# Design-based Inference in Observational Studies with Estimated Propensity Scores

Panos Toulis*

University of Chicago, Booth School of Business

24th Conference on Research on Economic Theory and Econometrics, Ios, Greece, 2026

* Joint work with Shunzhuang Huang, Jiangchuan Du and Azeem Shaikh

# Introduction

- In experimental studies, we may use the randomization of treatment as the logical basis for inference. Known as randomization inference.
- Originally proposed as an alternative to $t$-test by Fisher (1935).
- Key benefit is exact inference in finite samples for "sharp nulls".
- In this paper, we ask:

  **Q1:** How does this idea extend to observational studies?

  **Q2:** Are there any benefits over normal asymptotic inference?

  **A1:** Naive studentization can fail. Influence function correction is crucial.

  **A2:** Yes! Especially in "near-sharp null" regimes.

# Setup

Cross-sectional data with binary treatment.

There are $n$ units indexed by $i \in [n] = \{1, 2, \ldots, n\}$. For each $i$,

- $D_i \in \{0, 1\}$ binary treatment;
- $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ covariate;
- $Y_i(d) \in \mathbb{R}$ is the potential outcome for $d \in \{0, 1\}$;
- $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$ is the observed outcome.

Define the *propensity score* function as

$$\pi(x) = \mathbb{P}(D_i = 1 \mid X_i = x), \quad \pi_i := \pi(X_i) \text{ for short .}$$

# Main assumption

**Assumption.** $\{Y_i(0), Y_i(1), D_i, X_i\}_{i=1}^n \overset{iid}{\sim} P$ for some distribution $P$. Moreover:

1. Ignorability. $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$.
2. Overlap. $\eta \leqslant \pi(x) \leqslant 1 - \eta$, for some $\eta \in (0, 1/2)$ and all $x \in \mathcal{X}$.

- We may interpret these assumptions as defining an *artificial* randomized experiment with distribution:

$$\mathbb{P}(D \mid X) = \prod_{i=1}^n \pi(X_i)^{D_i} (1 - \pi(X_i))^{1-D_i}.$$

- This forms the basis for design-based inference; i.e., only treatment assignment is random, while covariates and potential outcomes are fixed.

# Target of inference

Our main focus is the weak null hypothesis:

$$H_0: \quad \mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(0)].$$

- This is equivalent to testing that the average treatment effect (ATE) is zero; e.g., $H_0 : \beta = 0$, in the linear regression model $Y_i = \beta D_i + \gamma' X_i + \epsilon_i$.
- We will consider a variety of estimators: OLS, doubly robust/double ML.
- To set the stage, we begin with the more common sharp null hypothesis.

# Sharp null hypothesis

A sharper version of the weak null is:

$$H_0^{\text{sharp}}: \quad Y_i(1) = Y_i(0), \ \forall i.$$

- Under mild assumptions, this maps to conditional independence testing:

$$H_0^{\text{CI}}: Y_i \perp\!\!\!\perp D_i \mid X_i.$$

- Testing the sharp null under known propensity scores $\pi_i := P(D_i \mid X_i)$ is possible through the classical Fisher Randomization Test (FRT).
- The conditional permutation test of Candès et al. (2018) for $H_0^{\text{CI}}$ is identical to FRT in this case.

# Fisher Randomization Test — known propensity scores

- Consider a test statistic $t(D, Y, X)$.
- Calculate the randomization distribution

$$\widetilde{F}_t(s) := \mathbb{P}\{t(\widetilde{D}, Y, X) \leqslant s \mid Y, X\},$$

where $\widetilde{D} \sim \otimes_{i=1}^{n} \mathrm{Ber}(\pi_i)$ is an independent copy of $D$.

**Definition.** The randomization test at level $\alpha \in (0, 1)$ is

$$\phi_\alpha(D, Y, X; \pi) = \mathbb{1}\left\{t(D, Y, X) > \widetilde{Q}_{1-\alpha,t}\right\}$$

where $\widetilde{Q}_{1-\alpha,t} = \inf_{s\in\mathbb{R}}\{\widetilde{F}_t(s) \geqslant 1 - \alpha\}$.

- When $\pi_i$s are known (e.g., in experimental studies), the FRT is **exact in finite samples** (Lehmann and Romano, 2005, Ch. 15).
- Fully nonparametric: does not require correct specification of $t(D, Y, X)$.

# Sharp null under estimated propensity scores

- When the $\pi_i$s are unknown, Rosenbaum (1984) proposed to condition on a sufficient statistic of $P(D \mid X)$, then perform a stratified permutation test.

  — Requires model for $P(D \mid X)$. Computationally challenging as well.

- When estimates $\widehat{\pi}_i$s are available, then we can them plug-in to form the randomization test $\phi_\alpha(D, Y, X; \widehat{\pi})$.

- The excess type I error rate of the generic test is bounded by the TV distance between $\{\hat{\pi}_i\}$ and $\{\pi_i\}$ (Berrett et al., 2019).

  — This bound is very loose, in general.

- Under more structure, the test $\phi_\alpha(D, Y, X; \widehat{\pi})$ can be asymptotically valid for the sharp null if $t(D, Y, X)$ is doubly robust (Niu et al., 2024).

