LECTURE 2: STRATIFIED AND PAIRED RANDOMIZED EXPERIMENTS AND POWER CALCULATIONS

Guido Imbens – Stanford University

Economics 272, GSB 507, Spring 2024

OUTLINE

1. Stratified Randomized Experiments
2. Re-randomization
3. Power Calculations

STRATIFIED RANDOMIZED EXPERIMENTS

- Setting: Suppose we have N units, we observe some covariates, and wish to evaluate a binary treatment.
- Question:
  - Should we randomize the full sample,
  - or should we stratify the sample first by pre-treatment variables before randomizing within strata, or even pair the units up and randomize within the pairs?
- 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."

MORE QUOTES

- 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."

MORE QUOTES

- 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."

WHAT TO MAKE OF THIS?

- How Do We Reconcile These Statements?
  - Be careful and explicit about goals: precision of estimators versus power of tests.
  - Be careful about estimands: what is the population of interest.

EXPECTED SQUARED ERROR CALCULATIONS FOR COMPLETELY RANDOMIZED VS STRATIFIED RANDOMIZED EXPERIMENTS

- Exact calculation in simple setting to clarify issues raised by quotes.
- Suppose we have a single binary covariate $X_i \in \{f, m\}$. Define

$$\tau(x) \equiv \mathbb{E}\left[Y_i(\mathrm{T}) - Y_i(\mathrm{C}) \middle| X_i = x \right], \qquad x \in \{f, m\}$$

  where the expectations denote expectations taken over the super-population.
- The estimand we focus on is the super-population version of the the finite sample average treatment effect,

$$\tau_{\mathrm{pop}} \equiv \mathbb{E}\left[Y_i(\mathrm{T}) - Y_i(\mathrm{C}) \right] = \mathbb{E}\left[\tau(X_i) \right]$$

NOTATION

- conditional means

$$\mu(w,x) \equiv \mathbb{E}\left[\left. Y_i(w) \right| W_i = w, X_i = x \right],$$

- conditional variances

$$\sigma^2(w,x) \equiv \mathbb{V}\left(\left. Y_i(w) \right| W_i = w, X_i = x \right),$$

  for $w = \mathrm{C}, \mathrm{T}$, and $x \in \{f, m\}$,

- conditional variance of treatment effect

$$\sigma^2_{\mathrm{CT}}(x) \equiv \mathbb{E}\left[\left. (Y_i(\mathrm{T}) - Y_i(\mathrm{C}) - (\mu(\mathrm{T},x) - \mu(\mathrm{C},x)))^2 \right| X_i = x \right],$$

THREE ESTIMATORS:

- First, simple difference:

$$\hat{\tau}_{\text{dif}} = \overline{Y}_{\text{T}}^{\text{obs}} - \overline{Y}_{\text{C}}^{\text{obs}}$$

- Second, use the regression function

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

  Then estimate τ 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:

$$\hat{\tau}_{\text{strat}} = \frac{N_{\text{C}f} + N_{\text{T}l}}{N} \left(\overline{Y}_{\text{T}f}^{\text{obs}} - \overline{Y}_{\text{C}f}^{\text{obs}} \right) + \frac{N_{\text{C}m} + N_{\text{T}m}}{N} \left(\overline{Y}_{\text{T}m}^{\text{obs}} - \overline{Y}_{\text{C}m}^{\text{obs}} \right)$$

DESIGN: LARGE (INFINITELY LARGE) SUPERPOPULATION

- We draw a stratified random sample of size N from this population.
- Share p come from $X_i = f$ subpopulation, and share 41-p $from\mathrm{X}_i = m$ subpopulation.
- Two experimental designs.
  - completely randomized design ($\mathcal{C}$) where qN units are randomly assigned to the treatment group, and the remaining $(1 - q)N$ are assigned to the control group.
  - stratified randomized design ($\mathcal{S}$) where $q\,pN$ are randomly selected from the $X_i = f$ subsample and assigned to the treatment group, and $q(1 - p)N$ 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: $\text{pr}(W_i = 1|X_i) = q$, for both types, $x = f, m$.

