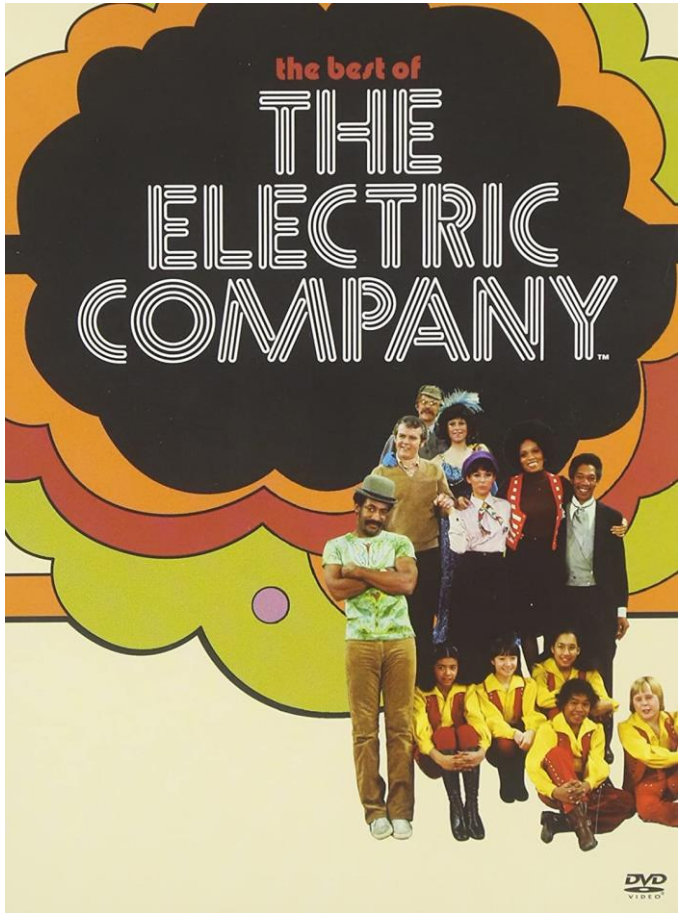
Lecture 8
Pairwise randomized experiments

Outline

- pairwise randomized experiment
  - Fisher's exact p-value
  - Neyman's repeated sampling approach
  - Regression analysis
  - How to find strata / pairs?
  - R example
- Suggested reading: Imbens and Rubin Section 10.1 -10.6; Peng's book Section 7.1-7.6

The Electric Company experiment



Would showing an educational television program at school improve reading?

Youngstown: two first-grade classes in each of eight schools.

Within each school, randomly assign one class to watch the show. The other continues the regular reading curriculum.

Unit: class. Outcome: class-average post-test score.
Baseline covariate: class-average pretest score.

Ball, Bogatz, Rubin & Beaton (1973); Imbens & Rubin, Chapter 10.

Pairing after stratification

Lecture 7: compare treated and control units within baseline strata.

$$N_j = 2, \qquad N_{t,j} = N_{c,j} = 1, \qquad q_j = \frac{1}{J}$$

Today: one treatment and one control in every pair.

Fisher: swap labels within each pair.
Neyman: average the pair differences and estimate uncertainty.
Regression: adjust remaining baseline differences.

With one observed outcome per arm, which variance estimate fails?

Eight paired comparisons

Each row is one school: a treated class and a control class.

School	Pretest C	Pretest T	Post-test C	Post-test T	Difference
1	12.9	12.0	54.6	60.6	6.0
2	15.1	12.3	56.5	55.5	−1.0
3	16.8	17.2	75.2	84.8	9.6
4	15.8	18.9	75.6	101.9	26.3
5	13.9	15.3	55.3	70.6	15.3
6	14.5	16.6	59.3	78.4	19.1
7	17.0	16.0	87.0	84.2	−2.8
8	15.8	20.1	73.7	108.6	34.9

How many differences are positive?

Does each difference equal a causal effect for that school?

The target and the observed pair difference

Pair j contains two fixed classes, $k = 1$ and 2, with potential outcomes $Y_{jk}(0), Y_{jk}(1)$.

$$\tau_j = \frac{1}{2}\sum_{k=1}^{2}[Y_{jk}(1) - Y_{jk}(0)], \qquad \tau_{fs} = \frac{1}{J}\sum_{j=1}^{J}\tau_j$$

Observed treatment-minus-control difference and its average:

$$D_j = Y_{t,j}^{obs} - Y_{c,j}^{obs}, \qquad \hat{\tau} = \overline{D} = \frac{1}{J}\sum_{j=1}^{J} D_j$$

Our target averages effects across classes, giving small and large classes equal weight.

