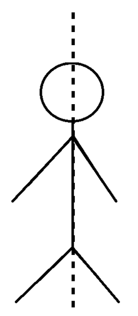
Question: Are conventional standard errors for randomized experiments optimal? What are the best ones?

Treatment Indicator $Z_i \in \{0,1\}$

Potential Outcomes $Y_i(0) \quad Y_i(1) \in \mathbb{R}$

Design-based Inference



Covariates $X_i \in \mathbb{R}^k$

Design-based Inference: only consider randomness due to the randomization

$$\boldsymbol{Z} = \{Z_1, \dots, Z_N\}$$

Option 1: Bernoulli Design

$$Z_i \overset{iid.}{\sim} \text{Bernoulli}(p)$$

Option 2: Simple Random

$$P(\boldsymbol{Z} = z) = \frac{1}{\binom{N}{n_1}} \quad \text{with } \sum_{i=1}^N z_i = n_1$$

Estimand: Sample ATE (SATE)

$$\tau = \frac{1}{N}\sum_{i=1}^N Y_i(1) - Y_i(0) = \bar{Y}(1) - \bar{Y}(0)$$

Regression functions $f_1 : \mathbb{R}^k \to \mathbb{R}$ and $f_0 : \mathbb{R}^k \to \mathbb{R}$

Regression Adjustment

$$f_q \in \arg\min_{f_q \in \mathcal{F}} \left\{ \sum_{i=1}^N (Y_i(q) - f_q(X_i))^2 \right\}$$

Population-adjusted potential outcomes $\varepsilon_i(q) := Y_i(q) - f_q(X_i) \quad$ for $q \in \{0,1\}$

Oracle estimator

$$\hat{\tau}_N^{\text{oracle}} = \frac{1}{n_T}\sum_{i=1}^N T_i(Y_i(1) - f_1(X_i)) - \frac{1}{n_{\bar{T}}}\sum_{i=1}^N \bar{T}_i(Y_i(0) - f_0(X_i)) + \frac{1}{N}\sum_{i=1}^N (f_1(X_i) - f_0(X_i))$$

Under Assumption 1:

$$\text{Var}(\hat{\tau}_N^{oracle}) = \frac{1}{N}\left(\frac{N-n_T}{n_T}s_N^2(1) + \frac{N-n_{\bar{T}}}{n_{\bar{T}}}s_N^2(0) + 2\frac{1}{N}\sum_{i=1}^N \varepsilon_{i,N}(1)\varepsilon_{i,N}(0)\right)$$

Need to use estimates of f_1, f_0

$$s_N^2(q) = \frac{1}{N}\sum_{i=1}^N \varepsilon_{i,N}(q)^2$$

Sample-adjusted outcomes $\hat{\varepsilon}_{i,N}(q) = Y_i(q) - \hat{f}_{q,N}(X_i)$

$$\hat{\tau}_N = \frac{1}{n_T}\sum_{i=1}^N T_i\hat{\varepsilon}_{i,N}(1) - \frac{1}{n_{\bar{T}}}\sum_{i=1}^N \bar{T}_i\hat{\varepsilon}_{i,N}(0) + \frac{1}{N}\sum_{i=1}^N (\hat{f}_{1,N}(X_i) - \hat{f}_{0,N}(X_i))$$

In the settings we consider, variation introduced by estimating f_1, f_0 is asymptotically negligible

Even if f_1, f_0 were known, the variance of the oracle estimator remains non-identified

Variance Bounds

$$\text{Var}(\hat{\tau}_N^{oracle}) = \frac{1}{N}\left(\frac{N-n_T}{n_T}s_N^2(1) + \frac{N-n_{\bar{T}}}{n_{\bar{T}}}s_N^2(0) + 2\frac{1}{N}\boxed{\sum_{i=1}^N \varepsilon_{i,N}(1)\varepsilon_{i,N}(0)}\right)$$

non-identified

Cauchy–Schwarz inequality

$$\text{Var}(\hat{\tau}_N)^{\text{CS}} = \frac{1}{N}\left(\frac{N-n_T}{n_T}s_N^2(1) + \frac{N-n_{\bar{T}}}{n_{\bar{T}}}s_N^2(0) + 2s_N(1)s_N(0)\right)$$

