

Experiments

Stat 256, Causality

Chad Hazlett

UCLA

`chazlett@ucla.edu`

Spring 2026

Identification vs. Estimation

Goal: learn about an *unobservable, counterfactual-dependent* quantity of interest using *observed* data.

Two inferential hurdles:

- 1 **Identification**: With data on an entire population, can you recover your QoI?
- 2 **Estimation**: With finite data on a sample, how well can you learn it?

Identification precedes estimation.

For each identification strategy we discuss:

- 1 the **identification assumption**
- 2 information from outside the data supporting it
- 3 testable implications (limited)
- 4 estimators
- 5 sensitivity to violations

Classical Randomized Experiment

Setup:

- Units: $i = 1, \dots, N$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: Y_{0i}, Y_{1i}
- Observed outcome: $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$ (consistency)
- $N_1 = \sum_{i=1}^N D_i$, $N_0 = N - N_1$

Random assignment can take several forms:

- **Complete randomization**: Exactly N_1 units treated (draw without replacement)
- **Simple (Bernoulli) randomization**: Each unit independently assigned with $\Pr(D_i = 1) = p$

Randomization = Ignorability

Random assignment of D_i guarantees:

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i$$

That is: the potential outcomes a unit *would have* under treatment or control are statistically independent of which group they are assigned to.

This is the **identification assumption** for experiments. Let's see what it buys us.

Identification of ATE Under Experiments

Quantity of Interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

Identification: Using $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i$,

$$\begin{aligned}\mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \quad (\text{by ignorability}) \\ &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \quad (\text{by consistency}) \\ &= \text{DIM estimand}\end{aligned}$$

Practice these two steps. They are the core of every ID argument we will see.

Identification of Full Distributions

With $(Y_1, Y_0) \perp\!\!\!\perp D$, we identify more than means.

Identification Result

Let $F_{Y_d}(y)$ denote the CDF of Y_d . Then

$$F_{Y_0}(y) = \Pr(Y_0 \leq y) = \Pr(Y_0 \leq y \mid D = 0) = \Pr(Y \leq y \mid D = 0)$$

$$F_{Y_1}(y) = \Pr(Y \leq y \mid D = 1)$$

So the effect at any quantile $\theta \in [0, 1]$ is identified:

$$\alpha_\theta = Q_\theta(Y_1) - Q_\theta(Y_0) = Q_\theta(Y \mid D=1) - Q_\theta(Y \mid D=0)$$

Note: the *quantile of the effect* $Q_\theta(Y_1 - Y_0)$ is **not** identified without further assumptions — the difference of quantiles is not the quantile of the difference.

SATE and PATE

- **SATE:** $\tau_{SATE} = \frac{1}{N} \sum_{i=1}^N (Y_{1i} - Y_{0i})$
 - Inferences limited to units in the study
 - Resampling notion can vary:
 - *Design-based:* $\mathbf{X}$, POs fixed; only treatment assignment is re-randomized
 - *Model-based:* $\mathbf{X}$, D fixed; residual influences on Y are random
- **PATE:** $\tau_{PATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$ over a superpopulation
 - Generalizes the above to also redraw units from a larger superpopulation/DGP
- If sampling is random from the population, $\mathbb{E}[\text{SATE}] = \text{PATE}$. We are often uninterested in distinguishing them.

Variance of the DIM: PATE

Start from the variance of the mean difference:

$$\mathbb{V}(\hat{\tau}) = \mathbb{V}(\bar{Y}_1) + \mathbb{V}(\bar{Y}_0) - 2\text{Cov}(\bar{Y}_1, \bar{Y}_0)$$

For the PATE (superpopulation view): treated and control units are independent draws, so $\text{Cov}(\bar{Y}_1, \bar{Y}_0) = 0$:

$$\mathbb{V}_{\text{Neyman}}(\hat{\tau}) = \frac{\sigma_{Y_1}^2}{N_1} + \frac{\sigma_{Y_0}^2}{N_0}$$

This is the **Neyman variance**.

Plug-in estimator: replace population variances with sample variances:

$$\widehat{SE}(\hat{\tau}) = \sqrt{\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}}, \quad S_d^2 = \frac{1}{N_d - 1} \sum_{i:D_i=d} (Y_i - \bar{Y}_d)^2$$

Variance of the DIM: SATE

For the SATE (finite-population / design-based view): the potential outcomes (Y_{1i}, Y_{0i}) are fixed. Randomness comes only from treatment assignment.