VARIANCES FOR $\hat{\tau}_{\text{dif}}$ FOR TWO DESIGNS

- p is share of $X_i = f$ subpopulation, q is probability of treatment

$$\mathbb{V}_{\mathcal{S}} = \mathbb{E}\left[\left(\hat{\tau}_{\text{dif}} - \tau_{\text{pop}}\right)^2 \middle| \mathcal{S} \right]$$

$$= \frac{q}{N} \cdot \left(\frac{\sigma^2(\mathrm{T}, f)}{p} + \frac{\sigma^2(\mathrm{C}, f)}{1-p} \right) + \frac{1-q}{N} \cdot \left(\frac{\sigma^2(\mathrm{T}, m)}{p} + \frac{\sigma^2(\mathrm{C}, m)}{1-p} \right)$$

$$\mathbb{V}_{\mathcal{C}} = \mathbb{E}\left[\left(\hat{\tau}_{\text{dif}} - \tau_{\text{pop}}\right)^2 \middle| \mathcal{C} \right] = q(1-q) \left((\mu(\mathrm{C}, f) - \mu(\mathrm{C}, m))^2 + (\mu(\mathrm{T}, f) - \mu(\mathrm{T}, m))^2 \right)$$

$$+ \frac{q}{N} \cdot \left(\frac{\sigma^2(\mathrm{T}, f)}{p} + \frac{\sigma^2(\mathrm{C}, f)}{1-p} \right) + \frac{1-q}{N} \cdot \left(\frac{\sigma^2(\mathrm{T}, m)}{p} + \frac{\sigma^2(\mathrm{C}, m)}{1-p} \right)$$

$$\mathbb{V}_{\mathcal{C}} - \mathbb{V}_{\mathcal{S}} = q(1-q) \cdot \left((\mu(\mathrm{C}, f) - \mu(\mathrm{C}, m))^2 + (\mu(\mathrm{T}, f) - \mu(\mathrm{T}, m))^2 \right) \geq 0$$

COMMENT 1:

- Stratified randomized design has lower expected squared error than completely randomized design.
- Strictly lower if the covariate predicts potential outcomes in population.
  - Exact finite sample result.
  - 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

$$\mathbb{E}\left[(\hat{\tau}_{\text{dif}} - \tau)^2 \middle| \mathbf{Y}(\mathrm{C}), \mathbf{Y}(\mathrm{T}), \mathbf{X}, \mathcal{C}\right]$$

$$\textit{versus} \qquad \mathbb{E}\left[(\hat{\tau}_{\text{dif}} - \tau)^2 \middle| \mathbf{Y}(\mathrm{C}), \mathbf{Y}(\mathrm{T}), \mathbf{X}, \mathcal{S}\right].$$

- It is possible that stratification leads to larger variances because of negative correlations within strata in a finite sample (consistent with Snedecor and Cochran quote). That is not possible on average, that is, over repeated samples.
- 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 $\hat{\tau}_{\text{post}}$, $\hat{\tau}_{\text{strat}}$, 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 $\hat{\tau}_{\text{post}}$ will have a lower variance than $\hat{\tau}_{\text{dif}}$.
    - However, for any fixed sample size, if the correlation is sufficiently weak, the variance of $\hat{\tau}_{\text{post}}$ will actually be strictly higher than that of $\hat{\tau}_{\text{dif}}$.

THINK THROUGH ANALYSES IN ADVANCE!

- “Design Trumps Analysis” – Donald Rubin
  - For *ex post* adjustment (*e.g.* regression) 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 that $\hat{\tau}_{\text{dif}} = \hat{\tau}_{\text{reg}}$ (*e.g.*, stratify, or re-randomize).

PAIRWISE RANDOMIZATION

- Klar & Donner argument
- Compare two designs with $4N$ units.
  - N strata with 4 units each ($\mathcal{S}$).
  - $2N$ pairs with 2 units each ($\mathcal{P}$).
- What are costs and benefits of $\mathcal{S}$ versus $\mathcal{P}$?
- Benefits of Pairing $\mathcal{P}$ over Stratification $\mathcal{S}$
  - 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

- Benefits of Stratification $\mathcal{S}$ over Pairing $\mathcal{P}$
  - 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 intra-class 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 can not 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
- Kohavi: stratify, but use variance as if completely randomized (will be conversative)

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? Is that ok?
- 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

$$\overline{X}_w = \frac{1}{N} \sum_{i:W_i=w} X_i, \qquad t_X = \frac{\overline{X}_\mathrm{T} - \overline{X}_\mathrm{C}}{\sqrt{s^2_{X,\mathrm{C}}/N + s^2_{X,\mathrm{T}}/N}}$$

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

TWO TYPES OF RE-RANDOMIZATION

- 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:
  - Articulate strategy *a priori*, so randomization inference is possible.
  - Do not search over all assignments for optimal value for criterion because then there is little randomness left.