AM-GM inequality

$$\text{Var}(\hat{\tau}_N)^{\text{Conv}} = \frac{1}{n_T}s_N^2(1) + \frac{1}{n_{\bar{T}}}s_N^2(0)$$

Joint distribution $\Gamma_N(\xi_0, \xi_1) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\{\varepsilon_{i,N}(0) \le \xi_0, \varepsilon_{i,N}(1) \le \xi_1\}$

Marginals $F_N(\xi) = 1/N\sum_{i=1}^N \mathbb{1}\{\varepsilon_{i,N}(0) \le \xi\} \qquad G_N(\xi) = 1/N\sum_{i=1}^N \mathbb{1}\{\varepsilon_{i,N}(1) \le \xi\}$

Notice, the variance only depends on the joint distribution

$$\text{Var}(\hat{\tau}_N^{oracle}) = \text{Var}_{\Gamma_N}(\hat{\tau}_N^{oracle})$$

Identification: Joint Γ_N is unknown but marginals F_N and G_N are known

Estimation: Marginals F_N and G_N need to be estimated from observed outcomes

Sharp Bounds on the Variance of General Regression Adjustment in Randomized Experiments

[1]**Jonas Mikhaeil**, [2]Don P. Green

(j.mikhaeil@columbia.edu)

[1]Department of Statistics, [2]Department of Political Science, Columbia University

$$\text{SVB}(F_N, G_N) = \sup_{\gamma \in \Pi(F_N, G_N)} \text{Var}_\gamma(\hat{\tau})$$

Sharp Variance Bounds

$$H_N^H(\varepsilon_1, \varepsilon_0) = \min\{G_N(\varepsilon_1), F_N(\varepsilon_0)\}$$

$$\sup_{\gamma \in \Pi(F_N, G_N)} \text{Cov}_\gamma(Y(1), Y(0)) = \mathbb{E}_{H_N^H}[Y(1)Y(0)] - \mathbb{E}_{F_N}[Y(1)]\mathbb{E}_{G_N}[Y(0)]$$

Sharp Variance Bound

$$V_N^H = \frac{1}{N}\left(\frac{N-n_T}{n_T}\mathbb{E}_{G_N}\left[\varepsilon_N(1)^2\right] + \frac{N-n_{\bar{T}}}{n_{\bar{T}}}\mathbb{E}_{F_N}\left[\varepsilon_N(0)^2\right] + 2\mathbb{E}_{H_N^H}\left[\varepsilon_N(1)\varepsilon_N(0)\right]\right)$$

(Marginals of sample-adjusted potential outcomes)

$$\hat{G}_N(\xi) = 1/\tilde{n}_T\sum_{i=1}^N T_i\mathbb{1}\{\hat{\varepsilon}_{i,N}(1) \le \xi\} \quad \hat{F}_N(\xi) = 1/\tilde{n}_{\bar{T}}\sum_{i=1}^N \bar{T}_i\mathbb{1}\{\hat{\varepsilon}_{i,N}(0) \le \xi\}$$

(Plug-in)

$$\tilde{n}_{\bar{T}} = \sum_{i=1}^N \bar{T}_i \qquad \tilde{n}_T = \sum_{i=1}^N T_i$$

Sharp Variance Bound Estimator

$$\hat{V}_N^H = \frac{1}{N}\left(\frac{N-n_T}{n_T}\mathbb{E}_{\hat{G}_N}\left[\hat{\varepsilon}_N(1)^2\right] + \frac{N-n_{\bar{T}}}{n_{\bar{T}}}\mathbb{E}_{\hat{F}_N}\left[\hat{\varepsilon}_N(0)^2\right] + 2\mathbb{E}_{\hat{H}_N^H}\left[\hat{\varepsilon}_N(1)\hat{\varepsilon}_N(0)\right]\right)$$

Assumption 1: *Bernoulli Randomized Experiment or Completely Randomized Experiment*

Assumption 2: *The joint distribution of population-adjusted potential outcomes H_N converges weakly to a distribution H with marginals G and F.*

Proposition 1 *Let Assumption 1 and 2 hold. Further, assume*

(a) The estimates of the outcome models are $o_p(1)$ consistent, that is,

$$\left(\frac{1}{N}\sum_{i=1}^N (f_{q,N}(X_i) - \hat{f}_{q,N}(X_i))^2\right)^{1/2} = o_p(1) \quad \text{for } q \in \{0,1\}.$$

(b) The population-adjusted potential outcomes $\varepsilon_{i,N}(q) = Y_i(q) - f_{q,N}$ are uniformly square-integrable, that is

$$\sup_N \frac{1}{N}\sum_{i=1}^N \varepsilon_{i,N}(q)^2\mathbb{1}[\varepsilon_{i,N}(q)^2 \ge \beta] \to 0$$

as $\beta \to \infty$ for $q \in \{0,1\}$.