Under complete randomization, the N_1 treated units are a sample *without replacement* from the N fixed units. Two consequences:

- $\mathbb{V}(\bar{Y}_d)$ is *smaller* than $\sigma_{Y_d}^2/N_d$ by a finite population correction factor $(1 - N_d/N)$
- $\text{Cov}(\bar{Y}_1, \bar{Y}_0) \neq 0$ because the two groups are complementary (selecting one determines the other)

Working through the finite-population sampling algebra, these combine to give:

$$\mathbb{V}_{SATE}(\hat{\tau}) = \frac{\sigma_{Y_1}^2}{N_1} + \frac{\sigma_{Y_0}^2}{N_0} - \frac{\sigma_{\tau}^2}{N}$$

where $\sigma_{\tau}^2 = \text{Var}(\tau_i) \geq 0$.

So the Neyman variance is **conservative** for the SATE. The gap is σ_{τ}^2/N : zero if effects are constant, and shrinking with N regardless.

Regression to Estimate ATE

Estimator (Regression with Experiments)

$$\begin{aligned}
 Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
 &= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
 &= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\hat{\tau}} D_i + \underbrace{\{(Y_{0i} - \bar{Y}_0) + D_i[(Y_{1i} - \bar{Y}_1) - (Y_{0i} - \bar{Y}_0)]\}}_{\varepsilon_i}
 \end{aligned}$$

- This does **not** assume constant effects: $\hat{\tau}$ averages over heterogeneous τ_i .
- The residual ε_i has different variances under $D = 1$ and $D = 0 \Rightarrow$ heteroskedasticity.
- We need a heteroskedasticity-robust SE.

The Sandwich Variance Estimator

For OLS with $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$:

$$\begin{aligned}\widehat{\text{Var}}(\hat{\beta} \mid \mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \widehat{\text{Var}}(\mathbf{Y} \mid \mathbf{X}) \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where $\hat{\Omega} = \text{diag}(\hat{\omega}_1, \dots, \hat{\omega}_N)$ estimates the conditional variance of each Y_i .

The choice of $\hat{\omega}_i$ defines the variant:

HC0: $\hat{\omega}_i = \hat{\varepsilon}_i^2$

HC1: $\hat{\omega}_i = \frac{N}{N-p} \hat{\varepsilon}_i^2$ (global degrees-of-freedom correction)

HC2: $\hat{\omega}_i = \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}$ (leverage correction, h_{ii} = i th diagonal of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$)

With HC0, the “meat” of the sandwich becomes:

$$\mathbf{X}'\hat{\Omega}\mathbf{X} = \sum_{i=1}^N \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i'$$

where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i'\hat{\beta}$ and $\mathbf{x}_i$ is the i th row of $\mathbf{X}$.

Sandwich for the Experimental Regression

The $[2, 2]$ element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}(\hat{\varepsilon}_i^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ is just

$$\widehat{\mathbb{V}}(\hat{\tau}) = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'\text{diag}(\hat{\varepsilon}_i^2)\tilde{D}(\tilde{D}'\tilde{D})^{-1}$$

where $\tilde{D}$ is the vector $D_i - \bar{D}$. This sorts and splits into two pieces:

$$= \frac{\sum_{i: D_i=1} \hat{\varepsilon}_i^2 \tilde{D}_i^2}{(\tilde{D}'\tilde{D})^2} + \frac{\sum_{i: D_i=0} \hat{\varepsilon}_i^2 \tilde{D}_i^2}{(\tilde{D}'\tilde{D})^2}$$

Noting that $\tilde{D}_i = N_0/N$ for treated, and $-N_1/N$ for controls, and that $\tilde{D}'\tilde{D} = N_1 N_0/N$, after some tedious algebra we get this to equal

$$= \frac{\widehat{\text{Var}}(Y|D=1)}{N_1} + \frac{\widehat{\text{Var}}(Y|D=0)}{N_0}$$

Further, if you construct $\widehat{\text{Var}}(Y|D=d)$ with the $N_d - 1$ correction then you have the HC2 correction and this matches Neyman exactly.

