Week 2: Design Effects

POL-GA 3202
Quantitative Field Methods
Prof. Cyrus Samii
NYU Politics

September 21, 2016

Last week:

- Basic randomization theory, sample size determination, and power analysis.

This week:

- Complex designs that use clustering and stratification.
- Consequent “design effects” and their implications for sample size requirements and power.

Design effects

- Sample size determination and variance estimation under simple random sampling (SI) or completely randomized experiments was pretty straightforward.
- But these designs are not always, or even often, used in practice.
- Does that render the things we learned last week useless?

Design effects

- In sampling, the "design effect" for a statistic Z and design δ is,

$$D^2(Z,\delta) = \frac{\sigma^2_{\delta,Z}}{\sigma^2_{SI,Z}} = \frac{\text{Variance of } Z \text{ under design } \delta}{\text{Variance of } Z \text{ under } SI}$$

- For experiments or observational studies, the same definition can be carried through: design effects are the proportional effects on the variance of target statistics (e.g., regression coefficients).

Design effects

Design effects provide information needed for sample size adjustments:

- Recall sample size determination formula for $\bar{Y}$: $n = \frac{z^2_{\alpha/2}\sigma^2_{\bar{Y}}}{\mu^2_Y r^2}$.
- *Consistent* design choices affect only $\sigma^2_{\bar{Y}}$.
- By implication, for δ versus SI, we have

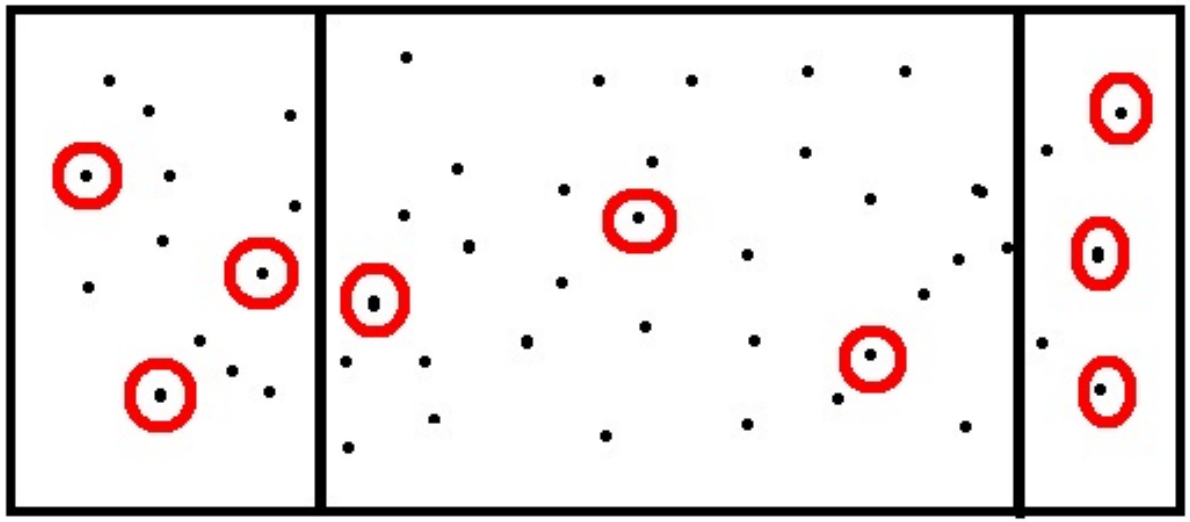
$$\frac{n_{0,\delta}}{n_{0,SI}} = \frac{(z^2_{\alpha/2}\sigma^2_{\delta,\bar{Y}})/(\mu^2_Y r^2)}{(z^2_{\alpha/2}\sigma^2_{SI,\bar{Y}})/(\mu^2_Y r^2)} = \frac{\sigma^2_{\delta,\bar{Y}}}{\sigma^2_{SI,\bar{Y}}} = D^2(\bar{Y},\delta).$$

- And so, $n_{0,\delta} = n_{0,SI}D^2(\bar{Y},\delta)$.
- Also, appropriate *s.e.* for a statistic from design δ is $D(\bar{Y},\delta)$ times the s.e. under SI.
- Same principles hold when comparing complex experiments to completely randomized experiments, for example.

Design effects

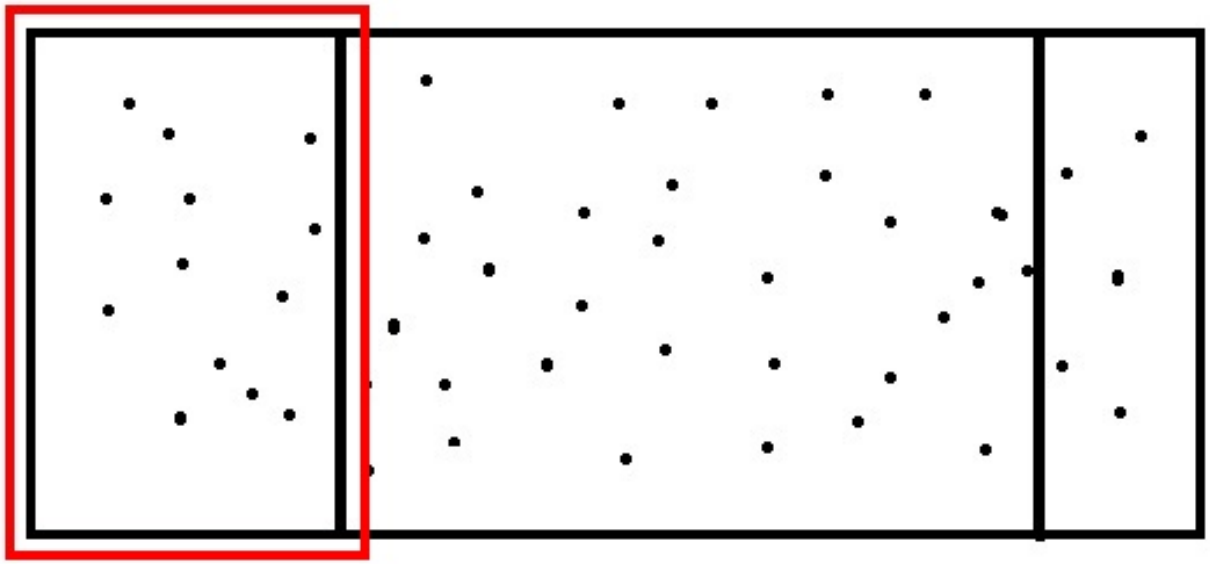
- Thus, we can use theoretical results from SI to get $n_{0,SI}$ and then scale that value up or down by $D^2(\bar{Y})$ to determine adequate sample size for any design.
- $D^2(\bar{Y}) \in (0, \infty)$.
- Some designs (e.g., stratified) can result in $D^2(\bar{Y}) < 1$, reducing necessary sample sizes relative to SI.
- Other designs (e.g., cluster) can result in $D^2(\bar{Y}) > 1$, increasing necessary sample sizes relative to SI.
- Where do we get $D^2(Z)$ values?
  - Theoretical results applied to real data.
  - Simulations using real data.

