Design of (Cluster) Randomized Experiments: Costs and Benefits of Stratification and Pairing

UCSB Field Experiments Conference
Saturday, March 24th, 2012, 3.00-4.00pm

Guido Imbens - Harvard University
Outline

1. Mixed Advice in the Literature

2. Power Comparisons for t-statistic Based Tests

3. Expected Squared Error Calculations

4. Practical Advice for Cluster Randomized Experiments
1. Mixed Advice in the Literature

- Suppose we have $N$ units, we observe some covariates on each unit, and wish to evaluate a binary treatment.

- Should we randomize the full sample, or should we stratify the sample first, or even pair the units up?

**Recommendation In Literature:**

- In large samples, and if the covariates are strongly associated with the outcomes, definitely stratify or pair.

- In small samples, with weak association between covariates and outcomes, the literature offers mixed advice.
Quotes from the Literature

Snedecor and Cochran (1989, page 101) write, comparing paired randomization and complete randomization:

“If the criterion [the covariate used for constructing the pairs] has no correlation with the response variable, a small loss in accuracy results from the pairing due to the adjustment for degrees of freedom. A substantial loss may even occur if the criterion is badly chosen so that member of a pair are negatively correlated.”
Box, Hunter and Hunter (2005, page 93) also suggest that there is a tradeoff in terms of accuracy or variance in the decision to pair, writing:

“Thus you would gain from the paired design only if the reduction in variance from pairing outweighed the effect of the decrease in the number of degrees of freedom of the $t$ distribution.”
Klar and Donner (1997) raise additional issues that make them concerned about pairwise randomized experiments (in the context of randomization at the cluster level):

“We shown in this paper that there are also several analytic limitations associated with pair-matched designs. These include: the restriction of prediction models to cluster-level baseline risk factors (for example, cluster size), the inability to test for homogeneity of odds ratios, and difficulties in estimating the intracluster correlation coefficient. These limitations lead us to present arguments that favour stratified designs in which there are more than two clusters in each stratum.”
Imai, King and Nall (2009) claim there are no tradeoffs at all between pairing and complete randomization, and summarily dismiss all claims in the literature to the contrary:

“Claims in the literature about problems with matched-pair cluster randomization designs are misguided: clusters should be paired prior to randomization when considered from the perspective of efficiency, power, bias, or robustness.”

and then exhort researchers to randomize matched pairs.

“randomization by cluster without prior construction of matched pairs, when pairing is feasible, is an exercise in selfdestruction.”
How Do We Reconcile These Statements?

- Be careful and explicit about goals: precision of estimators versus power of tests.

- Be careful about estimands: population versus sample, average over clusters or average over individuals.
2. Power Comparisons for t-statistic Based Tests

The basic calculation underlying the concern with pairwise randomization is based on calculation of t-statistics.

Randomly sample $N$ units from a large population. Covariate $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$. We then draw another set of $N$ units, with exactly the same values for the covariates.

Conditional potential outcome distribution does not depend on $X_i$:

$$Y_i(0)|X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{and} \quad Y_i(1) = Y_i(0) + \tau$$

Completely randomized design ($\mathcal{C}$): randomly pick $N$ units to receive the treatment.

Pairwise randomized design ($\mathcal{P}$): pair the units by covariate and randomly assign one unit from each pair to the treatment.
The natural estimator under both designs is

\[ \hat{\tau}_{\text{dif}} = \overline{Y}_1^{\text{obs}} - \overline{Y}_0^{\text{obs}}. \]

Its distribution under the two designs is the same as well (because covariate is independent of outcomes):

\[ \hat{\tau}_{\text{dif}}|C \sim \mathcal{N}(\tau_{\text{SP}}, \frac{2 \cdot \sigma^2}{N}) \quad \text{and} \quad \hat{\tau}_{\text{dif}}|P \sim \mathcal{N}(\tau_{\text{SP}}, \frac{2 \cdot \sigma^2}{N}) \]
The natural estimator for the variance for the estimator given
the pairwise randomized experiment is

\[
\hat{\mathbb{V}}_P = \frac{1}{N \cdot (N - 1)} \sum_{i=1}^{N} (\hat{\tau}_i - \hat{\tau})^2 \sim \frac{2 \cdot \sigma^2}{N} \cdot \chi^2(N - 1)
\]