Example: Effect of Job Training on Earnings

- Treatment: $N_1 = 7,487$, $\bar{Y}_1 = \$16,199$, $S_1 = \$17,038$
- Control: $N_0 = 3,717$, $\bar{Y}_0 = \$15,040$, $S_0 = \$16,180$
- $\hat{\tau} = 16,199 - 15,040 = \$1,159$
- $\widehat{SE} = \sqrt{\frac{17,038^2}{7,487} + \frac{16,180^2}{3,717}} \approx \$330$
- $t = 1,159 / 330 \approx 3.5$, $p \approx 0.0005$

All exactly the same thing:

- DIM with Neyman SE
- Welch (unequal variance) t -test
- `lm(earnings ~ D) with HC2 SEs via sandwich::vcovHC(fit, type="HC2")`

Fisher's Exact Test

You already know the large- N test:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \quad (\text{weak null})$$

Fisher's Exact Test uses a stronger hypothesis:

$$H_0 : Y_{1i} = Y_{0i} \quad \forall i \quad (\text{sharp null of no effect})$$

Key idea: Under the sharp null, we “observe” all potential outcomes — so we can compute the test statistic under every possible treatment assignment.

Procedure:

- 1 Compute $\hat{\tau}_{\text{obs}}$ from actual assignment
- 2 For every (or many) alternative assignments $\omega \in \Omega$, compute $\hat{\tau}(\omega)$ using the filled-in PO schedule
- 3 $p\text{-value} = \Pr_{\omega}(|\hat{\tau}(\omega)| \geq |\hat{\tau}_{\text{obs}}|)$

Fisher's Exact Test: Small-Sample Example

i	Y_{1i}	Y_{0i}	D_i
1	3	?	1
2	1	?	1
3	?	0	0
4	?	1	0
$\hat{\tau}_{\text{obs}}$			1.5

What do we know given $H_0 : Y_{1i} = Y_{0i} \ \forall i$?

i	Y_{1i}	Y_{0i}	D_i
1	3	3	1
2	1	1	1
3	0	0	0
4	1	1	0
$\hat{\tau}_{\text{obs}}$			1.5

Chern null fills in the full schedule. Now enumerate (4) assignments

Scaling Up

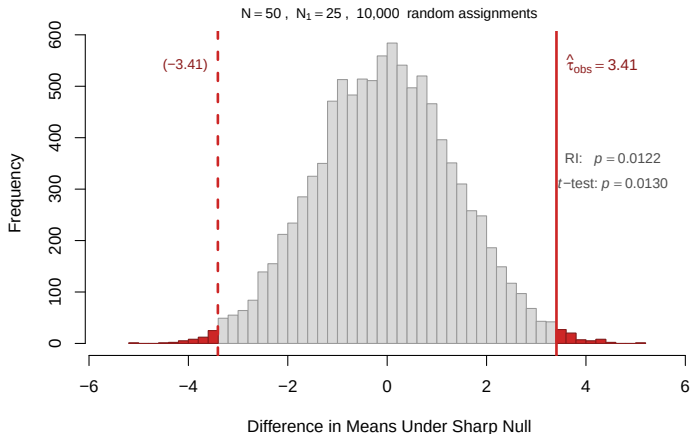
With 4 units and 2 treated: $\binom{4}{2} = 6$ assignments — enumerate all.

With 50 units and 25 treated: $\binom{50}{25} \approx 1.26 \times 10^{14}$.

Solution: **sample** from Ω .

- Re-draw treatment assignment many times (e.g. 10,000)
- Compute $\hat{\tau}(\omega)$ each time under the sharp null
- Compare observed $\hat{\tau}$ to this simulated null distribution

Randomization Inference in Practice



Histogram: 10,000 re-randomizations under the sharp null.

Solid line: observed $\hat{\tau}$. Shaded: values at least as extreme $\Rightarrow$ p-value.

How Do These Inferences Differ?

Finite-sample exactness.

- The RI p-value is exact under the sharp null — no CLT required.
- When N is large, the null distribution is approximately normal and you get nearly the same answer as the t -test.

Sharp null $\neq$ weak null, but close in practice.

- The sharp null says $Y_{1i} = Y_{0i}$ for *every* unit.
- But a summary statistic like the DIM is only sensitive to effects that *accumulate* — not to effects that cancel. Rejecting the sharp null via the DIM really tells you “effects exist *and* don’t average to zero,” which is close to the weak null.

