k, and then extend to multiple districts. For a single district k, the simplest cluster-average estimator for the average impact would be

$$\hat{\beta}_{k-cluster} = \frac{1}{J_{k1}} \sum_{j=1}^{J_k} Z_{jk} \bar{Y}_{jk} - \frac{1}{J_{k0}} \sum_{j=1}^{J_k} (1 - Z_{jk}) \bar{Y}_{jk},$$

where J_{k1} and J_{k0} are the number of clusters assigned to treatment or control status, respectively. The $\bar{Y}_{jk}$ are the observed mean outcomes of the clusters. This is a version of the aggregation approach, above. This estimator targets the average of the cluster average effects, in district k.

The person-average estimator would weight the clusters by size, giving

$$\hat{\beta}_{k-person} = \frac{1}{N_{k1}} \sum_{j=1}^{J_k} Z_{jk} N_{jk} \bar{Y}_{jk} - \frac{1}{N_{k0}} \sum_{j=1}^{J_k} (1 - Z_{jk}) N_{jk} \bar{Y}_{jk},$$

where $N_{k1} = \sum Z_{jk} N_{jk}$ are the number of individuals assigned to treatment, and N_{k0} is defined similarly. These estimators can either work with the cluster aggregates or the individual data: they boil down to cluster average outcomes (although when adjusting for individual-level covariates there are some minor technical details, discussed below).

Extending to multiple districts

The classic design-based estimator approach will estimate impacts within each district k, and then average those estimates across districts. For example, a district-size-weighted average of the person-average $\hat{\beta}_k$ would give the estimated effect for the average person across the sample. This estimator would be

$$\hat{\beta} = \sum_{k=1}^{K} \frac{1}{N} N_k \hat{\beta}_{k-person},$$

where N_k is the total size (in individuals) of district k and N is the total number of individuals.

For the district-average of person averages, we might weight differently again:

$$\hat{\beta} = \sum_{k=1}^{K} \frac{1}{K} \hat{\beta}_{k-person}.$$

This would be the average of the district average impacts. We might use this if we were interested in how much a random district would shift its population, on average, if given treatment. Other weightings options are also possible, such as the district-average of the cluster-average impact.

Uncertainty estimation

As discussed in the introduction, uncertainty estimation depends on whether the population model used is finite or super-population. Designed based inference will specifically target different models in the standard error formula.

The fully finite context is the most well-known for design-based inference; see Middleton and Aronow (2015) or Schochet (2015). If we view the clusters as sampled from a fixed set of superpopulations, one for each district, with individuals within the clusters considered fixed (so we are sampling fixed *sets* of individuals), we can use Schochet et al. (2021) for uncertainty estimation. Other models are possible; see Pashley and Miratrix (2021a, 2021b) for further discussion in the context of individual randomization with blocking.

In general, if we consider the districts as fixed (regardless of what is happening inside of them), we can use design-based uncertainty estimators within each district (with weights to align with the ATE estimator), and then aggregate based on district size, again using weights depending on our target estimand and how we calculated our ATE. In particular, the standard errors in this case would simply be the aggregation of the district-level standard errors (similar to the interacted linear model case):

$$SE(\beta) = \left(\sum_{k=1}^{K} w_k^2 SE(\beta_k)^2 \right)^{1/2},$$

where the w_k are the weights used in calculating an ATE of $\beta = \sum_k w_k \beta_k$, with $\sum w_k = 1$. Because randomization is independent within district, the above formula holds directly.

The Middleton and Aronow estimators

Middleton and Aronow note that the design based estimators given above for the individual-average effect are actually biased. They then propose a class of unbiased estimators for the average treatment effect for cluster randomized trials. With multiple districts, we can apply their approach to each district and then average, pooling the standard error as discussed above (assuming we are holding the districts as fixed).

To see the bias concern, consider the expected value of the prior simple difference estimator:

$$\begin{aligned}
E[\hat{\beta}_k] &= E\left[\frac{1}{N_{k1}}\sum_{j=1}^{J_k} Z_{jk}N_{jk}\bar{Y}_{jk} - \frac{1}{N_{k0}}\sum_{j=1}^{J_k}(1-Z_{jk})N_{jk}\bar{Y}_{jk}\right] \\
&= \sum_{j=1}^{J_k} E\left[\frac{1}{N_{k1}}Z_{jk}\right]N_{jk}\bar{Y}_{jk}(1) - \sum_{j=1}^{J_k} E\left[\frac{1}{N_{k0}}(1-Z_{jk})\right]N_{jk}\bar{Y}_{jk}(0) \\
&= \sum_{j=1}^{J_k} E\left[\frac{1}{N_{k1}}Z_{jk}\right]Y^T_{jk}(1) - \sum_{j=1}^{J_k} E\left[\frac{1}{N_{k0}}(1-Z_{jk})\right]Y^T_{jk}(0),
\end{aligned}$$

where $Y^T_{jk} = N_{jk}\bar{Y}_{jk}$ is the sum of the outcomes for cluster j in district k.