The variance estimator for the completely randomized design,
exploiting homoskedasticity, is

\[
\hat{\mathbb{V}}_C = \frac{2}{N} \left( \frac{(N - 1) \cdot s^2(0) + (N - 1) \cdot s^2(1)}{2N - 2} \right) \sim \frac{2 \cdot \sigma^2}{N} \cdot \chi^2(2 \cdot N - 2)
\]

Under the normality assumption both variance estimators are
independent of \( \hat{\tau}_{\text{dif}} \).

\[
\mathbb{E} \left[ \hat{\mathbb{V}}_P \right] = \mathbb{E} \left[ \hat{\mathbb{V}}_C \right] = \frac{2 \cdot \sigma^2}{N} \quad \mathbb{V} \left( \hat{\mathbb{V}}_P \right) = 2 \cdot \mathbb{V} \left( \hat{\mathbb{V}}_C \right) = \frac{8 \cdot \sigma^4}{N^2 \cdot (N - 1)}
\]


This leads to the t-statistics

\[ t_P = \frac{\hat{\tau}_{\text{dif}}}{\sqrt{\hat{\nabla}_P}}, \quad \text{and} \quad t_C = \frac{\hat{\tau}_{\text{dif}}}{\sqrt{\hat{\nabla}_C}}. \]

If we wish to test the null hypothesis of \( \tau = 0 \) against the alternative of \( \tau \neq 0 \) at level \( \alpha \), we would reject the null hypothesis if \( |t| \) exceeds the critical value \( c_\alpha \) (different for the two designs)

\[ c_P^\alpha = q^t_{1-\alpha/2}(N - 1), \quad c_C^\alpha = q^t_{1-\alpha/2}(2N - 2) \]
Proposition For any $\tau \neq 0$, and for any $N \geq 2$ the power of the test based on the t-statistic $t_C$ is strictly greater than the power based on the t-statistic $t_P$ if the covariate is independent of the potential outcomes.

(at $N = 1$ we cannot test the hypothesis without knowledge of the variances)

By extension, the power for the test based on the completely randomized design is still greater than the power based on the pairwise randomized experiment if the association between the covariate and the potential outcomes is weak, at least in small samples.

This is the formal argument against doing a pairwise (or by extension) a stratified randomized experiment if the covariates are only weakly associated with the potential outcomes.
Limitations

- Test comparison relies on normality. Without normality we cannot directly rank the power, and the actual size of the tests need not even be equal to the nominal size.

- Homoskedastic case is most favorable to completely randomized experiment (but features most often in power comparisons). In the case of heteroskedasticity, the loss in power for pairwise randomized experiment is less.
3. Expected Squared Error Calculations for Completely Randomized vs Stratified Randomized Experiments

Suppose we have a single binary covariate $X_i \in \{f, m\}$. Define

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)| X_i = x]$$

where the expectations denote expectations taken over the superpopulation.

$$\tau_{FS} = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)).$$

The estimand we focus on is the (super-)population version of the finite sample average treatment effect,

$$\tau_{SP} = \mathbb{E}[\tau_{FS}] = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau(X_i)]$$
Notation

\[ \mu(w, x) = \mathbb{E} [ Y_i(w) | W_i = w, X_i = x ], \]

\[ \sigma^2(w, x) = \mathbb{V} ( Y_i(w) | W_i = w, X_i = x ), \]

for \( w = 0, 1 \), and \( x \in \{ f, m \} \), and

\[ \sigma_{01}^2(x) = \mathbb{E} \left[ ( Y_i(1) - Y_i(0) - (\mu(1, x) - \mu(0, x)))^2 \middle| X_i = x \right], \]
Three Estimators: $\hat{\tau}_{\text{dif}}$, $\hat{\tau}_{\text{reg}}$, and $\hat{\tau}_{\text{strat}}$

First, simple difference:

$$\hat{\tau}_{\text{dif}} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}}$$

Second, use the regression function

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \beta \cdot 1_{X_i=f} + \epsilon_i.$$  

Then estimate $\tau$ by least squares regression. This leads to $\hat{\tau}_{\text{reg}}$.