Let $\mathcal{H}$ be the collection of all bivariate distributions with marginals G and F, then

$$NV_N^H \to \frac{1-\pi_T}{\pi_T}\mathbb{E}_G[\varepsilon(1)^2] + \frac{1-\pi_{\bar{T}}}{\pi_{\bar{T}}}\mathbb{E}_F[\varepsilon(0)^2] + 2\sup_{h \in \mathcal{H}}\mathbb{E}_h[\varepsilon(1)\varepsilon(0)]$$

$$NV_N^L \to \frac{1-\pi_T}{\pi_T}\mathbb{E}_G[\varepsilon(1)^2] + \frac{1-\pi_{\bar{T}}}{\pi_{\bar{T}}}\mathbb{E}_F[\varepsilon(0)^2] + 2\inf_{h \in \mathcal{H}}\mathbb{E}_h[\varepsilon(1)\varepsilon(0)],$$

and $(\hat{V}_N^L - V_N^L, \hat{V}_N^H - V_N^H) = o_p(1/N)$.

Corollary Linear Regression

Oracle Model $(\beta_{q,N}, \gamma_{q,N}) = \underset{\beta_q, \gamma_q}{\text{argmin}} \sum_{i=1}^N \left(Y_i(q) - \gamma_q - \beta_q^\top X_i\right)^2 \quad$ for $q = 0, 1$

Estimator $\hat{\tau}_N(\hat{\beta}_1, \hat{\beta}_0) = \frac{1}{n_1}\sum_{i=1}^N Z_i(Y_i(1) - \hat{\beta}_1^\top X_i) - \frac{1}{n_0}\sum_{i=1}^n (1 - Z_i)(Y_i(0) - \hat{\beta}_0^\top X_i)$

[Lin, 2013, Ann. Appl. Stat.]

Under classical regularity assumptions:

$$\hat{\tau}_N(\hat{\beta}_1, \hat{\beta}_0) \pm z_{1-\alpha/2}\sqrt{\hat{V}_N^H}$$

is the **asymptotically narrowest Wald-type confidence interval** with nominal coverage

Proof

The proof has three steps:

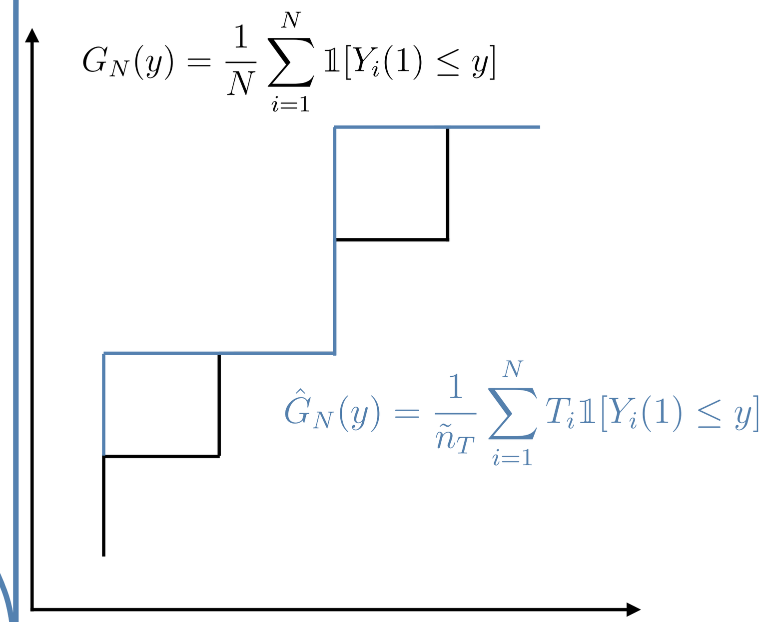
i) Convergence of the marginals $\hat{G}_N, \hat{F}_N$