IMBALANCE STATISTIC (MORGAN & RUBIN, 2012)

- Given a K-component covariate vector $\mathbf{X}$, and with marginal assignment probability p, we have a quadratic form $\psi\{0,1\}^N \times \mathbb{X}^N \mapsto \{0,1\}$ that measures the imbalance of the covariates:

$$\psi(\mathbf{W},\mathbf{X}) = N p(1-p)\left(\overline{X}_{\mathrm{C}} - \overline{X}_{\mathrm{T}}\right)^\top \Sigma_X^{-1}\left(\overline{X}_{\mathrm{C}} - \overline{X}_{\mathrm{T}}\right).$$

- Given the randomization this statistic is approximately distributed as $\mathcal{X}^2(K)$.
- Re-randomize until $\psi(\mathbf{W},\mathbf{X}) \leq a$. Then the covariance matrix of $\mathbf{X}_{\mathrm{C}} - \mathbf{X}_{\mathrm{T}}$ conditional on $\psi(\mathbf{W},\mathbf{X}) \leq a$ is ν_a times the unconditional covariance matrix.
- Here ν_a is equal to the ratio of the probabilities of chi-squared random variables being less than a:

$$\nu_a = \frac{\mathrm{pr}(\mathcal{X}^2_{K+2} \leq a)}{\mathrm{pr}(\mathcal{X}^2_K \leq a)} \leq 1.$$

- For example, with $K = 2$, $a = 0.1$, $\nu_a \approx 0.025$.

- Suppose the outcomes follow a linear model, with

$$Y_i(w)|X_i \sim \mathcal{N}(\alpha + \beta^\top X_i, \sigma^2),$$

with R^2 equal to $1 - \sigma^2/(\beta^\top \Sigma \beta + \sigma^2)$,.

- Then the variance for the difference-in-means estimator $\hat{\tau} = \overline{Y}_\text{T} - \overline{Y}_\text{C}$ after re-randomization is

$$\mathbb{V}(\hat{\tau}, \nu_a) = \frac{\sigma^2}{N p(1 - p)} + \frac{\nu_a}{N p(1 - p)} \beta^\top \Sigma \beta.$$

(Morgan & Rubin, 2012, p. 1274)

CAUTIONARY NOTE REGARDING RE-RANDOMIZATION

- Suppose with $2N$ units, X_i earnings, $2N - 1$ units have $X_i \in [0, 10]$, and one 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.
- Recommendation Instead of re-randomization, lay out acceptable set of assignments.

POWER CALCULATIONS: TESTING A MEAN AGAINST ZERO

- Suppose we have a random sample $X_1, \ldots, X_N$ from a normal distribution with mean μ and variance σ^2.
- We wish to test the null hypothesis

$$H_0: \quad \mu = 0, \qquad\qquad \text{against } H_a: \quad \mu \neq 0.$$

- We want the probability of a type I error to be less than α (e.g., $\alpha = 0.05$).
- We want the test to have power β, that is the probability of rejecting the null hypothesis, when the null is false with $\mu = \mu_0$, should be β.
- Let's say $\beta = 0.8$, for $\mu_0 = 0.1 \cdot \sigma$ (conventional, but arbitrary choices).

- We base the test on the t-statistic

$$T = \frac{\overline{X}}{\sqrt{S_X^2/N}},$$

- where

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \text{and } S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2.$$

- So the question is now:

  what is the minimum sample size N to achieve power β against alternative μ_0/σ given size α for this test?