The problem is that

$$E[\frac{1}{N_{k1}}Z_{jk}] \neq \frac{1}{N_k},$$

because the N_{k1} term is random: if we end up with a large cluster in treatment, it is larger, and if we end up with a small one, it is smaller.

Alternatively, consider that the mean of the treatment-side in the above is the sum of all outcomes in the treatment group divided by total number of students in the treatment group. Then, the expected value of this across randomizations is

$$E\left[\frac{1}{N_{k1}}\sum_{j=1}^{J_k} Z_{jk}N_{jk}\bar{Y}_{jk}\right] = E\left[\frac{\sum_j N_{jk}Z_{jk}\bar{Y}_{jk}(1)}{\sum_j N_{jk}Z_{jk}}\right] \neq \frac{E\left[\sum_j N_{jk}Z_{jk}\bar{Y}_{jk}(1)\right]}{E\left[\sum_j N_{jk}Z_{jk}\right]} = \bar{Y}_k(1).$$

In other words, because the number of units in the treatment group is random, the expectation does not go through, and we do not end up getting an unbiased estimate of the average of all the individuals' treatment-side potential outcomes as would be needed for an overall unbiased estimate. That being said, the bias of this estimator is often not large, and the estimator is clearly targeting the average treatment impact of the students in the sample, not the average of the cluster average impacts.

One fix is to make an estimator that has the *expected* number of treated or control units as the denominators, i.e., replace N_{k1} with $E[N_{k1}] = pN_k$, with p_k the fixed proportion of clusters in district k assigned to treatment. This is a type of Horvitz-Thompson estimator, taking the cluster totals as the potential outcomes, and is often very unstable.

Middleton and Aronow propose to instead use a Raj difference estimator where we adjust all the outcomes to try and stabilize the consequence of different sized clusters. In particular, define

$$U_{jk} = Y^T_{jk} - \alpha\left(N_{jk} - \frac{N_k}{J}\right)$$

where α is a prior estimate of the regression coefficient from regressing Y^T_j on N_j. Technically α needs to be known in advance; more on this regression coefficient later.

The term $N_{jk} - \frac{N_k}{J}$ is the difference between the size of cluster j and the average cluster size in district k; we are essentially adjusting our outcome totals to what they would be if the cluster were average in size.

We then estimate impact as

$$\hat{\beta} = \frac{1}{J_{1k}} \sum_j Z_{jk} U_{jk} - \frac{1}{J_{0k}} \sum_j (1 - Z_{jk}) U_{jk}.$$

In expectation, the adjustments are 0, meaning the overall estimand is unbiased.

To estimate α we propose to simply regress the cluster totals onto cluster sizes across all the districts in a single step. For any given district, the value of α will be primarily driven by the other districts and thus we can mostly avoid the reuse issues discussed in (Middleton and Aronow, 2015). Alternatively, out-of-sample estimation can be used, where α is estimated via the other districts for each district k; this would be more computationally intensive.

To get estimated standard errors we have, again taking from (Middleton and Aronow, 2015),

$$\widehat{SE}(\hat{\beta})^2 = \frac{J_k^2}{N_k^2} \frac{J_k}{J_k - 1} \left[\frac{S_{k1}^2}{J_{k1}} + \frac{S_{k0}^2}{J_{k0}} \right],$$

with sample standard deviations for treatment group z of

$$S_z^2 = \frac{1}{J_{kz} - 1} \sum_{j : Z_j = z} \left(U_{jk}^T - \overline{U_{jkz}^T} \right)^2.$$

S_z^2 is sample standard deviation of the adjusted cluster totals and $\overline{U_{jkz}^T}$ is the average cluster total in treatment group z. This is simply an appropriately scaled Neyman standard error.

References

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). estimatr: Fast estimators for design-based inference. R package version 1.0.0.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. 6(1-2):312.

Miratrix, L. W., Weiss, M. J., and Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, 14(1):270–308.

Schochet, P. Z. (2013). Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference. *Journal of Educational and Behavioral Statistics*, 38(3):219 – 238. Good illustration of using the Neyman model for understanding other procedures.

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., and Kautz, T. (2021). Design-based ratio estimators and central limit theorems for clustered, blocked rcts. *Journal of the American Statistical Association*, pages 1–12.