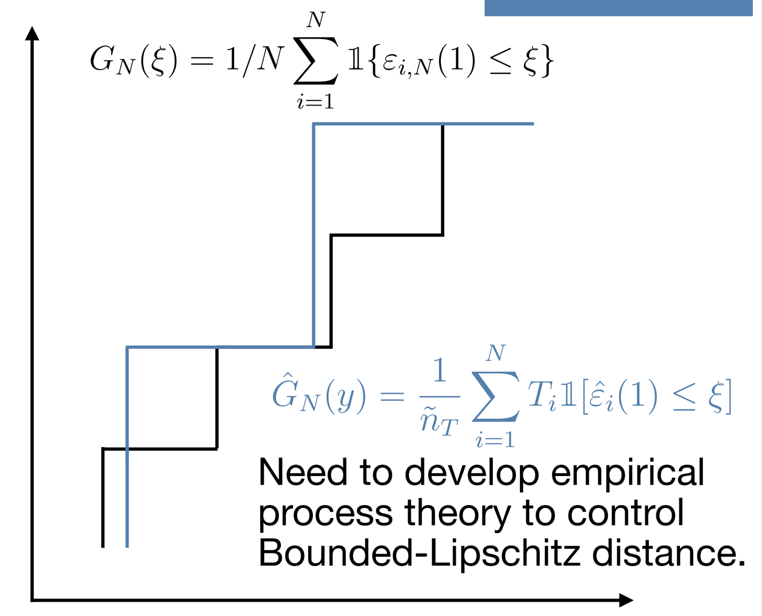
ii) Integration to the limit for $\mathbb{E}_{\hat{G}_N}[\hat{\varepsilon}(1)]$

iii) Convergence of extremal joint $\hat{H}_N^H(\varepsilon_1, \varepsilon_0) = \min\{\hat{G}_N(\varepsilon_1), \hat{F}_N(\varepsilon_0)\}$

Step i) is challenging:

Empirical Process Theory





Need to develop empirical process theory to control Bounded-Lipschitz distance.

Bounded-Lipschitz distance

$$d_{\text{BL}}(\mu, \nu) = \sup\left\{\int \phi d\mu - \int \phi d\nu \,;\, \|\phi\|_\infty + \|\phi\|_{Lip} \le 1\right\}$$ metrizes weak convergence

Empirical Process Theory for Design-based Inference

Proposition 1 (P-Glivenko-Cantelli by entropy) *Let $\mathcal{F}$ be a class of measurable functions with envelope F such that $P_N(\mathcal{F}) \le \infty$. Let $\mathcal{F}_M$ be the class of functions $f\mathbb{1}[F \le M]$ for all $f \in \mathcal{F}$. Then under Assumption 1*

$$\|\hat{P}_N - P_N\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n_T}\sum_{i=1}^N T_i f(y_i) - \frac{1}{N}\sum_{i=1}^N f(y_i) \to 0 \quad \textit{in probability}$$

if there exists an $M > 0$ such that

$$\frac{1}{N}\log N(\varepsilon, \mathcal{F}_M, L_1(P_N)) \xrightarrow{P} 0$$

for every $\varepsilon > 0$.

Lemma 1 *Let $BL := \{f : \mathbb{R} \to [-1,1] | f$ is 1-Lipschitz$\}$ be the class of all 1-bounded-Lipschitz functions. Let P_N be such that $\mathbb{E}_{P_N}[|Y|] = \frac{1}{N}\sum_{i=1}^N |y_i| \le C$ for some constant C. Then*