Designs and their effects: stratified sampling



- Motivated by,
  - Ensuring reliable estimates over strata.
  - Helping in administration of survey.
  - Allowing different sub-designs in different strata.
  - Gaining efficiency.
- Leads one to ask,
  - How should I allocate a sample over strata?
  - How many strata should I use?

Designs and their effects: cluster sampling



- Motivated by,
  - No pre-existing unit-level frame.
  - Reducing cost in administering survey.
- Leads one to ask,
  - How many clusters?
  - How many units per cluster?

Stratified sampling: the method

If intelligently used, stratification almost always results in a smaller variance for the estimated mean or total than is given by a comparable simple random sample.
(Cochran, 1977, p. 99).

- Partition U into L non-overlapping and exhaustive cells, $U_1, \ldots, U_L$, each of size $N_1, \ldots, N_L$.
- Take SRS of size n_h inside each.

Stratified random sampling: choices

- How should I create strata?
- How should I allocate over strata?
  - *Proportional to* N_h?
  - *Equal* $(n_1 = ... = n_L)$?
  - *"Optimal" for population inference*?
  - *"Optimal" for contrasts across strata*?

Stratified random sampling: estimation

- Estimate the population mean as follows:

$$\bar{Y}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \left(\frac{1}{n_h} \sum_{i \in h} Y_{ih} \right) = \sum_{h=1}^{L} W_h \bar{Y}_h, \text{ where } W_h = \frac{N_h}{N}.$$

- Then, $\mathrm{E}[\bar{Y}_{st}] = \mu_Y$, the population mean.
- In addition, $\mathrm{Var}[\bar{Y}_{st}] = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right)$, the weighted sum of within-stratum variances.

Stratified random sampling: design effects

- Design effect is given by,

$$D^2(\bar{Y}, ST) = \frac{\text{Var}_{ST}[\bar{Y}_{st}]}{\text{Var}_{SI}[\bar{Y}]} = \frac{\sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)}{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}$$

- Under proportional sampling within each stratum, $n_h = n(N_h/N)$, and the design effect reduces to,

$$D^2(\bar{Y}, ST_{prop}) = \frac{\sum_{h=1}^{L} W_h S_h^2}{S^2} = \frac{\sum_{h=1}^{L} W_h S_h^2}{\sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h (\mu_{Y_h} - \mu_Y)^2} \leq 1,$$

  where the expansion of the denominator is the ANOVA identity (Cochran, 1977, p. 100).

- Thus, ST_{prop} dominates SI in terms of efficiency.

Stratified random sampling: design effects

- It is possible to lose efficiency from stratification:
  - if the strata do not reduce within variation relative to overall variation, or
  - heterogenous strata are under-sampled relative to more homogenous strata.
- Typically, one gains from stratification though.
- Computing $D^2(\bar{Y}, ST)$ from a given sample requires obtaining an approximation for $\text{Var}_{SI}[\bar{Y}]$.
- Cochran (1977, pp. 136-8) provides a formula for a consistent estimator:

$$\widehat{\text{Var}}_{SI}[\bar{Y}] = \frac{N-n}{n(N-1)}\left[\frac{1}{N}\left(\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi}\right) - \bar{y}_{st} + s^2(\bar{y}_{st})\right].$$

Stratified random sampling: number of strata

- Cochran (1977, pp. 132-134) shows how returns diminish in the number of stratification cells for a given variable.
- Suppose you are interested in the mean of Y and you stratify on L well-placed bins over X. Let ρ be correlation between X and Y.

TABLE 5A.12

$V(\bar{y}_{st})/V(\bar{y})$ AS A FUNCTION OF L FOR THE LINEAR REGRESSION MODEL AND FOR SOME ACTUAL DATA

	Linear Regression Model				Data, Set		
	$\rho =$						
L	0.99	0.95	0.90	0.85	1	2	3
2	0.265	0.323	0.392	0.458	0.197	0.295	0.500
3	0.129	0.198	0.280	0.358	0.108	0.178	0.375
4	0.081	0.154	0.241	0.323	0.075	0.142	0.244
5	0.059	0.134	0.222	0.306	0.065	0.105	0.241
6	0.047	0.123	0.212	0.298	0.050	0.104	0.212
∞	0.020	0.098	0.190	0.277	—	—	—

		Type of Data		
Set	Data	x	y	Source
1	College enrollments	1952	1958	Cochran (1961)
2	City sizes	1940	1950	Cochran (1961)
3	Family incomes	1929	1933	Dalenius and Gurney (1951)

Stratified random sampling: “optimal allocation”

- “Optimal allocation” over strata tries to induce the most beneficial design effect.
- The optimum in this respect will differ depending on goals of the analysis:
  - Population level inference, or
  - Between-stratum contrasts.
- Bear in mind that such optima allow gains in precision at the cost of administrative or analytical convenience.

Stratified random sampling: "optimal allocation" for *population inference*

- Suppose a budget, $C = c_0 + \sum_{h=1}^{L} c_h n_h$.
- Choose $n_1, \ldots, n_L$ to minimize $\text{Var}\left[\bar{Y}_{st}\right]$ s.t. C.
- The solution yields (Cochran, 1977, p. 98),

$$\frac{n_h}{n} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} (N_h S_h / \sqrt{c_h})},$$

  increasing in N_h, S_h, decreasing in c_h.
- When c_h fixed over h, it drops out, and the allocation is proportional the product of population sizes and standard deviations ("Neyman allocation").
- Given fixed C, we can plug back into the C function to obtain

$$n = \frac{(C - c_0) \sum_{h=1}^{L} (N_h S_h / \sqrt{c_h})}{\sum_{h=1}^{L} (N_h S_h \sqrt{c_h})}$$

Stratified random sampling: "optimal allocation" for *between stratum contrasts*

- Suppose two strata, $h = 1, 2$, and you want, $\Delta_{12} = \bar{Y}_1 - \bar{Y}_2$.
- By independence across strata, $\text{Var}[\Delta] \approx \frac{S_i^2}{n_1} + \frac{S_2^2}{n_2}$ (omitting fpc's).
- Assuming budget, $C = c_0 = c_1 n_1 + c_2 n_2$, $\text{Var}[\Delta_{12}]$ smallest when,

$$\frac{n_1}{n} = \frac{S_1/\sqrt{c_1}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}} \text{ and } \frac{n_2}{n} = \frac{S_2/\sqrt{c_1}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}$$

- When many strata/contrasts, minimize average $\text{Var}[\Delta_{gh}]$ with,

$$n_h \propto \frac{S_h}{\sqrt{c_h}}$$

- Other approaches might be to minimize variance for "most important" contrasts, etc.

Stratified random sampling: choices

Recap:

- Stratification is generally beneficial.
- Stratification involves two choices: how to create them, and how to allocate over them.
- In making these choices, we trade between design effects and administrative and analytical convenience.

Stratified random sampling: choices

- How should I create strata?
  - More strata ensures a more representative sample.
  - Within-stratum homogeneity reduces design effects
  - But these benefits are diminishing in number of strata.
  - The higher the correlation between stratification variables and outcomes, the larger the efficiency gains.
  - You can use things like cluster analysis to create strata when you have lots of variables available.