The third estimator we consider is based on first estimating the average treatment effects within each stratum, and then weighting these by the relative stratum sizes:

$$\hat{\tau}_{\text{strat}} = \frac{N_{0f} + N_{1l}}{N} \cdot (\bar{Y}_{1f}^{\text{obs}} - \bar{Y}_{0f}^{\text{obs}}) + \frac{N_{0m} + N_{1m}}{N} \cdot (\bar{Y}_{1m}^{\text{obs}} - \bar{Y}_{0m}^{\text{obs}}).$$
Large (infinitely large) superpopulation.

We draw a stratified random sample of size $4N$ from this population, where $N$ is integer. Half the units come from the $X_i = f$ subpopulation, and half come from the $X_i = m$ subpopulation.

Two experimental designs. First, a completely randomized design ($\mathcal{C}$) where $2N$ units are randomly assigned to the treatment group, and the remaining $2N$ are assigned to the control group.

Second, a stratified randomized design ($\mathcal{S}$) where $N$ are randomly selected from the $X_i = f$ subsample and assigned to the treatment group, and $N$ units are randomly selected from the $X_i = m$ subsample and assigned to the treatment group.

In both designs the conditional probability of a unit being assigned to the treatment group, given the covariate, is the same: $\Pr(W_i = 1|X_i) = 1/2$, for both types, $x = f, m$. 
\[ V_s = \mathbb{E} \left[ (\hat{\tau}_{\text{dif}} - \tau_{\text{SP}})^2 \right | S] = \frac{q}{N} \cdot \left( \frac{\sigma^2(1, f)}{p} + \frac{\sigma^2(0, f)}{1 - p} \right) + \frac{1 - q}{N} \cdot \left( \frac{\sigma^2(1, m)}{p} + \frac{\sigma^2(0, m)}{1 - p} \right) \]

\[ V_c = \mathbb{E} \left[ (\hat{\tau}_{\text{dif}} - \tau_{\text{SP}})^2 \right | C] = q(1 - q)(\mu(0, f) - \mu(0, m))^2 + \frac{q\sigma^2(0, f)}{(1 - p)N} + \frac{(1 - q)\sigma^2(0, m)}{(1 - p)N} \]

\[ + q(1 - q)(\mu(1, f) - \mu(1, m))^2 + \frac{q\sigma^2(1, f)}{pN} + \frac{(1 - q)\sigma^2(1, m)}{pN} \]

\[ V_c - V_s = q(1 - q) \cdot \left( (\mu(0, f) - \mu(0, m))^2 + (\mu(1, f) - \mu(1, m))^2 \right) \geq 0 \]
Comment 1:

Stratified randomized design has lower expected squared error than completely randomized design.

Strictly lower if the covariate predict potential outcomes.

- True irrespective of sample size
Comment 2: For this result it is important that we compare the **marginal** variances, not **conditional** variances. There is no general ranking of the conditional variances

$$E[(\hat{\tau}_{\text{dif}} - \tau_{FS})^2 \mid Y(0), Y(1), X, \mathcal{C}] \quad E[(\hat{\tau}_{\text{dif}} - \tau_{FS})^2 \mid Y(0), Y(1), X, S]$$

It is possible that stratification leads to larger variances because of negative correlations within strata in a finite sample (Snedecor and Cochran quote). That is not possible on average, that is, over repeated samples.