RI vs. the bootstrap.

- (Paired) bootstrap resamples *units* — “what if we’d drawn different people?” (superpopulation).
- RI re-randomizes *treatment* within the fixed sample — “what if the same people got different assignments?” (design-based).

Bottom line: RI most valuable in **small samples** where CLT inference is suspect.

When Observations Are Not Independent

Everything so far assumed (conditionally) independent observations. Often violated:

- Voters in the same precinct
- Students in the same classroom
- Repeated measures on the same person

Two distinct problems:

- 1 **Clustered assignment.** Treatment assigned at group level (village, school, precinct). This inflates the *actual* variance of $\hat{\tau}$ — you have less independent information than N suggests.
- 2 **Correlated residuals.** Suppose there are cluster level influences, U_c , driving within-group correlation in ε_i . When combined with cluster-assignment or non-random assignment allowing D correlated with U_c within clusters, this will make HC SE wrong.

A DGP for Clustering

Units i belong to clusters $c \in \{1, \dots, C\}$:

$$Y_i = \tau D_i + U_{c(i)} + \varepsilon_i$$

- U_c is a **cluster-level** component (precinct culture, school quality, ...)
- ε_i is individual-level noise, independent across units

Individual-level randomization: each cluster has both treated and control units. The within-cluster DIM cancels U_c from the comparison — both arms share the same U_c .

Cluster-randomization: $D_i = D_{c(i)}$. Now in any given sample, U_c and D are correlated (the treated clusters happen to have higher or lower U_c than the control clusters), shifting $\hat{\tau}$ away from τ . Across samples that shift changes direction, inflating $\text{Var}(\hat{\tau})$.

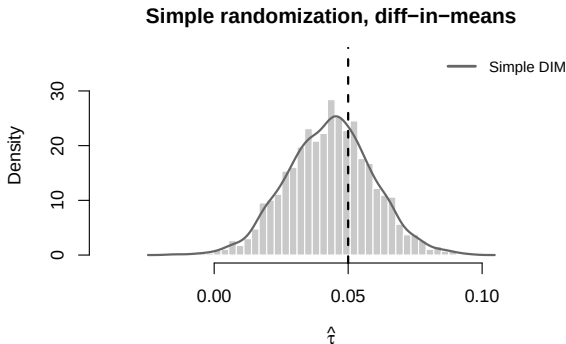
Design effect quantifies the loss. With average cluster size m and intra-cluster correlation ρ (fraction of $\text{Var}(Y)$ from U_c):

$$\text{DEFF} = 1 + (m - 1)\rho, \quad N_{\text{eff}} = \frac{N}{\text{DEFF}}.$$

Example: $N = 1000$, $m = 20$, $\rho = 0.2 \Rightarrow \text{DEFF} = 1 + 19 \times 0.2 = 4.8$, so $N_{\text{eff}} \approx 208$.

Simulation: Simple DIM

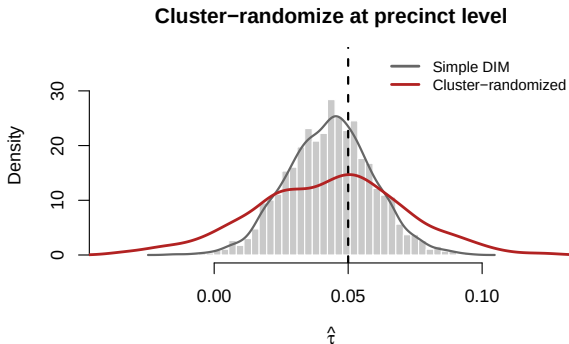
Running example: a voter-turnout experiment with $N = 1000$ voters across 50 precincts, $\tau = 0.01$.



- Blue: sampling distribution of $\hat{\tau}$ under individual-level randomization.
- $SD(\text{actual}) = 0.016$; Mean HC2 SE = 0.016. ✓

Simulation: Cluster-Randomize at the Precinct Level

Same DGP, but now everyone in a precinct gets the same D .