Stratified random sampling: choices

- How should I create strata?
  - More strata ensures a more representative sample.
  - Within-stratum homogeneity reduces design effects
  - But these benefits are diminishing in number of strata.
  - The higher the correlation between stratification variables and outcomes, the larger the efficiency gains.
  - You can use things like cluster analysis to create strata when you have lots of variables available.
- How should I allocate over strata?
  - *Proportional to* N_h: self-weighting, and therefore analytically convenient; some efficiency gains.
  - *Equal* $(n_1 = ... = n_L)$: Not self-weighting, but administratively convenient; efficient for contrasts between strata, but not necessarily for aggregate inference.
  - *"Optimal" for population inference*: Minimizes variance for aggregate estimates, however may not be optimal from the perspective of contrasts across strata.
  - *"Optimal" for contrasts across strata*: Maximizes power of tests for equality across strata (power analysis), but may come at a cost for aggregate inference.

Stratified random sampling: other considerations

- Subpopulation analyses:
  - You can have different designs for each subpopulation.
  - That way, you maximize precision based on available info for each subpopulation.
- All of these principles carry over to experimental design, observational studies, etc.

Using Covariates to Design an Experiment

Consider a simple model:

$$Y_{di} = \alpha + \rho d + \gamma X_i + \lambda d * X_i + \varepsilon_i,$$

with $d \in \{0,1\}$ and $\text{Cov}(X_i, \varepsilon_i) = 0$. For each i we observe,

$$Y_i = \alpha + \rho D_i + \gamma X_i + \lambda D_i * X_i + \varepsilon_i,$$

with the support of D_i being $\mathscr{D} = \{0,1\}$. The true ATE is,

$$\beta = \text{E}[Y_{1i} - Y_{0i}] = \rho + \lambda \text{E}[X_i].$$

Using Covariates to Design an Experiment

Suppose we take a random sample, apply a completely randomized experiment, and take the difference in means:

$$\hat{\beta} = \text{mean}(Y_i|D_i = 1) - \text{mean}(Y_i|D_i = 0)$$

Then,

$$V_{\hat{\beta}} = \frac{\text{Var}[Y_{1i}]}{N_1} + \frac{\text{Var}[Y_{0i}]}{N_0} = \frac{(\gamma+\lambda)^2\sigma_X^2 + \sigma_\varepsilon^2}{N_1} + \frac{\gamma^2\sigma_X^2 + \sigma_\varepsilon^2}{N_0}$$

- Variation in X_i contributes to variation in $\hat{\beta}$.
- If we can control variance in X_i without introducing bias, our design will have more power.

Using Covariates to Design an Experiment

One idea: try to balance covariates as much as possible.

- Randomization introduces uncertainty about balance.
- Wouldn't it be better to maximize balance directly?
- This is a subject of some controversy.

Using Covariates to Design an Experiment

Recall our target estimand:

$$\beta = \mathrm{E}[Y_{1i} - Y_{0i}] = \rho + \lambda \mathrm{E}[X_i].$$

Now,

$$\begin{aligned}\hat{\beta} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[\alpha + \rho + (\gamma+\lambda)X_i + \varepsilon_i] - \frac{1}{N_0}\sum_{i=1}^{N}(1-D_i)(\alpha+\gamma X_i + \varepsilon_i) \\ &= \rho + \lambda\bar{X}_1 + \gamma(\bar{X}_1 - \bar{X}_0) + (\bar{\varepsilon}_1 - \bar{\varepsilon}_0)\end{aligned}$$

which implies

$$|\hat{\beta} - \beta| = \left|\lambda(\bar{X}_1 - \mathrm{E}[X_i]) + \gamma(\bar{X}_1 - \bar{X}_0) + (\bar{\varepsilon}_1 - \bar{\varepsilon}_0)\right|.$$

Minimizing $|\bar{X}_1 - \bar{X}_0|$ tends to reduce $|\hat{\beta} - \beta|$.

Using Covariates to Design an Experiment

Toy example where we need 2 treated, 2 control:

Unit	X	ε
A	10	2
B	8	5
C	2	5
D	0	2

- To minimize $\bar{X}_1 - \bar{X}_0$, assign A and D to same condition.
- Then regardless of which pair gets treatment

$$|\hat{\beta} - \beta| = |\lambda(5-5) + \gamma(5-5) + (2-5)| = 3.$$

- Deterministic assignment is biased.
- Also, it may not be the least biased deterministic assignment—e.g., assigning B and D to treatment yields $|\hat{\beta} - \beta| = |-\lambda - 2\gamma| < 3$ if, e.g., $\lambda = \gamma = .5$.
- Random assignment is unbiased, however.

Using Covariates to Design an Experiment

- But bias is not always the most important thing. Consider MSE.
- For the deterministic design, it is just $3^2 = 9$.
- But for randomized design:

Treated	$(\hat{\beta} - \beta)^2$
AB	$16[\lambda^2 + 4\gamma^2 + 4\lambda\gamma]$
AC	$\lambda^2 + 4\gamma^2 + 4\lambda\gamma$
AD	9
BC	9
BD	$\lambda^2 + 4\gamma^2 + 4\lambda\gamma$
CD	$16[\lambda^2 + 4\gamma^2 + 4\lambda\gamma]$
MSE:	$\frac{17}{3}[\lambda^2 + 4\gamma^2 + 4\lambda\gamma] + 3$

- It's quite possible for best balanced design to dominate randomized design in terms of MSE.
- E.g., with $\lambda = \gamma = .5$ MSE under randomization is 15.75.

Using Covariates to Design an Experiment

Randomization is not "optimal" relative to direct balancing:

- Classic debate between Basu, Rubin, et al. (1980).
- More recently: Bruhn & McKenzie (2009), Kasy (2012).

Using Covariates to Design an Experiment

Randomization is motivated by broader considerations:

1. Transparent and testable (post-randomization balance checks):
   - "...in this crooked world, he else can he avoid the charge of doctoring his own data?..." (Basu 1980, p. 594)
   - "...more difficult for an experimenter to 'cheat'..." (Gelman et al. 2004, p. 225).

Using Covariates to Design an Experiment

Randomization is motivated by broader considerations:

1. Transparent and testable (post-randomization balance checks):
   - "...in this crooked world, he else can he avoid the charge of doctoring his own data?..." (Basu 1980, p. 594)
   - "...more difficult for an experimenter to 'cheat'..." (Gelman et al. 2004, p. 225).
2. When X_i is vector valued, not clear how to assess imbalance. Different metrics may imply different rankings of solutions.

Using Covariates to Design an Experiment

Randomization is motivated by broader considerations:

1. Transparent and testable (post-randomization balance checks):
   - "...in this crooked world, he else can he avoid the charge of doctoring his own data?..." (Basu 1980, p. 594)
   - "...more difficult for an experimenter to 'cheat'..." (Gelman et al. 2004, p. 225).
2. When X_i is vector valued, not clear how to assess imbalance. Different metrics may imply different rankings of solutions.
3. Eliminates need to model outcome-covariate relationships.

Using Covariates to Design an Experiment

Randomization is motivated by broader considerations:

1. Transparent and testable (post-randomization balance checks):
   - "...in this crooked world, he else can he avoid the charge of doctoring his own data?..." (Basu 1980, p. 594)
   - "...more difficult for an experimenter to 'cheat'..." (Gelman et al. 2004, p. 225).
2. When X_i is vector valued, not clear how to assess imbalance. Different metrics may imply different rankings of solutions.
3. Eliminates need to model outcome-covariate relationships.
4. Even if one wants to use models to estimate effects, randomization guarantees a consistent benchmark for model checking (Gelman et al. 2004, p. 224).