There does not appear to be any way to exploit this possibility of negative correlations within strata, so it is largely of theoretical importance.

In practice it means that if the primary interest is in the most precise estimate of the average effect of the treatment, **stratification dominates complete randomization**, even in small samples.
**Comment 3:** Under a stratified design the three estimators \( \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{strat}}, \) and \( \hat{\tau}_{\text{dif}} \) are identical, so their variances are the same.

Under a completely randomized experiment, the estimators are generally different. In sufficiently large samples, if there is some correlation between the outcomes and the covariates that underly the stratification, the regression estimator \( \hat{\tau}_{\text{reg}} \) will have a lower variance.

However, for any fixed sample size, if the correlation is sufficiently weak, the variance of \( \hat{\tau}_{\text{reg}} \) will actually be strictly higher than that of \( \hat{\tau}_{\text{dif}} \).
Think through analyses in advance

Thus for *ex post* adjustment there is a potentially complicated tradeoff: in small samples one should not adjust, and in large samples one should adjust if the objective is to minimize the expected squared error.

If one wishes to adjust for differences in particular covariates, do so by design: randomize in a way such taht $\hat{r}_{dif} = \hat{r}_{reg}$ (e.g., stratify, or rerandomize).
4. Analytic Limitations of Pairwise Randomization

Compare two designs with $4N$ units.

- $N$ strata with 4 units each ($S$).
- $2N$ pairs with 2 units each ($P$).

What are costs and benefits of $S$ versus $P$?
Benefits of Pairing

- The paired design will lead to lower expected squared error than stratified design in finite samples. (similar argument as before.)

- In sufficiently large sample power of paired design will be higher (but not in very small samples, similar argument as before).
Difference with Stratified Randomized Experiments

- Suppose we have a stratum with size $\geq 4$ and conduct a randomized experiment within the stratum with $\geq 2$ treated and $\geq 2$ controls.

- Within each stratum we can estimate the average effect and its variance (and thus intraclass variance). The variance may be imprecisely estimated, but we can estimate it without bias.

- Suppose we have a stratum (that is, a pair) with 2 units. We can estimate the the average effect in each pair (with the difference in outcomes by treatment status), but we cannot estimate the variance.
Difference with Stratified Randomized Experiments (ctd)

- From data on outcomes and pairs alone we cannot establish whether there is heterogeneity in treatment effects.

- We can establish the presence of heterogeneity if we have data on covariates used to create pairs (compare “similar” pairs).

- Efficiency gains from going from strata with 4 units to strata with 2 units is likely to be small.
Recommendation

- Use small strata, rather than pairs (but not a big deal either way)

- Largely agree with Klar & Donner
5. Re-Randomization

Sometimes researchers randomize assignment to treatment, then assess the (im)balance the specific assignment would generate, and decide to re-randomize if the initial assignment failed to lead to sufficient balance.

What to make of that?

Re-randomization can improve precision of estimates and power of tests considerably, but needs to be done carefully to maintain ability to do inference.
Re-randomization is conceptually similar to completely randomized experiment:

Consider a sample of $2N$ units.

Randomize treatment to each unit by flipping a fair coin.

Re-randomize till the number of treated units is exactly equal to $N$.

This leads to the same design as randomly selecting $N$ units for assignment to treatment in a completely randomized experiment.
Formal Analysis of Re-Randomization

Suppose we have $2N$ units. We observe a $K$-vector of covariates $X_i$. Without taking into account the covariate values, $N$ units are randomly selected to receive the treatment, and the remaining units are assigned to the control group.

Calculate

$$
\overline{X}_0 = \frac{1}{N} \sum_{i:W_i=0} X_i, \quad \overline{X}_1 = \frac{1}{N} \sum_{i:W_i=1} X_i, \quad t_X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{s^2_{X,0}/N + s^2_{X,1}/N}}
$$

What to do if $|t_X|$ is large, if discovered before assignment is implemented?
Two Cases

- Decide *a priori* to randomize $M$ times, and implement assignment vector that minimizes some criterion e.g., minimize the maximum of the t-statistics for the $K$ covariates.