We extend these results to the weak null case.

# Weak null under IPW estimator

- We center our analysis of the weak null around the IPW estimator:

$$t^{\text{ipw}}(D, Y, X; \pi) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i D_i}{\pi_i} - \frac{Y_i (1 - D_i)}{1 - \pi_i} \right).$$

- This unbiasedly estimates the ATE and is suitable for testing the weak null:

$$t^{\text{ipw}}(D, Y, X; \pi) \xrightarrow{\text{p}} \mathbb{E}[Y(1) - Y(0)].$$

- Not efficient but can be "augmented" with doubly robust methods.
- Denote by $\phi_\alpha^{\text{ipw}}(D, Y, X; \pi)$ the test based on $t^{\text{ipw}}$.

- Under the weak null and mild regularity conditions, we show that

$$\sqrt{n}\, t^{\mathrm{ipw}}(D, Y, X; \pi) \xrightarrow{\mathsf{d}} N(0, V_{\mathrm{samp}}).$$

- However, under the randomization distribution,

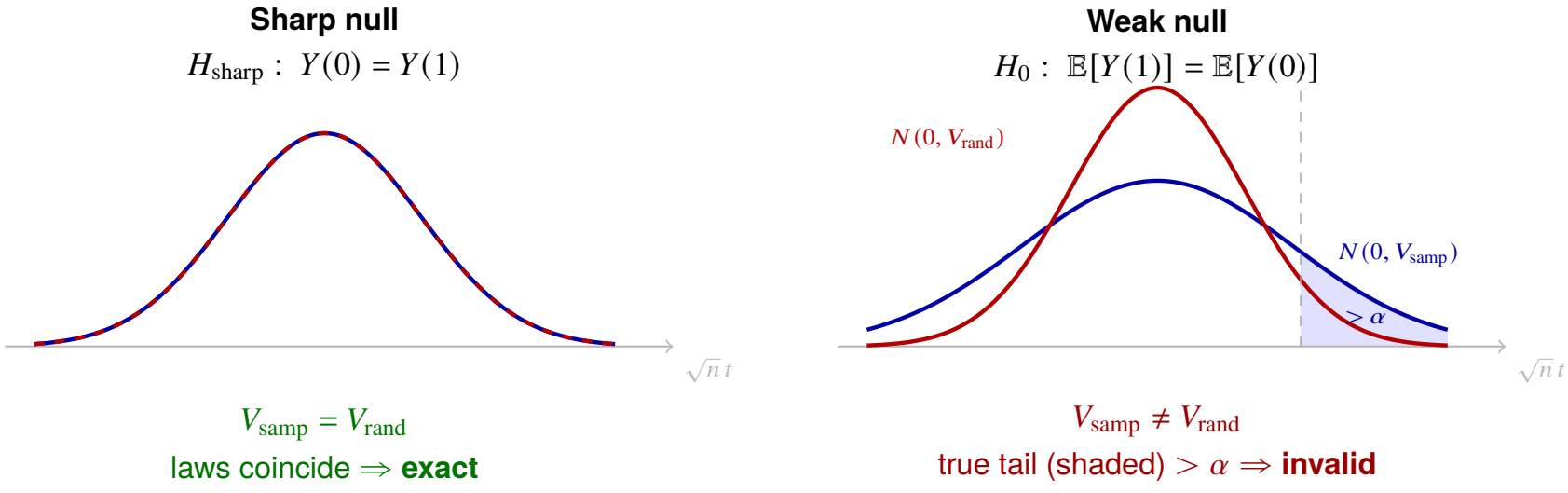
$$\sqrt{n}\, t^{\mathrm{ipw}}(\widetilde{D}, Y, X; \pi) \xrightarrow{\mathsf{d}} N(0, V_{\mathrm{rand}}), \quad V_{\mathrm{rand}} \neq V_{\mathrm{samp}}.$$

**Theorem 2.** Under the weak null, for any $\alpha \in (0, 1)$, we have

$$\lim_{n \to \infty} \mathbb{E}[\phi_\alpha^{\mathrm{ipw}}(D, Y, X; \pi)] = 1 - \Phi(\rho z_{1-\alpha}) \ , \ \rho^2 = V_{\mathrm{rand}}/V_{\mathrm{samp}}.$$

- The IPW-based randomization test is **not valid** in general.

# Illustration: Why simple IPW fails under the weak null




- No mismatch if, in addition, $\pi_i = c$ or $\mathbb{V}\text{ar}[Y(0)] = \mathbb{V}\text{ar}[Y(1)]$.
- This is directly analogous to the seminal work by Romano (1990).