- Red: sampling distribution under cluster-randomization — much wider!
- SD(actual) = 0.028, but Mean HC2 SE = 0.016. **HC2 is wrong.**

The SE Is Wrong — but CR-SEs Fix It

Design / Estimator	SD(actual)	HC2 SE	CR-SE (precinct)
Simple DIM	0.016	0.016	0.016
Cluster-randomized	0.028	0.016	0.028

- HC2 assumes independent residuals. But U_c + cluster-assignment means residuals within a cluster are **identical** in their U_c component and identical in D — the worst case for dependence.
- Cluster-robust SE (clustered at precinct level) recovers the actual sampling variance. In R:


```
sandwich::vcovCL(fit, cluster = precinct_id)
```
- U_c type influences can mess up HC SEs **without** cluster-assignment if $cov(U_c, D) \neq 0$ within clusters
 - issue in non-randomized setting (e.g. observational panel data)
 - occurs even if U_c observed, e.g. cluster-level covariates.

Cluster-Robust SEs: The Formula

The cluster-robust (“CR”) sandwich estimator:

$$\widehat{\mathbf{V}}_{CR}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{c=1}^C \mathbf{x}'_c \hat{\varepsilon}_c \hat{\varepsilon}'_c \mathbf{x}_c \right) (\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{X}_c$, $\hat{\varepsilon}_c$ are the rows and residuals for cluster c .

Allows **arbitrary** intra-cluster correlation; assumes clusters are independent.

Practical notes:

- Cluster on the source of dependence in your residuals, not on every grouping you can think of.
- With few clusters ($C \lesssim 30\text{--}50$), CR-SEs are downward biased. Alternatives: wild cluster bootstrap (`fwildclusterboot` in R), RI at the cluster level, or aggregate to cluster means.

Unbiased Is Not Enough

Under random assignment, $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$ is unbiased for the ATE.

But “unbiased” is a statement about averages over hypothetical repetitions. In any *single* experiment, $\hat{\tau}$ can be far from τ .

What we care about is **root mean squared error**:

$$\text{RMSE}(\hat{\tau}) = \sqrt{\mathbb{V}(\hat{\tau}) + \text{Bias}^2} = \sqrt{\mathbb{V}(\hat{\tau})} \quad (\text{when unbiased})$$

So the goal is to **reduce variance**. Two levers:

- 1 **Design**: blocking (build balance in before randomization)
- 2 **Analysis**: covariate adjustment (remove imbalance after the fact)

Both exploit the same idea: a pre-treatment covariate X that predicts Y can absorb residual variation.

Blocking: Build the Balance In

Idea: pre-stratify on a strong predictor of Y , then randomize treatment *within* each stratum (block).

- Guarantees exact balance on the blocking variable, in your sample — not just in expectation.
- Within-block variation in Y is smaller, so the within-block DIM is sharper.
- Randomization still balances unobserved influences, in expectation.

Estimator: within each block j , compute $\hat{\tau}_j$. Then weight by block size:

$$\hat{\tau}_B = \sum_{j=1}^J \frac{N_j}{N} \hat{\tau}_j$$

Equivalently (when treatment probability is the same in every block): OLS with block fixed effects.

Estimation: regression with block dummies

Two regressions for the same data:

$$(CR) \quad Y_i = \alpha + \tau_{CR} D_i + \epsilon_i$$

$$(BR) \quad Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^J \beta_j B_{j[i]} + \epsilon_i^*$$

where $B_{j[i]}$ is a dummy for the j -th block.

Under iid sampling, the sampling variances are:

$$\mathbb{V}(\hat{\tau}_{CR}) = \frac{\sigma_{\epsilon}^2}{\sum_i (D_i - \bar{D})^2}, \quad \mathbb{V}(\hat{\tau}_{BR}) = \frac{\sigma_{\epsilon^*}^2}{\sum_i (D_i - \bar{D})^2 (1 - R_D^2)}$$

where R_D^2 is from regressing D on the block dummies.

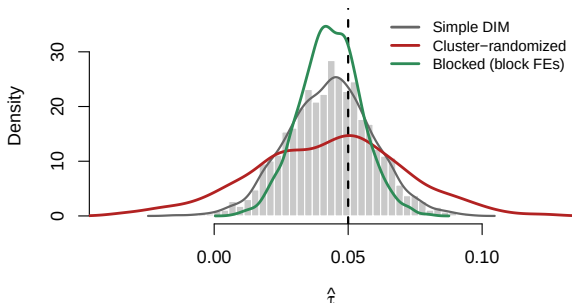
When treatment probability is equal in every block, $R_D^2 = 0$ and the denominators match. So we get pure benefit from $\sigma_{\epsilon^*}^2 < \sigma_{\epsilon}^2$ — the residual variance after block dummies absorb between-block variation in Y .