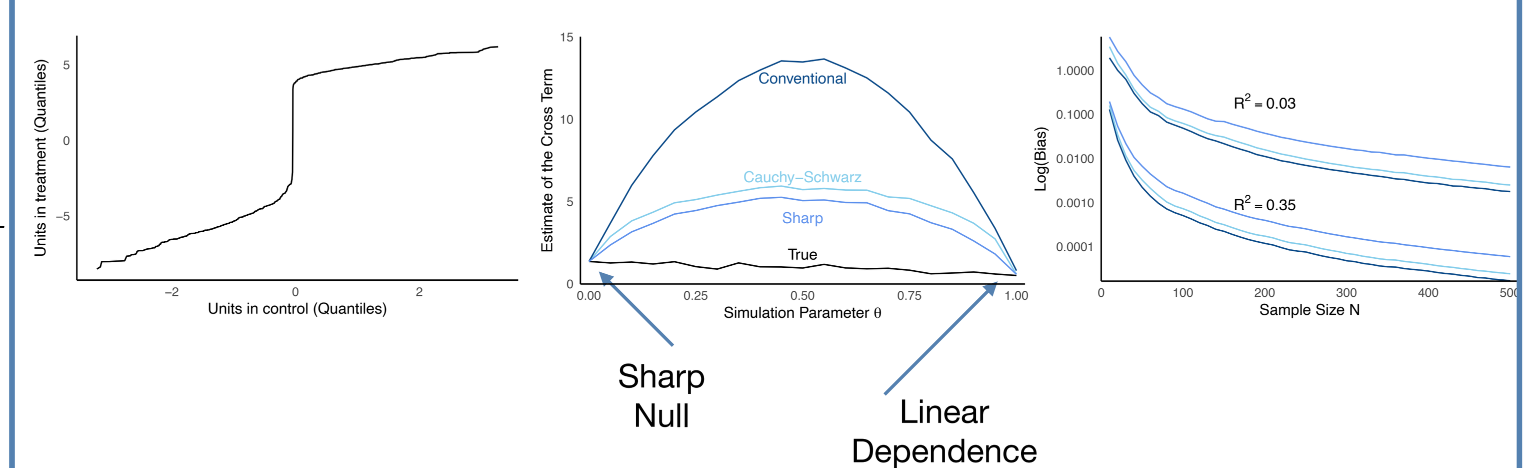
$$\log N(\varepsilon, BL, L_1(P_N)) \le A\frac{C}{\varepsilon}$$

for some constant A and for all $\varepsilon > 0$.

Let $p_i \overset{i.i.d}{\sim} \text{Bernoulli}(\theta)$, $e_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, and $Y_i(0) = \alpha_0 + \beta_0 x_i + e_i$. We then define

$$Y_i(1) = \begin{cases} Y_i(0) & \text{if } p_i = 0 \\ 10 + 0.5e_i & \text{else.} \end{cases}$$

Simulation



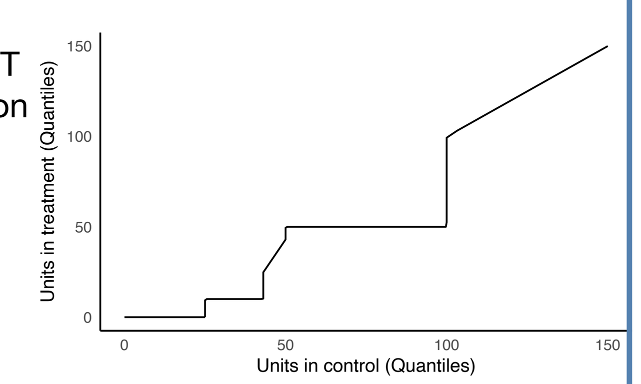
Empirical Example

Re-analyze the randomized experiment Aronow, et al. [2014] analyzed

1,561 subjects called with a fundraising appeal

Treatment (781): Caller who identified themselves as LGBT
Control (780): No mention of the caller's LGBT identification

Dataset features a set of covariates (age, sex, political affiliation) that is jointly significant.



	Unadjusted		Adjusted	
	Variance Estimate	Ratio	Variance Estimate	Ratio
Conventional	0.199	0.938	0.197	0.940
Cauchy-Schwarz	0.195	0.954	0.194	0.956
Sharp	0.186	-	0.185	-

Summary

Variances in randomized experiments are non-identified
Conventional variance bounds are loose

Randomized experiments often crucially rely on covariate adjustment
We derive consistent estimates of sharp variance bounds for general regression adjustment

We provide the asymptotically narrowest Wald-type confidence intervals for general regression adjustment

Open Directions

More than 2 levels of treatment (e.g. Factorial Design)

More complicated designs

https://arxiv.org/abs/2411.00191