The benefits of probability proportional to size sampling in cluster randomized experiments

Michael J. Higgins

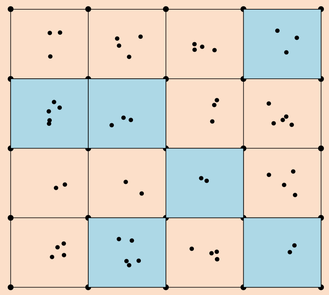
Postdoctoral Fellow, Department of Politics
Princeton University

Introduction

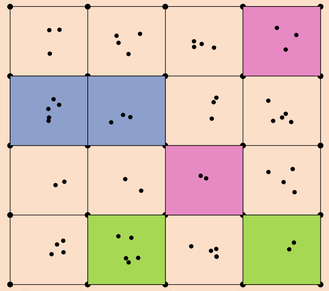
Cluster-randomized experiments (CREs), are performed when treatment is randomized across clusters (groups) of units of interest instead of across the individual units. The number of clusters in the experiment and the number of observations obtained within each cluster are typically restricted by a budget constraint. The assignment of treatment to clusters makes analysis difficult; under the Neyman-Rubin Causal Model, no estimator an average treatment effect currently exists that is both unbiased and invariant to location shifts in potential outcomes. We show that, when the quantity of interest is the population average treatment effect (PATE), such estimators can be obtained by initially sampling clusters with probability proportional to size (PPS).

Cluster randomized experiments procedure

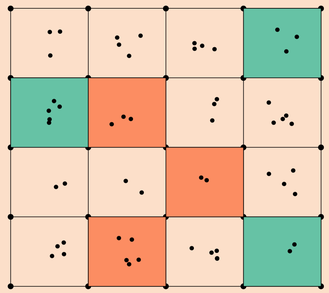
1) Sample clusters from a population (16 clusters, 56 units in total)



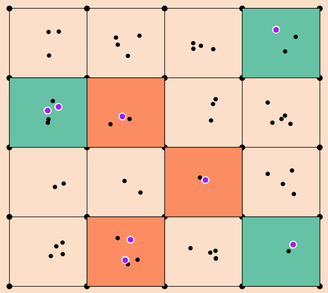
2) Form blocks of sampled clusters



3) Assign treatment to clusters



4) Sample units from clusters



Common settings for CREs

CREs are used when assigning treatment to units is infeasible or when necessary to avoid interference between treatment groups.

- **Developing countries:** Villages receive treatment.
- **Education:** Classrooms receive treatment.
- **Medical Trials:** Treatment given to clinics/medical practices.

Quantity of interest

We are interested in inference on:

$$PATE = \frac{1}{n}\sum_{c=1}^{b}\sum_{k=1}^{n_c} y_{kc1} - \frac{1}{n}\sum_{c=1}^{b}\sum_{k=1}^{n_c} y_{kc0}$$

A location shift involves adding a constant α to all potential outcomes:

$$y^*_{kci} = y_{kci} + \alpha$$

Estimators are invariant: values do not change when outcomes are shifted.

y_{kci}: Potential outcome of unit k in cluster c under treatment i.
$n/b/n_c$: Number of units/clusters/units in cluster c.

PPS sampling without replacement

A probability-proportional-to size-sample of size s drawn without replacement (PPSWOR) is any such sample satisfying:

$$P(\text{Sample cluster } c) = sn_c/n.$$

However, this condition does not uniquely define a sampling scheme. Generally, joint probabilities of being sampled must also be specified:

$$\pi_{cc'} \equiv P(\text{Sample clusters } c, c').$$

Sunter (1986) provides an efficient method for drawing a PPSWOR sample, which is implemented in the `R` package `SunterSampling`.

Estimators

Under PPSWOR sampling, the quantity:

$$\hat{\mu}_i = \sum_{c=1}^{b} \frac{S_c T_{ci}}{\#T_i} \sum_{k=1}^{n_c} \frac{y_{kci} S_{kc}}{s_c}$$

is an unbiased estimator of the population mean of units under treatment i, μ_i, that is a LINEAR function of potential outcomes. Hence, $\widehat{PATE} = \hat{\mu}_1 - \hat{\mu}_0$ is an unbiased and location-invariant estimator of the PATE.

The variance and covariance of this estimator are:

$$\begin{aligned}\text{Var}(\hat{\mu}_i) &= \mathbb{E}\left(\frac{1}{\#T_i}\right)\left(\sigma^2_{i,bet} + \sum_{c=1}^{b}\frac{n_c}{n}\left(\sigma^2_{c,i,with}\right)\right) \\ &\quad + \mathbb{E}\left(\frac{1}{\#T_i} - 1\right)\left(\sum_{c=1}^{b}\sum_{c'\neq c}\frac{\pi_{cc'}\mu_{ct}\mu_{c'i}}{s(s-1)} - \mu_i^2\right), \\ \text{cov}(\hat{\mu}_1, \hat{\mu}_0) &= \frac{1}{s(s-1)}\sum_{c=1}^{b}\sum_{c'\neq c}^{b}\pi_{cc'}\mu_{c1}\mu_{c'0} - \mu_1\mu_0.\end{aligned}$$

Where $\sigma^2_{i,bet}$ denotes the across-cluster variance for treatment i, and $\sigma^2_{c,i,with}$ denotes the finite sample variance within cluster c.

Current practice

- **Sampled clusters:** Either simple random sample or non-random selection.
- **Blocking:** Matched pair on size of cluster.

Under simple random sampling of clusters and complete randomization of treatment, the unbiased Horvitz-Thompson estimator is not generally invariant under location shifts which can artificially inflate variances. Blocking does not fix this problem.

Current practice is for researchers to use matched pair blocking and use either difference-in-means (DIM) or Des-Raj (Middleton and Aronow 2014) estimators for the PATE. DIM will generally be biased if outcomes are correlated with cluster sizes. Des-Raj requires the inclusion of an additional tuning parameter; estimation of this parameter will bias the estimator.

n_{si}: number of units that receive treatment i.
S_c/S_{kc}: Cluster/unit sampling indicators
$T_{ci}/\#T_i$: Treatment indicator/number of clusters receiving treatment i.
s_c: Number of units/clusters/units in cluster c.

Simulation results

We simulate potential outcomes under a model with several cluster-level covariates and unit-level covariates, where treatment effects depend on cluster sizes. We consider three designs:

1. Clusters sampled using PPSWOR, block on all cluster-level covariates.
2. Clusters sampled using SRS, block on all cluster-level covariates.
3. Clusters sampled using SRS, block only on size.

Our simulation includes 16 distinct clusters with sizes varying between 30 and 600 units. 15 units are sampled from each cluster. Under each design, we perform 10,000 CREs.

PPSWOR sampling eliminates the bias seen when using a SRS of clusters. In this case, standard errors are slightly higher for PPSWOR sampling.

mjh5@princeton.edu