Using Covariates to Design an Experiment

Randomization is motivated by broader considerations:

1. Transparent and testable (post-randomization balance checks):
   - "...in this crooked world, he else can he avoid the charge of doctoring his own data?..." (Basu 1980, p. 594)
   - "...more difficult for an experimenter to 'cheat'..." (Gelman et al. 2004, p. 225).
2. When X_i is vector valued, not clear how to assess imbalance. Different metrics may imply different rankings of solutions.
3. Eliminates need to model outcome-covariate relationships.
4. Even if one wants to use models to estimate effects, randomization guarantees a consistent benchmark for model checking (Gelman et al. 2004, p. 224).
5. Provides basis for statistical inference by establishing assignment mechanism specification "that all scientists will accept" (Rubin 1980, p. 591).

Using Covariates to Design an Experiment

A way to use covariates with random assignment is block randomization.

- Experimental analogue to stratified random sampling.
- Create blocks, then do random assignment within each block.
- Precision gains are based on within-block outcome (or treatment effect) homogeneity.
- Assignment status is *independent* for units in different blocks. Each block is like a separate experiment.
- Useful for subgroups that you will analyze separately.

Block Randomization: The Method

- Partition the sample, $\mathscr{S}$, into B blocks,, $\mathscr{S}_1, \ldots, \mathscr{S}_B$, each of size $N_1, \ldots, N_B$.
- Assign N_{1b} units to treatment completely at random within each block and $N_{0b} = N_b - N_{1b}$ to control.
- We don't assume that the same proportion is treated in each stratum, and so some weighting will be necessary to account for variable assignment probabilities.
- When $N_{1b} = N_{0b} = 1$, we have a "matched pairs" design.

Block Randomization: Estimation

- Simple sample ATE estimator is a stratified difference in means,

$$\hat{\beta}_S = \sum_{b=1}^{B} \frac{N_b}{N} \hat{\beta}_b$$

- Useful to restate in terms of unit level weighting:

$$\begin{aligned}
\hat{\beta}_S &= \sum_{b=1}^{B} \frac{N_b}{N} \left[\frac{1}{N_{1b}} \sum_{i=1}^{N_b} D_{ib} Y_{1ib} - \frac{1}{N_{0b}} \sum_{i=1}^{N_b} (1 - D_{ib}) Y_{0ib} \right] \\
&= \left[\frac{1}{N} \sum_{b=1}^{B} \sum_{i=1}^{N_b} \frac{D_{ib}}{N_{1b}/N_b} Y_{1ib} \right] - \left[\frac{1}{N} \sum_{b=1}^{B} \sum_{i=1}^{N_b} \frac{1 - D_{ib}}{N_{0b}/N_b} Y_{0ib} \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} D_i \frac{1}{\pi_{b[i]}} Y_i - \frac{1}{N} \sum_{i=1}^{N} (1 - D_i) \frac{1}{1 - \pi_{b[i]}} Y_i,
\end{aligned}$$

where $\pi_{b[i]}$ is the treatment probability in stratum b.

Block Randomization: Estimation

- So, the sample average treatment effect is estimated via a weighted difference in means, where a unit's weight is,

$$w_i = D_i \frac{1}{\pi_{b[i]}} + (1 - D_i) \frac{1}{1 - \pi_{b[i]}}.$$

- When a fixed proportion, π, is assigned to treatment within each block, the estimate is simply the usual difference in means,

$$\hat{\beta}_S = \frac{1}{\pi N} \sum_{i=1}^{N} D_i Y_i - \frac{1}{(1-\pi)N} \sum_{i=1}^{N} (1 - D_i) Y_i,$$

although there are gains in variance reduction from blocking.

Block Randomization: Estimation

Assuming fixed blocks, our block randomized experiment is essentially B independent experiments. Therefore by analogy to our analysis of completely randomized experiments,

$$\text{Var}\left[\hat{\beta}_S|\mathscr{S}\right] = \sum_{b=1}^{B}\left(\frac{N_b}{N}\right)^2 \text{Var}\left[\hat{\beta}_b\right] \leq \sum_{b=1}^{B}\left(\frac{N_b}{N}\right)^2\left(\frac{S_{1b}^2}{N_{1b}} + \frac{S_{0b}^2}{N_{0b}}\right) \equiv V_{\hat{\beta}_S},$$

Expression holds at equality under (i) constant effects or (ii) random sampling within blocks.

Block Randomization: Estimation

Estimation via OLS:

- Weighted OLS using w_i and block-level FEs.
- Equivalently, "centered-interactions" model (Imbens & Wooldridge 2009, p. 28).

Block Randomization: Variance and Power

Block randomization is analogous to stratified sampling:

- Not stratified *means* but stratified *differences in means*.
- Optimal allocation over strata (e.g., a Neyman allocation) would be based on the standard error of the block-level difference-in-means.

Block Randomization: Variance and Power

Design effect tells us (i) precision gains from blocking and (ii) how to rescale a sample size estimate based on power analysis for completely randomized experiment (e.g., w/ `sampsi`):

$$D^2(\hat{\beta}, Block) = \frac{V_{\hat{\beta}_S}}{V_{\hat{\beta}}} = \sum_{b=1}^{B} \left(\frac{N_b}{N}\right)^2 \frac{\text{Var}[\hat{\beta}_b]}{\text{Var}[\hat{\beta}]} = \sum_{b=1}^{B} \left(\frac{N_b}{N} \frac{s.e.[\hat{\beta}_b]}{s.e.[\hat{\beta}]}\right)^2 .$$

Block Randomization: Variance and Power

You can approximate the design effect with data from past studies, although no single approximations is perfect. If you have data from a past block randomized experiment,

- Approximate $s.e.[\hat{\beta}_b]$'s with standard errors for treatment effect from regressions *within* each of the blocks.
- Approximate $s.e.[\hat{\beta}]$ with standard error for treatment effect from weighted OLS regression that excludes block dummies.

With data from a completely randomized experiment, create blocks and then compute as above *as if* you had block randomized experiment.

Block Randomization: Variance and Power

We can also express the design effect in term of more rudimentary quantities:

$$D^2(\hat{\beta}, Block) = \sum_{b=1}^{B} \frac{\left(\frac{N_b}{N}\right)^2 \left(\frac{S_{1b}^2}{N_{1b}} + \frac{S_{0b}^2}{N_{0b}}\right)}{\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}}$$

$$= \sum_{b=1}^{B} \underbrace{\frac{\pi(1-\pi)}{\pi_b(1-\pi_b)}}_{\text{allocation variability}} \underbrace{\frac{S_{1b}^2 N_{0b} + S_{0b}^2 N_{1b}}{S_1^2 N_0 + S_0^2 N_1}}_{\text{departure from optimal allocation}} .$$

We can thus approximate the design effect using data from past studies, exploring the implications of different design choices for π's and N's.