- Re-randomize until the criterion meets some threshold: e.g., with two covariates, until both t-statistics are below 1.
  
  (need to be careful here: the threshold should be feasible).

**Key:**

1. articulate strategy *a priori*, so randomization inference is possible.

2. Do **not** search over all assignments for optimal value for criterion because then there is little randomness left.
Cautionary Note

- Suppose with $2N$ units, $X_i$ earnings, $2N - 1$ units have $X_i \in [0, 10]$, and 1 unit has $X_i = 1000$.

- Minimizing t-statistic leads to one treatment group containing individual with $X_i = 1000$ and $N - 1$ individual with lowest earnings, and other group containing $N$ richest individuals after very richest individual.

- Irrespective of design estimation of ave effect is difficult.

- Rank-based tests may still have substantial power.

- Maybe remove outlier unit for estimation purposes.
6. Power Calculations for Unit-level Randomization

6.1 Power Calculations: Testing a Mean against Zero

Suppose we have a random sample $X_1, \ldots, X_N$ from a normal distribution with mean $\mu$ and variance $\sigma^2$.

We wish to test the null hypothesis $H_0: \mu = 0$, against $H_a: \mu \neq 0$.

We want the probability of a type I error to be less than $\alpha$ (e.g., $\alpha = 0.05$). We want the test to have power $\beta$, that is the probability of rejecting the null hypothesis, when the null is false with $\mu = \mu_0$, should be $\beta$. Let's say $\beta = 0.8$, for $\mu_0 = 0.1 \cdot \sigma$. 
We base the test on the t-statistic

\[ T = \frac{\bar{X}}{\sqrt{S^2_X/N}}, \]

where

\[ \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i, \quad \text{and} \quad S^2_X = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2. \]

So the question is now, given \( \alpha, \beta, \tau = \mu_0/\sigma \), what is the minimum sample size?
If the size of the test is $\alpha$, we reject the null hypothesis if

$$|T| \geq \Phi^{-1}(1 - \alpha/2),$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$
Now let us look at the probability of rejecting the null hypothesis when the true mean is \( \mu_0 = 0.1 \cdot \sigma \).

\[
\Pr \left( |T| > \Phi^{-1} \left( 1 - \alpha/2 \right) \right) \approx \Pr \left( T > \Phi^{-1} \left( 1 - \alpha/2 \right) \right)
\]

\[
\approx \Phi \left( -\Phi^{-1} \left( 1 - \alpha/2 \right) + \frac{\mu_0}{\sqrt{\sigma^2/N}} \right) \approx \beta
\]

leading to

\[
N = \left( \frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma} \right)^2
\]
For example, if $\alpha = 0.05$ (we test at the 5% level), $\mu_0/\sigma = 0.1$ (the effect is an increase of 10% of a standard deviation), and we wish to detect such an effect with probability 0.8 (probability of type II error should be less than $1 - \beta = 0.2$).

Then we need a sample size of

$$N = \left( \frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma_X} \right)^2$$

$$= \left( \frac{\Phi^{-1}(0.8) + \Phi^{-1}(0.975)}{0.1} \right)^2 = 784.89,$$

so the minimum sample size is 785.
6.2 Testing a Difference of Means

Random sample, $Y_1, \ldots, Y_N$, and a treatment indicator $W_1, \ldots, W_N$, sample size $N$.

Interest in hypothesis that the average treatment effect $\tau_{SP} = \mathbb{E}[Y_i(1) - Y_i(0)]$ is zero:

$$H_0: \tau_{SP} = 0, \quad \text{against} \quad H_a: \tau_{SP} \neq 0.$$  

The size of the test $\alpha$, power $\beta$ against an alternative that the average treatment effect is $\tau_{SP} = d \cdot \sigma$, $\sum_i W_i/N = 1/2$ is proportion of treated units.