(DoF cost: $n-k-1$ vs $n-2$ in denominator of $\hat{\sigma}^2$. Negligible unless many, small blocks.)

Simulation: + Blocking

Block on quartiles of a pre-treatment predictor (e.g., `past_turnout`).

Add blocking on past-turnout quartiles



- Green: distribution of $\hat{\tau}$ under blocking.
- $SD(\text{actual}) = 0.011$; tighter than simple DIM (0.016) — and dramatically tighter than cluster-randomized (0.028).
- The gain: blocks balance important X in every sample.

Covariate Adjustment: The Lin Estimator

Consider this regression:

$$Y_i = \alpha + \tau D_i + \beta' X_i + \varepsilon_i$$

Lin (2013): demean X and interact with D :

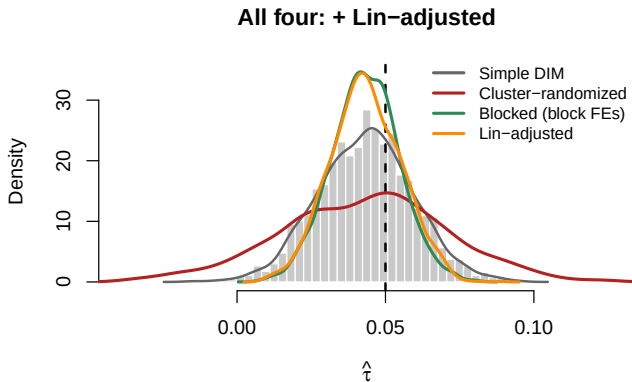
$$Y_i = \alpha + \tau_{Lin} D_i + \beta'_1 (X_i - \bar{X}) + \beta'_2 D_i \cdot (X_i - \bar{X}) + \varepsilon_i$$

- τ_{Lin} is directly interpretable as the ATE.
- Avoids misspecification biases under heterogeneous effects that otherwise lead to “weird weights” and Freedman’s $O(1/N)$ bias critique.
- Use HC2 SEs. (Pre-specify the covariates if you are worried)

Preview: this is nothing but *regression imputation* / *G-computation* / the *T-learner* — fit $\hat{\mu}_d(X)$ separately per arm, impute both POs, average. We will return to this.

Best practice: “Block on what you can, adjust for what you can’t.”

Simulation: all four designs/estimators



- Orange: Lin-adjusted under individual randomization.
- SDs: Simple DIM = 0.016; Cluster = 0.028; Blocked = 0.011; Lin = 0.012.
- Blocking and Lin adjustment give similar gains here — both exploit the same predictor (`past_turnout`).

Takeaways

- **Randomization = ignorability.** Identification uses two steps: ignorability, then consistency.
- **DIM = Neyman = Welch t = OLS with HC2.**
- **Randomization inference** is exact in finite samples and design-based. Rejecting the sharp null via DIM is close to rejecting the weak null.
- **Clustering** inflates variance and/or breaks SE estimation. Cluster-robust SEs handle it with enough clusters.
- **Blocking and covariate adjustment** reduce variance — unbiasedness alone is not the goal. Block on what you can, adjust for what you can't.
- **Next:** What if we can't randomize?

Appendix

Why Neyman Is Conservative for SATE

Recall the exact variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\sigma_{Y_1}^2}{N_1} + \frac{\sigma_{Y_0}^2}{N_0} - \frac{\sigma_\tau^2}{N}$$

The Neyman approximation drops $-\sigma_\tau^2/N$:

$$\mathbb{V}_{\text{Neyman}}(\hat{\tau}) = \frac{\sigma_{Y_1}^2}{N_1} + \frac{\sigma_{Y_0}^2}{N_0}$$

Since $\sigma_\tau^2 \geq 0$, we have $\mathbb{V}_{\text{Neyman}} \geq \mathbb{V}(\hat{\tau})$.

How large is the gap?

- If effects are constant ($\tau_i = \tau$ for all i), then $\sigma_\tau^2 = 0$ and Neyman is exact.
- If effects vary, Neyman overestimates. The gap is σ_τ^2/N , which shrinks with N .

Either way, CIs based on Neyman SEs have *at least* nominal coverage for the SATE.