Block Randomization: Variance and Power

When treatment and control variances are equal, design effect is,

$$D^2(\hat{\beta}, Block_{eq}) = \sum_{b=1}^{B} \frac{\pi(1-\pi)}{\pi_b(1-\pi_b)} \frac{N_b}{N} \frac{S_b^2}{S^2}.$$

When the terms of this expression are constant over b, then this reduces to average of within-block outcome variances divided by total outcome variance, which is approximated by,

$$1 - \rho_{ICC,Y}.$$

Block Randomization: Variance and Power

To determine sample size, multiply the design effect times the result from the computation based on the completely randomized experiment:

$$\begin{aligned} N_{block} &= D^2(\hat{\beta}, Block) N_{complete} \\ &= D^2(\hat{\beta}, Block) \left[\frac{(t_{\alpha/2} + t_{1-\kappa})^2 \left(\frac{S_1^2}{p} + \frac{S_0^2}{1-p} \right)}{\beta^2} \right], \end{aligned}$$

where $(t_{\alpha/2} + t_{1-\kappa})^2 \approx 8$ by conventional standards.

Block Randomization: Variance and Power

In practice, you should explore different ways estimating design effects:

- Simulating experiments and "placebo trials" (Bruhn & McKenzie, 2009).
- Doing simulations with fake outcome data that have distributions that resemble data from studies in your area of application.

Block Randomization: Variance and Power

- Sample sizes from analysis of completely randomized experiment is *conservative* if you are actually blocking.
- Conservative approach might be preferred, because it provides degrees of freedom.
- Pinning down design effects to reduce sample size is appropriate only when you face constraints.

Block Randomization: Variance and Power

- Generally, smaller N_b will decrease D^2 so long as the more refined blocks make some progress in reducing within-block variance (cf. Imbens 2011).
- This suggests that we should use good covariates to create as refined blocks as possible.
- This is the motivation for the matched pair design.

Block Randomization: Matched Pairs?

So, should we *always* seek to pair-match?

- There is debate, centering on difference between *efficiency* and *power*.
- Greevy et al. (2004) and Imai et al. (2009) say yes.
- Imbens (2011) says no:
  - Complications in variance estimation for matched pairs undermine the precision benefits of going from blocks of 4 to matched pairs
  - So, Imbens finds no gains from going beyond blocks of 4.
- (Of course, we *could* just use a matched pair design, and then in the analysis *pair the pairs* to create sets of 4 for variance estimation. Then, one gets the efficiency of matched pairs with power of what Imbens is suggesting.)

Block Randomization with Lots of Covariates

What to do with lots of covariates??

- 1. Barrios (2014) approach:
  - Get data that contain both covariates and primary outcome for population being studied (e.g., baseline data or recent survey in study locale).
  - Fit a rich model with these data.
  - If covariate set is really large, you can use, e.g., LASSO.
  - Use the model to predict "prognostic scores" for your experimental units.
  - Rank units by these prognostic scores and create blocking strata with adjacently-ranked units.

Block Randomization with Lots of Covariates

What to do with lots of covariates??

- 2. Bruhn & McKenzie (2009) "big stick":
  - Only accept randomizations that meet a balance criterion.
  - e.g, min naive *p*-value for KS-test .6, say.
  - Can make conditions more stringent on more important variables.
  - Yields a "restricted randomization."
  - Note: this is usually a *very good idea*!
  - May result in varying probabilities of treatment and complex covariances.
  - All of this can be handled by simulating the randomization distribution (e.g., taking 5000 draws from the restricted randomization).

Block Randomization with Lots of Covariates

What to do with lots of covariates??

- 3. Coarsening, clustering, or "non bipartite matching":
  - Resources in R:

```
blockTools
cem
experiment
hclust
nbpMatching
```

Block Randomization: From sample to population

- Suppose units themselves were sampled from population with probabilities that vary.
- Then the unit-level *treatment* weight should be premultiplied by inverse of the *sampling* probability:

$$\omega_{i,pop} \equiv \frac{1}{\Pr[i \in \mathscr{S}]} w_i$$

- The weighted analysis accounting for both unequal sampling and treatment assignment probabilities is needed for design-consistent population inference.

Block Randomization: Recap

Block Randomization: Recap

- Block whenever you can and at as refined a level as you can.

Block Randomization: Recap

- Block whenever you can and at as refined a level as you can.
- There are two brakes on increasing the amount of information that you use to block: (i) Imbens's sample size floor ($N_b \geq 4$ and $N_{1b}, N_{0b} \geq 2$), and (ii) when blocking on more information causes you to compromise on the balance you get with important variables (cf. Bruhn & McKenzie, pp. 211-212).

Block Randomization: Recap

- Block whenever you can and at as refined a level as you can.
- There are two brakes on increasing the amount of information that you use to block: (i) Imbens's sample size floor ($N_b \geq 4$ and $N_{1b}, N_{0b} \geq 2$), and (ii) when blocking on more information causes you to compromise on the balance you get with important variables (cf. Bruhn & McKenzie, pp. 211-212).
- Stratify on subgroups that you want to analyze separately. This helps to ensure adequate power for the subgroup analyses.

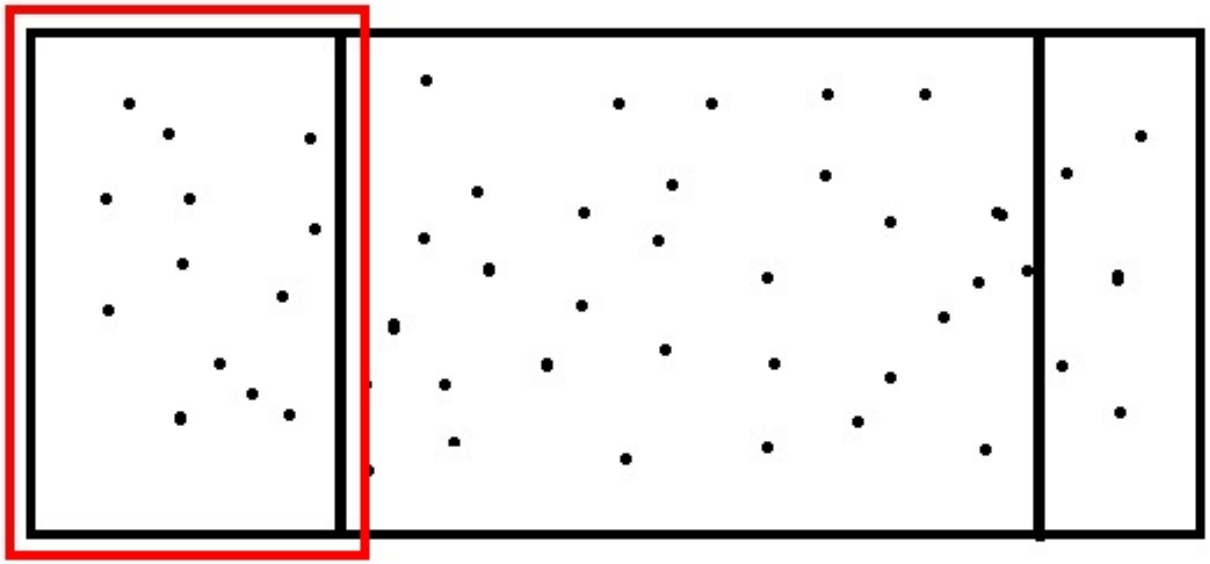
Block Randomization: Recap

- Block whenever you can and at as refined a level as you can.
- There are two brakes on increasing the amount of information that you use to block: (i) Imbens's sample size floor ($N_b \geq 4$ and $N_{1b}, N_{0b} \geq 2$), and (ii) when blocking on more information causes you to compromise on the balance you get with important variables (cf. Bruhn & McKenzie, pp. 211-212).
- Stratify on subgroups that you want to analyze separately. This helps to ensure adequate power for the subgroup analyses.
- Use weights to account for unequal treatment or sampling rates across blocks.