- If the size of the test is α, we reject the null hypothesis if

$$|T| \geq \Phi^{-1}\left(1 - \alpha/2\right), \qquad \textit{e.g.}\ \Phi^{-1}\left(1 - 0.05/2\right) = 1.96$$

- Here $\Phi(\cdot)$ is the cumulative distribution function for a standard Normal random variable

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx$$

- We use the fact that

$$\sqrt{N} \cdot \frac{\overline{X} - \mu_0}{\sigma} \approx \mathcal{N}(0, 1)$$

- Now let us look at the power, the probability of rejecting the null hypothesis when the true mean is $\mu_0 = 0.1 \cdot \sigma$.

$$\mathrm{pr}\left(|T| > \Phi^{-1}(1-\alpha/2)\right) \approx \mathrm{pr}\left(T > \Phi^{-1}(1-\alpha/2)\right)$$

$$\vdots$$

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

leading to

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

- For example, $\alpha = 0.05$ (we test at the 5% level), $\mu_0/\sigma = 0.1$ (the effect is an increase of 10% of a standard deviation),
- We wish to detect such an effect with probability 0.8 (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.

TESTING A DIFFERENCE OF MEANS WITH UNEQUAL SAMPLE SIZES AND EQUAL VARIANCES

- Now let us look at a more general/interesting case.
- We have a random sample, outcomes $Y_1, \ldots, Y_N$, and a treatment indicator $W_1, \ldots, W_N$, with sample size N.
- We are interested in testing the hypothesis that the average treatment effect is zero:

$$H_0 : \ \mathbb{E}[Y_i(1) - Y_i(0)] = 0, \qquad \text{against } H_a : \ \mathbb{E}[Y_i(1) - Y_i(0)] \neq 0.$$

- The size of the test is again α, and we want power β, against an alternative that the average treatment effect is $\tau = d \cdot \sigma$.
- Let $\gamma = \sum_i W_i/N$ be the proportion of treated units. We look for the minimum sample size $N = N_0 + N_1$, as a function of α, β, τ, σ^2, and γ.

RESULT

- Required Sample Size

$$N = \frac{\left(\Phi^{-1}\left(\beta\right) + \Phi^{-1}\left(1 - \alpha/2\right)\right)^2}{(\tau^2/\sigma^2) \cdot \gamma \cdot (1 - \gamma)}.$$

- For example, suppose we choose $\gamma = 0.5$ (equal sample sizes), $\alpha = 0.05$ (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

$$N = \frac{\left(\Phi^{-1}\left(0.8\right) + \Phi^{-1}\left(0.975\right)\right)^2}{0.1^2 \cdot 0.5^2}$$

$$= 4 \times \frac{\left(\Phi^{-1}\left(0.8\right) + \Phi^{-1}\left(0.975\right)\right)^2}{0.1^2} = 4 \times 785.89 \approx 3,119.6.$$

LEWIS-RAO PAPER ON POWER FOR ONLINE AD CAMPAIGNS

- Test with power $\beta = 0.8$, test at level $\alpha = 0.05$. Equal size control group and treatment group, $\gamma = 0.5$.
- We are looking at effect of size τ =\$0.35, because ads are cheap. Mean value of outcome is \$7, standard deviation σ =\$75.

  item Then:

$$N = \frac{\left(\Phi^{-1}\left(\beta\right) + \Phi^{-1}\left(1 - \alpha/2\right)\right)^2}{(\tau^2/\sigma^2)\gamma(1-\gamma)} = 1.44M.$$

- Need very large experiments because coefficient of variation is large, and meaningful treatment effect is small!

REFERENCES

- Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94, no. 4 (2004): 991-1013.
- Box, George EP, William H. Hunter, and Stuart Hunter.*Statistics for experimenters.* Vol. 664. New York: John Wiley and sons, 1978.
- Donner, Allan, Neil Klar, and Neil S. Klar. *Design and analysis of cluster randomization trials in health research.* Vol. 27. London: Arnold, 2000.
- Imai, Kosuke, Gary King, and Clayton Nall. "The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation." (2009): 29-53.
- Lewis, Randall A., and Justin M. Rao. "The unfavorable economics of measuring the returns to advertising." *The Quarterly Journal of Economics* 130, no. 4 (2015): 1941-1973.

REFERENCES (CTD)

- Morgan, Kari Lock, and Donald B. Rubin. "Rerandomization to improve covariate balance in experiments." *Annals of Statistics* 40, no. 2 (2012): 1263-1282.
- Snedecor, G. W., and W. G. Cochran. *Statistical methods.*(1989).