Assume a well-defined intervention and no interference between class units.

The assignment mechanism within pairs

Form pairs before assignment. Independently randomize one unit per pair.

Pair 1: A, B	**Pair 2: C, D**	**Probability**
Treat A	Treat C	1/4
Treat A	Treat D	1/4
Treat B	Treat C	1/4
Treat B	Treat D	1/4

$$\text{Allowed assignments: } 2^J, \qquad \Pr(W = w \mid \text{pairs}) = 2^{-J}$$

Which assignments would shuffling all 16 labels incorrectly allow?

Fisher's test by sign flips

- Sharp null H_0: $Y_{jk}(1) = Y_{jk}(0)$ for every class

- Statistic: absolute mean difference

$$T = |\overline{D}|, \qquad T_{\text{obs}} = 13.425$$

- Under each reassignment:

$$D_j^* = \begin{cases} D_j & \text{keep labels} \\ -D_j & \text{swap labels} \end{cases}$$

  - Recompute $T^* = |\overline{D}^*|$

- As each pair has exactly one treatment and one control
  - We don't need to consider different weights
  - No worry of Simpson's paradox

- Rank statistics
  - Use population ranks:
    $T = |\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|$
  - Use within-pair ranks

$$T^{\text{rank,pair}} = \left| \frac{2}{N} \sum_{j=1}^{N/2} \left(\mathbf{1}_{Y_{j,1}^{\text{obs}} > Y_{j,0}^{\text{obs}}} - \mathbf{1}_{Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}} \right) \right|$$

Different statistics use different information

Statistic	Observed value	Two-sided exact p
Absolute mean difference	13.425	8/256 = 0.03125
Difference in global mean ranks	3.75	8/256 = 0.03125
Within-pair ranks (sign statistic)	0.50	74/256 = 0.28906

- Rank v.s. within-pair rank
  - Both can reduce the sensitivity to outliers
  - Using within-pair ranks can have more power when there is substantial variation in the level of the outcomes between pairs
  - Otherwise, using within-pair ranks loses power as it treats small within-pair differences (which may be due to random noises) equally with large within-pair differences

Which statistic would you expect to be more powerful in each scenario:
(1) positive differences are much larger than negative differences
(2) most pairs show modest positive differences, but a few extreme negative differences dominate the mean?

Neyman's approach: target and point estimator

- Target: SATE or PATE τ

$$\tau_j = \frac{1}{2}\sum_{k=1}^{2}[Y_{jk}(1) - Y_{jk}(0)], \qquad \tau_{fs} = \frac{1}{J}\sum_{j=1}^{J}\tau_j$$

- Estimator: average the observed treatment-minus-control differences

$$D_j = Y_{t,j}^{\text{obs}} - Y_{c,j}^{\text{obs}}, \qquad \hat{\tau} = \overline{D} = \frac{1}{J}\sum_{j=1}^{J} D_j$$

- Randomization gives each assignment within a pair probability 1/2:

$$\mathrm{E}_W[D_j] = \frac{1}{2}\{Y_{j1}(1) - Y_{j2}(0)\} + \frac{1}{2}\{Y_{j2}(1) - Y_{j1}(0)\} = \tau_j$$

$$\mathrm{E}_W[\hat{\tau}] = \tau_{fs}$$

- Paired experiment is a special stratified randomized experiment

Neyman's approach: estimating uncertainty

- One treated and one control unit per pair: we cannot compute the within-arm sample variances used in Lecture 7.

$$\mathrm{Var}_W(\hat{\tau}) = \frac{1}{J^2}\sum_{j=1}^{J}\mathrm{Var}_W(D_j)$$

Assignments are independent across pairs.

- Use the sample variance ${s_D}^2$ of the J observed pair differences:

$$\hat{V} = \frac{{s_D}^2}{J} = \frac{1}{J(J-1)}\sum_{j=1}^{J}(D_j - \overline{D})^2$$

- Each D_j has mean τ_j instead of τ, which may differ across the fixed pairs:

$$\mathrm{E}_W[\hat{V}] = \mathrm{Var}_W(\hat{\tau}) + \frac{1}{J(J-1)}\sum_{j=1}^{J}(\tau_j - \tau_{fs})^2$$

  - $\hat{V}$ is conservative in expectation, and unbiased if all τ_j are equal.

If the two units in every pair have identical control potential outcomes, can the Neyman variance estimator still be conservative? Why?

Neyman inference: the Electric Company example

Eight schools give $J = 8$ pair differences. Use $\mathrm{SE} = \sqrt{\hat{V}}$.