Block Randomization: Recap

- Block whenever you can and at as refined a level as you can.
- There are two brakes on increasing the amount of information that you use to block: (i) Imbens's sample size floor ($N_b \geq 4$ and $N_{1b}, N_{0b} \geq 2$), and (ii) when blocking on more information causes you to compromise on the balance you get with important variables (cf. Bruhn & McKenzie, pp. 211-212).
- Stratify on subgroups that you want to analyze separately. This helps to ensure adequate power for the subgroup analyses.
- Use weights to account for unequal treatment or sampling rates across blocks.
- Using block-FEs implies an assumption that blocks are fixed.

Cluster Sampling



Cluster sampling: the method

- Partition U into N non-overlapping and exhaustive cells, $U_1, \ldots, U_N$, each of size $M_1, \ldots, M_N$.
- Take sample of *cells* of size n. These are the "clusters" or "primary sampling units" (PSUs).
- Either survey everyone in each PSU or take SRS's of size m_i within each PSU.
- The units selected inside each cluster are the "secondary sampling units" (SSUs).

Cluster sampling: the rationale

- Reduces administrative cost.
- Basis of multistage sampling for situations when there is no pre-existing frame.

Cluster sampling: choices

- How should I create clusters?
- How many clusters should I sample?
- How many units should I sample from within each cluster?

Cluster sampling: estimation (classical approach)

- First, suppose that clusters are chosen via SRS.
- For the sake of analysis, suppose we estimate the population *total*, $\tau_Y = \sum_i \sum_j y_{ij}$, as,

$$\hat{\tau}_{cl} = \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} Y_{ij} = N \frac{1}{n} \sum_{i=1}^{n} M_i \bar{Y}_i = N \bar{\hat{\tau}},$$

  where $\bar{\hat{\tau}}$ is mean of cluster total estimates.
- Clearly, $E[\hat{\tau}_{cl}] = \tau_Y$.
- Variance is given by,

$$\text{Var}\,[\hat{\tau}_{CL}] = \underbrace{N(N-n)\frac{S_\tau^2}{n}}_{\text{Variance from PSU totals } (\tau_i \text{s})} + \underbrace{\frac{N}{n} \sum_{i=1}^{N} M_i (M_i - m_i) \frac{S_i^2}{m_i}}_{\text{Variance from SSU values } (y_{ij} \text{s})}.$$

- To estimate *mean*, divide by $M = \sum_{i=1}^{N} M_i$:

$$\bar{Y}_{cl} = \frac{1}{M} \hat{\tau}_{cl} \text{ and } \text{Var}\,[\bar{Y}_{cl}] = \frac{1}{M^2} \text{Var}\,[\hat{\tau}_{cl}]$$

Cluster sampling: design effects

- To simplify, suppose all clusters are of size $\bar{M}$ and that we select all units in a cluster (so, $m_i = M_i = \bar{M}$). Then,

$$\text{Var}\left[\hat{\tau}_{CL}\right] = N(N-n)\frac{S_\tau^2}{n} = N(N-n)\frac{\bar{M}\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(\mu_{yi}-\mu_y)^2}{n}.$$

- Compare to the *SI* design (ie, SRSWOR) that selects $n\bar{M}$ units, for which we estimate τ_Y as $\hat{\tau}_{SI} = n\bar{M}\bar{Y}$, with variance,

$$\begin{aligned}\text{Var}\left[\hat{\tau}_{SI}\right] &= (n\bar{M})^2\frac{S^2}{n\bar{M}}\left(\frac{N\bar{M}-n\bar{M}}{N\bar{M}}\right) = N(N-n)\bar{M}\frac{S^2}{n} \\ &= N(N-n)\bar{M}\frac{\frac{1}{N\bar{M}-1}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\mu_y)^2}{n}.\end{aligned}$$

- Again, to estimate means, we just divide by $M = n\bar{M}$.
- Putting it together to get design effect:

$$D^2(\bar{Y}, CL) = \frac{\text{Var}\left[\hat{\tau}_{CL}\right]}{\text{Var}\left[\hat{\tau}_{SI}\right]} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(\mu_{yi}-\mu_y)^2}{\frac{1}{N\bar{M}-1}\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\mu_y)^2}$$

where μ_y is the overall mean of y.

Cluster sampling: design effects

- Another way to express this is in terms of the the population *intra-cluster correlation* (ICC) for y:

$$\begin{aligned}\rho_y &= \frac{\mathrm{E}\left[(y_{ij}-\mu_y)(y_{ij'}-\mu_y)\right]}{\mathrm{E}\left[(y_{ij}-\mu_y)^2\right]} \\ &= \frac{\sum_{i=1}^{N}\frac{1}{\bar{M}-1}\sum_{j=1}^{\bar{M}}\sum_{j'\neq j}(y_{ij}-\mu_y)(y_{ij'}-\mu_y)}{\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(y_{ij}-\mu_y)^2}.\end{aligned}$$

- Within-variance $\to 0$ implies homogeneity within clusters.
- As this happens ρ_y goes to its maximum of 1.
- So, ρ_y measures the degree of within-cluster homogeneity for the variable y.
- Then as Thompson (2012, ch. 12-13) shows:

$$D^2(\bar{Y}, CL) \approx 1 + (m' - 1)\rho_y,$$

where $m' = \sum_{i=1}^{n} m_i^2 / \sum_{i=1}^{n} m_i$.

- Design effect increases in cluster size (holding overall sample size fixed) and ICC.

Cluster sampling: design effects

- So long as cluster sizes do not vary too much, Kalton et al. (2005, pp. 106-7) approximate $D^2(\bar{Y}, CL)$ by using the average cluster size, $\bar{m}$, rather than m', yielding:

$$D^2(\bar{Y}, CL) \approx 1 + (\bar{m} - 1)\rho_y$$

- For stratified designs, Kalton et al. propose an approximation that uses the average of $\rho_{y,h}$'s over strata, $\bar{\rho}_y \equiv \frac{1}{L}\sum_{h=1}^{L} \rho_{y,h}$, and the average cluster size, $\bar{m}$, yielding

$$D^2(\bar{Y}, STCL) = \frac{\mathrm{Var}_{CL}[\bar{Y}_{st}]}{\mathrm{Var}_{SI}[\bar{Y}]} \approx 1 + (\bar{m} - 1)\bar{\rho}_y.$$

Cluster sampling: design effects

- $\text{Var}\left[\hat{\tau}_{CL}\right]$ shows that increases in m_i's and n drive down variance:

$$\text{Var}\left[\hat{\tau}_{CL}\right] = \underbrace{N(N-n)\frac{S^2_{\tau_i}}{n}}_{\text{Between-cluster variance}} + \underbrace{\frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{S_i^2}{m_i}}_{\text{Within cluster variance}}.$$

- However, $D^2(\bar{Y}, CL)$ shows that if we hold total sample size $(n * \bar{m})$ fixed, larger clusters leads to efficiency *losses*:

$$D^2(\bar{Y}, CL) = \frac{\text{Var}_{CL}[\bar{Y}_{cl}]}{\text{Var}_{SI}[\bar{Y}]} \approx 1 + (m' - 1)\rho_y,$$

  where $m' = \sum_{i=1}^{n} m_i^2 / \sum_{i=1}^{n} m_i$.
- You gain more precision by increasing *number of clusters*, not size of clusters.
- We are penalized by within-cluster homogeneity (ρ_y). If we have a choice, clusters should *heterogenous*.