We look for the minimum sample size $N = N_0 + N_1 = N_0 \cdot (1 + \gamma)$, as a function $N(\alpha, \beta, d, \sigma^2)$. 
**Result:** Total sample size required is

\[
N = 4 \cdot \left( \Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2) \right)^2 / d^2.
\]

For example, suppose we choose \( \alpha = 0.05 \) (test at 0.05 level), \( \tau/\sigma = 0.1 \) (looking for effect of 0.1 standard deviation), \( \beta = 0.8 \), (power of 0.8). Then

\[
N = \frac{\left( \Phi^{-1}(0.8) + \Phi^{-1}(0.975) \right)^2}{0.1^2 \cdot 0.5^2} = 3,139.6.
\]

Page 54-55 of Cohen give values for \( N_0 = N_1 = N/2 \) for this case for different values of \( \beta \) (power), \( \alpha \) (\( a_2 \) in Cohen’s book), and different values for \( d \).
7. Additional Design Issues for Clustered Randomization

Conflicting goals:

- Estimation and power for $\tau_\lambda$
- Estimation and power for $\tau_{FS}$

Additional inputs required for power analysis:

- intra cluster correlation
- distribution of cluster sizes $N_i$ (for power for $\tau_\lambda$)
- Relative cost of sampling units in same or different clusters
Bear in mind:

- $\tau_\lambda$ and $\tau_{FS}$ are both weighted averages of potential outcomes.

- In medical trials typically convenience samples (corresponding to indifference between weighted and unweighted averages).

- Finding evidence that either differs from zero is evidence of some causal effects.

**Recommendation:** Focus on estimation of $\tau_\lambda$, but focus power on $\tau_{FS}$. 
Design Choices that affect power

- Sampling ($S_i = 1$ if sampled, $S_i = 0$ otherwise)
  - Distribution of cluster sizes in sample $\Pr(S_i = 1|N_i)$
  - Sample sizes for each cluster $\Pr(n_i|S_i = 1, N_i)$

- Randomization
  - Probability of treatment assignment given cluster sizes
    $\Pr(W_i = 1|N_i, N_1, \ldots, N_M)$
**Sampling**

Suppose

- Clusters all large,

- Intracluster correlation substantial

**Recommendation**

- Sample equal (small) numbers from all clusters.

- Sample clusters proportional to cluster size.
Suppose intracluster correlation is zero (or small)

- No inferential cost from sampling many units from each cluster.

- Reduction in sampling cost from sampling many units from each cluster.
Randomization

• Stratify by cluster size, with many small strata.

Three benefits:

– General benefits from stratification on covariates (improving precision)

– Remove finite sample bias from estimator.

– Identifies effect size by cluster size.
Power Calculations for Cluster Randomized Experiments

• Focus power calculations on $\tau_{FS}$

• Use power calculations for unit-level randomization, using only combined variance for cluster averages.
• Number of units sampled per cluster is equal to $n$ for all clusters.

• Outcome variance within clusters is $\sigma^2_{\text{within}}$. Variance of true cluster means is $\sigma^2_{\text{between}}$.

• Effect size is $\tau$.

**Result:** Total number of clusters required is

$$M = 4 \cdot \left( \Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2) \right)^2 \cdot \left( \frac{\sigma^2_{\text{within}}}{n} + \sigma^2_{\text{between}} \right) \frac{\tau^2}{\tau^2}.$$ 

Required number of clusters does not change much if cluster size varies if we focus on power for $\tau_{FS}$ rather than $\tau_{\lambda}$.
Summary of Design Recommendations for Cluster Randomization

- Stratify or pair by cluster size (and possibly by powerful predictors of potential outcomes)
  - optimal for precision of estimation
  - potential loss of power small.

- Test null of no effect using statistic with power, based on randomization distribution.

- Estimate both $\tau_{FS}$ and $\tau_{\lambda}$