Analysis	Estimate	Estimated SE
Average of eight pair differences	13.43	4.64
Two-sample formula ignoring pairs	13.43	7.85

Pair	Outcome for Control Unit	Outcome for Treated Unit	Difference
1	54.6	60.6	6.0
2	56.5	55.5	−1.0
3	75.2	84.8	9.6
4	75.6	101.9	26.3
5	55.3	70.6	15.3
6	59.3	78.4	19.1
7	87.0	84.2	−2.8
8	73.7	108.6	34.9
Mean	67.2	80.6	13.4
(S.D.)	(12.2)	(18.6)	(13.1)

Approximate 95% confidence intervals for the average effect

$$\hat{\tau} \pm 1.96\sqrt{\hat{V}} = 13.425 \pm 1.96 \times 4.636 = [4.3, 22.5]$$

$$\hat{\tau} \pm t_{0.975,7}\sqrt{\hat{V}} = 13.425 \pm 2.365 \times 4.636 = [2.5, 24.4]$$

Why is the paired SE smaller?

- Differences remove shared school variation.
- Ignoring the pairing includes that between-school variation in the SE.

Regression: model and notation

Two units per pair cannot support separate covariate regressions.
Pool slopes across pairs and allow each pair its own baseline level.

How to build a reasonable regression framework?

- Working model for $Y_{jk}(w)$, conditional on X and the pair indicators, with w = 0, 1:

$$\mathrm{E}[Y_{jk}(w) \mid X] = \alpha_j + \beta^{\mathrm{T}} X_{jk} + w\{\tau + \gamma^{\mathrm{T}}(X_{jk} - \mu_X)\}$$

Parameter	Meaning
α_{j}	Pair-specific baseline level
β	Covariate slopes for the control outcome
τ	PATE
γ	How the treatment effect changes with X

Randomization links this model to the observed treated and control outcomes:

$$\mathrm{E}[Y_{jk}{}^{\mathrm{obs}} \mid W_{jk} = w, X] = \mathrm{E}[Y_{jk}(w) \mid X]$$

Regression: deriving the pair-difference model

Within pair j, t and c label the treated and control units.

$$\Delta X_j = X_{t,j} - X_{c,j}, \qquad \overline{X_j} = \frac{X_{t,j} + X_{c,j}}{2}$$

Subtract the two conditional means. The shared intercept α_j cancels:

$$\mathrm{E}[D_j \mid W, X] = \tau + \gamma^{\mathrm{T}}(X_{t,j} - \mu_X) + \beta^{\mathrm{T}}\Delta X_j$$

$$= \tau + \gamma^{\mathrm{T}}(\overline{X_j} - \mu_X) + (\beta + \frac{\gamma}{2})^{\mathrm{T}}\Delta X_j$$

Fit one row per pair: regress D_j on $\left(\overline{X_j} - \bar{X}\right)$ and ΔX_j, with an intercept.
The intercept estimates τ. The slope on ΔX_j is $\beta + \gamma/2$.

If γ = 0, the conditional treatment effect does not vary with X:

$$D_j = \tau + \beta^{\mathrm{T}}\Delta X_j + \varepsilon_j, \qquad \mathrm{E}[\varepsilon_j \mid W, X] = 0$$

Regression: the Electric Company example

X_{jk} is the class-average pretest score. Fit the γ = 0 model to J = 8 schools.

$$D_j = \tau + \beta \Delta X_j + \varepsilon_j$$

Regression across schools	Intercept estimate	Model SE	Residual df
D ~ 1	13.43	4.64	7
D ~ ΔX	8.99	1.41	6

If treated classes had higher pretest scores, would you expect adjustment to increase or decrease the estimated effect? Why?

The fitted intercept removes the estimated contribution of baseline imbalance:

$$\hat{\tau}_{\text{adj}} = \overline{D} - \hat{\beta}\overline{\Delta X} = 13.425 - 5.371 \times 0.825 \approx 8.99$$

Treated classes started 0.825 pretest points higher on average.
Pretest differences predict outcome differences, reducing the residual variation.

How to perform stratification / pairing

- Implementation based on convenience

- Univariate blocking: discrete or discretized variable
- Multivariate blocking: Mahalanobis distance

$$D(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \widehat{\mathbb{V}(\mathbf{X})}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