Cluster sampling: PPS

- "Probability proportional to size" (PPS) sampling is a common method of cluster sampling.
- Select with probability, $\pi_{i'} = \alpha \frac{M_i'}{\sum_{i=1}^{N} M_j}$.
- Then, if $m_i = m$, constant, we have a self-weighting sample:

$$\pi_{i'j} = \pi_{i'} \pi_{j|i'} = \alpha \frac{M_{i'}}{\sum_{i=1}^{N} M_i} \frac{m}{M_{i'}} = \frac{\alpha m}{\sum_{i=1}^{N} M_i} = \frac{\alpha m}{\bar{M} N},$$

  providing much analytical convenience, in addition to administrative convenience.
- If we only have an estimate of M_i, say $\tilde{M}_i$, $\pi_{i'j} = \frac{\alpha m}{\bar{M} N} \frac{\tilde{M}_i}{M_i}$, varying over clusters. This introduces some bias unless we can obtain the true $\frac{\alpha m}{\bar{M} N} \frac{\tilde{M}_i}{M_i}$ and use it in a weighted analysis.
- If M_i or $\tilde{M}_i$ is correlated with Y, then PPS is also highly efficient.

Clustering in experiments

Very similar results obtain for cluster randomized experiments...

Cluster randomization: The method

- Suppose we have J clusters all of size n. They may have been obtained via a sample from some large population of clusters, or they may be a convenience sample of clusters.
- Some fraction, p, of the clusters are assigned to treatment ($D_j = 1$) and the remaining $1-p$ fraction are assigned to control ($D_j = 0$).
- Unit i within cluster j is characterized by potential outcomes, (Y_{1ij}, Y_{0ij}). The experiment reveals $Y_{ij} = D_{j[i]}Y_{1ij} + (1 - D_{j[i]})Y_{0ij}$.
- The estimand is the unit level average treatment effect: $\mathrm{E}[Y_{1ij} - Y_{0ij}]$.
- Our estimator, $\hat{\beta}$, is the coefficient on $D_{j[i]}$ from an OLS regression of the Y_{ij}'s on the $D_{j[i]}$'s and possibly some covariates. When there are no covariates, this is just the difference in *unit-level* treatment and control means.

Cluster randomization: The method

- Because we have assumed constant cluster sizes, $\hat{\beta}$ is unbiased and consistent for the ATE.
- Under variable cluster sizes, $\hat{\beta}$ is not unbiased, but it is consistent (see Quant II notes on robust inference).

Cluster randomization: Variance and Power

- Suppose the case of no covariates. Then,

$$\hat{\beta} = \bar{Y}_1 - \bar{Y}_0,$$

  the difference in the unit level treatment and control means.
- Then, the randomization+sampling variance is given by,

$$\text{Var}[\hat{\beta}] = \text{Var}[\bar{Y}_1] + \text{Var}[\bar{Y}_0].$$

- (As we have seen before, if we hold the sample fixed, then this quantity will exceed the randomization-only variance. As such, it provides a conservative approximation.)
- Essentially, we have a variance for two *cluster-sample* means. Therefore, results (e.g., design effects) from cluster sampling will carry through directly to this setting.
- Sampling here refers to sampling *of* and *within* clusters.

Cluster randomization: Variance and Power

- To derive the variance, first consider a generic potential outcome, Y_{dij}, with $d = 0, 1$. We can decompose Y_{dij} into,

$$Y_{dij} = \mu_d + u_{dj} + e_{dij} \tag{1}$$

  where $\mu_d = \bar{Y}_d$, $u_{dj} = \bar{Y}_{dj} - \bar{Y}_d$, and $e_{dij} = Y_{dij} - (\mu_d + u_{dj})$. By construction, $\mathrm{E}[u_{dj}] = 0$ and $\mathrm{E}[e_{dij}] = 0$.
- $\mathrm{Cov}[u_{dj}, e_{dij}] = 0$, so $\mathrm{Var}[Y_{dij}] = \mathrm{Var}[u_{dj}] + \mathrm{Var}[e_{dij}] = \sigma^2_{db} + \bar{\sigma}^2_{dw}$, where σ^2_{db} is the variance of the cluster-level means, and $\bar{\sigma}^2_{dw}$ is the average of the within-cluster variances.
- Also, $\mathrm{Cov}[e_{dij}, e_{di'j}] = 0$ for $i \neq i'$, and so $\mathrm{Cov}[Y_{dij}, Y_{di'j}] = \mathrm{Var}[u_{dj}] = \sigma^2_b$.
- Moment conditions are by *construction*, not by assumption.

Cluster randomization: Variance and Power

For $\text{Var}[\bar{Y}_1]$:

$$\begin{aligned}
\text{Var}[\bar{Y}_1] &= \text{Var}\left[\frac{1}{pJ}\sum_{j:D_j=1}\frac{1}{n}\sum_{i=1}^{n}Y_{ij}\right] \\
&= \frac{1}{p^2J^2}\sum_{j:D_j=1}\frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n}Y_{ij}\right] \\
&= \frac{1}{pJ}\left(\frac{1}{pJ}\sum_{j:D_j=1}\frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n}Y_{ij}\right]\right) \\
&= \frac{1}{pJ}\left(\frac{\bar{\sigma}^2_{1w}}{n}+\sigma^2_{1b}\right).
\end{aligned}$$

Cluster randomization: Variance and Power

We may observe that,

$$\begin{aligned}\frac{1}{pJ}\left(\frac{\bar{\sigma}^2_{1w}}{n}+\sigma^2_{1b}\right) &= \frac{1}{pJn}(\bar{\sigma}^2_{1w}+\sigma^2_{1b}+n\sigma^2_{1b}-\sigma^2_{1b}) \\ &= \frac{1}{pJn}[\bar{\sigma}^2_{1w}+\sigma^2_{1b}+(n-1)\sigma^2_{1b}] \\ &= \frac{\mathrm{Var}[Y_1]}{N_1}\underbrace{[1+(n-1)\rho_{ICC,1}]}_{\text{Aha!}}\end{aligned}$$

Similar for $\mathrm{Var}[\bar{Y}_0]$.

Cluster randomization: Variance and Power

- Putting it all together we have,

$$\text{Var}[\hat{\beta}] = \frac{\text{Var}[Y_1]}{N_1}[1+(n-1)\rho_{ICC,1}] + \frac{\text{Var}[Y_0]}{N_0}[1+(n-1)\rho_{ICC,0}],$$

where the red components sum the usual unit contributions to the variance and the blue components are the variance inflation factors due to clustering.