Greedy algorithms

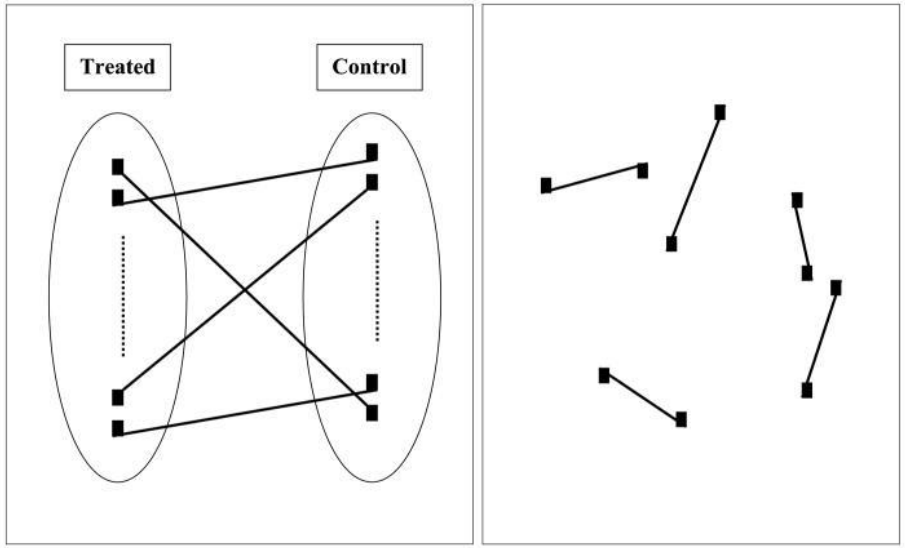
- Matching: pair two units with the shortest distance, set them aside, and repeat
- Blocking: randomly choose one unit and choose N_j units with the shortest distances, set them aside, and repeat

But the resulting matches may not be optimal

For values 0, 2, 3, 5, which pairs does greedy matching choose?
Can another pairing have a smaller total absolute distance?

Optimal matching

- D: $N \times N$ matrix of pairwise distance or a cost matrix
- Optimal matching
  - Binary $N \times N$ matching matrix: M with $M_{ij} \in \{0,1\}$
  - Optimization problem:
    $\min_{M} \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} D_{ij}$ subject to $\sum_{i=1}^{N} M_{ij} = 1$ for all j
    where we set $D_{ii} = \infty$ for all i
  - M also need to be symmetric



Bipartite Matching

Nonbipartite Matching

- Nonbipartite matching

- Computational cost $O(n^3)$
- Derigs' algorithm: implemented in the R package nbpMatching
  https://cran.r-project.org/web/packages/nbpMatching/

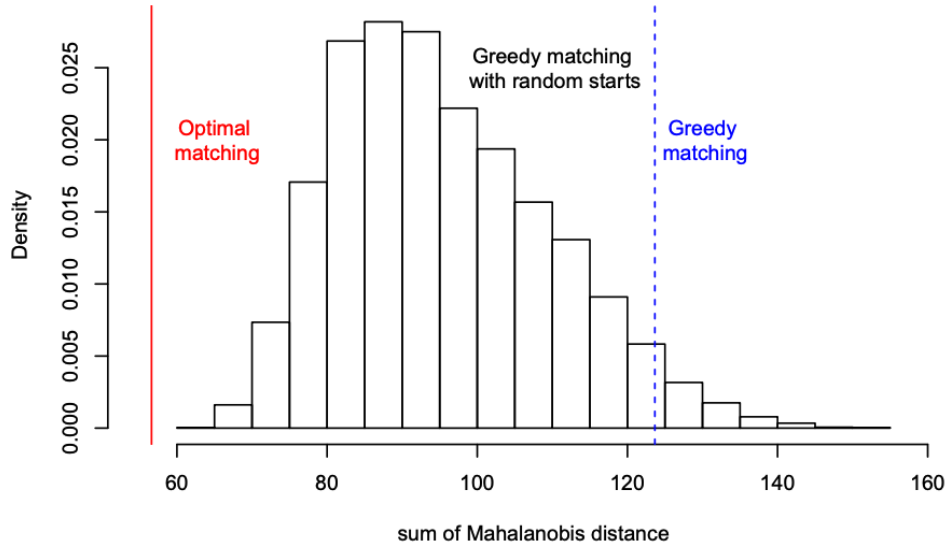
Seguro Popular: Mexico's public health-insurance program

Units: health-facility catchment areas. Match on baseline characteristics.

74 pairs randomized; household outcomes surveyed in 50 pairs.

Assignment: encouragement to enroll in Seguro Popular health insurance (at the community level), together with improvements to health services.

Actual household enrollment did not perfectly follow assignment.



Matching illustration: 100 clusters in 50 pairs.

King et al. (2009), The Lancet.

Paired experiments and treatment receipt

The design determines the assignments used for Fisher's test.

Average pair differences estimate SATE. Their variation across pairs supplies a computable, conservative variance estimate.

Baseline adjustment can improve precision, with additional modeling choices.

Next: assignment can differ from treatment received.
What does randomization identify when people do not comply?