Cluster randomization: Variance and Power

To get the same precision as a completely randomized design:

- Inflate the treatment group by $[1+(n-1)\rho_{ICC,1}]$.
- Inflate the control group by $[1+(n-1)\rho_{ICC,0}]$.
- If the $\rho_{ICC,0} = \rho_{ICC,1} = \rho_{ICC}$, then the design effect is,

$$D^2(\hat{\beta}, CL_{const.icc}) = 1+(n-1)\rho_{ICC},$$

and you would inflate both treatment and control groups by this amount.

Cluster randomization: Variance and Power

Implementation:

- Estimate ρ_{ICC} from auxiliary data. Get a bunch of estimates to see what the "best" and "worst" cases might look like.
- Obtain sample size estimate under the assumption of a completely randomized experiment.
- Apply the inflation factors to your treatment and control group sample sizes.
- In Stata there is a ado called `sampclus` that does exactly this: you start with an approximation from `sampsi`, then use `sampclus` to apply the inflation factor.
- Another popular tool is the `Optimal Design` program (you can download it for free), which requires similar input as `sampsi` and `sampclus`.

Cluster randomization: Variance and Power

A crucial point:

power and consistency are dictated by number of clusters,

not size of clusters:

$$\text{Var}[\hat{\beta}] = \frac{1}{pJ}\left(\frac{\bar{\sigma}^2_{1w}}{n} + \sigma^2_{1b}\right) + \frac{1}{(1-p)J}\left(\frac{\bar{\sigma}^2_{0w}}{n} + \sigma^2_{0b}\right)$$

Expression goes to zero in J but not n, because increasing n does nothing to tame between cluster variance.

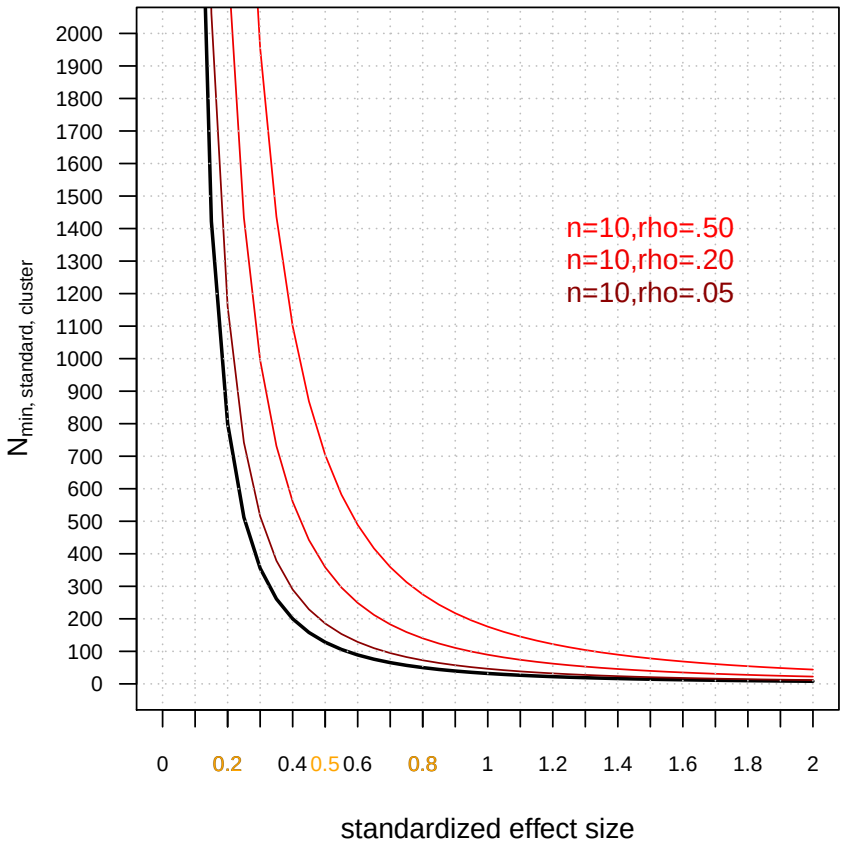
Cluster randomization: Variance and Power

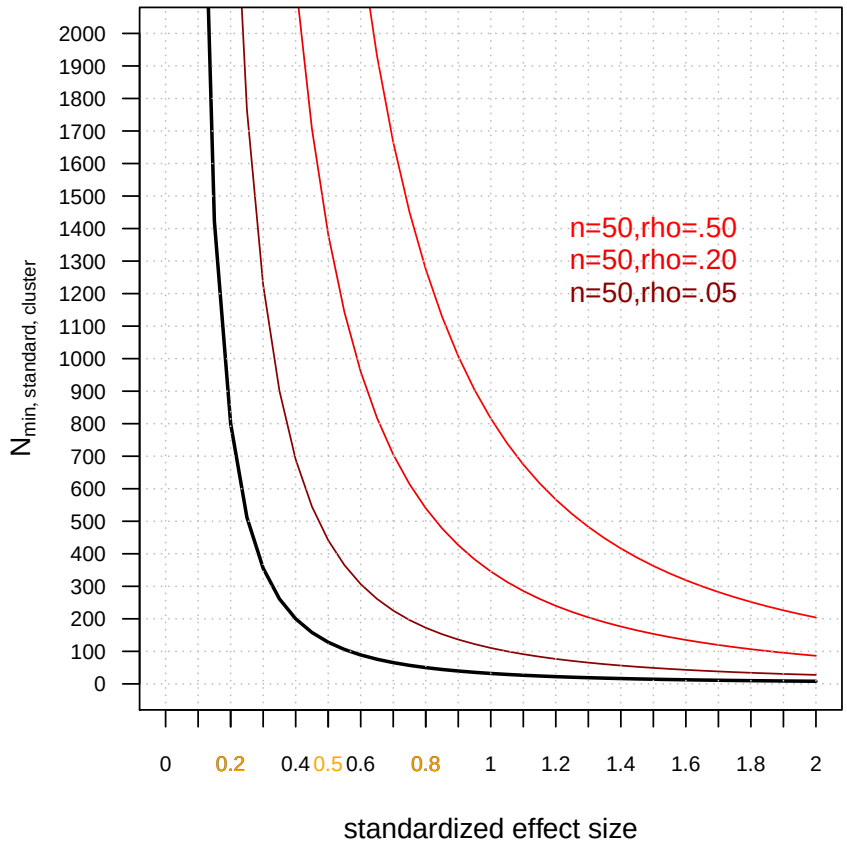
The modification under varying cluster sizes is (Donner et al., 1981):

$$\text{Var}\left[\bar{Y}_d\right] = \frac{\text{Var}\left[Y_d\right]}{\sum_{j:D_j=d} n_j}\left[1 + \left(\frac{\sum_{j:D_j=d} n_j^2}{\sum_{j:D_j=d} n_j} - 1\right)\rho_{ICC,d}\right].$$

Another "conservative" way to proceed would be to use the equal cluster size formula but to use the largest cluster size for n. (Most available software assumes equal cluster sizes, so you'd have to do something like this.)

Cluster randomization: Variance and Power





Summary

- Stratification is a way to use pre-existing information to make a design more efficient.
- Clustering is sometimes necessary for administrative reasons and it makes a design less efficient.
- Design effects quantify the efficiency consequences.
- Stratification and clustering can be combined ("complex" design).