## Studentization

- The asymptotic variance is estimable, however:

$$\widehat{V}(D, Y, X; \pi) = n^{-1} \sum_{i=1}^{n} [Y_i^2 D_i / \pi_i^2 + Y_i^2 (1 - D_i)/(1 - \pi_i)^2]$$
$$\xrightarrow{\mathsf{p}} V_{\text{samp}}.$$

- Moreover, this estimator under the randomization distribution yields:

$$\widehat{V}(\widetilde{D}, Y, X; \pi) \xrightarrow{\mathsf{p}} V_{\text{rand}}.$$

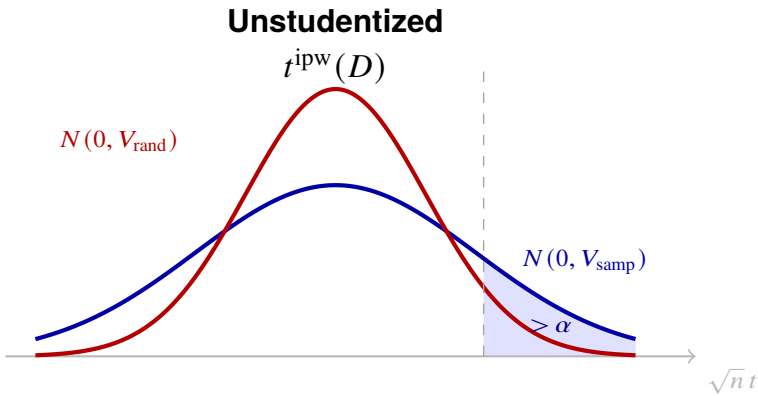
- This suggests using the studentized statistic

$$t^{\text{sipw}}(D) := \frac{t^{\text{ipw}}(D, Y, X; \pi)}{\widehat{V}(D, Y, X; \pi)^{1/2}}.$$

**Theorem 3.** Under the weak null, let $\phi_\alpha^{\text{sipw}}(D, Y, X; \pi)$ denote the randomization test based on $t^{\text{sipw}}(D)$. Then, for any $\alpha \in (0, 1)$,

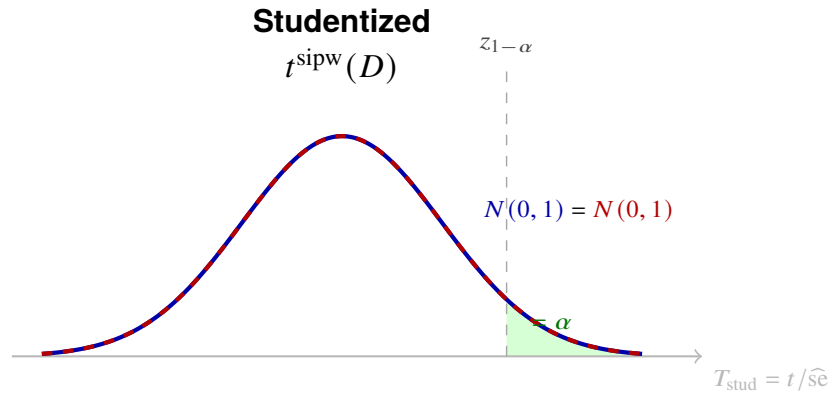
$$\lim_{n \to \infty} \mathbb{E}[\phi_\alpha^{\text{sipw}}(D, Y, X; \pi)] = 1 - \Phi(z_{1-\alpha}) = \alpha \ .$$

# Illustration




$V_{\text{samp}} \neq V_{\text{rand}}$

reference $\neq$ truth $\Rightarrow$ **invalid**




both laws $\longrightarrow N(0, 1)$

reference = truth $\Rightarrow$ **valid**

- Studentization creates a pivotal quantity and restores validity.
- Known technique from earlier work (Romano, 1990; Chung and Romano, 2013; Wu and Ding, 2021). However, only applicable in randomized studies (known $\pi_i$) and non-IPW test statistics.
- It's a challenge to show that $t^{\text{sipw}}(\widetilde{D}) \xrightarrow{\text{d}} N(0, 1)$ under $\widetilde{D}$.
  In fact, we will show next a case where naive studentization fails.

# Weak null — estimated propensity scores

- Let $\pi(x) = \pi(x;\beta)$ for some $\beta \in \mathbb{R}^k$ and known form; e.g., logistic.
- Write $\widehat{\pi}_i := \pi(X_i;\widehat{\beta}_n)$, for $\widehat{\beta}_n$ the MLE, and plug-in to IPW:

$$t^{\text{ipw}}(D,Y,X;\widehat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i D_i}{\widehat{\pi}_i} - \frac{Y_i(1-D_i)}{1-\widehat{\pi}_i}\right).$$

- From standard M-estimation theory

$$\sqrt{n}\, t^{\text{ipw}}(D,Y,X;\widehat{\beta}_n) \xrightarrow{\text{d}} N(0,V_{\text{ipw}}), \text{ where } V_{\text{ipw}} = V_{\text{samp}} - H_\beta I_\beta^{-1} H_\beta \,,$$

  for some terms $H_\beta, I_\beta$ relating to Fisher information.

- *Side note:* Interestingly, IPW is more efficient when using the estimated $\widehat{\pi}_i$s rather than the true $\pi_i$s (Hirano et al., 2003).

# Asymptotic variance

- Estimating the IPW variance is well known (Lunceford and Davidian, 2004):

$$\widehat{V}^{\mathrm{ps}}(D, Y, X; \widehat{\pi}) = n^{-1} \sum_{i=1}^{n} \left( \frac{Y_i D_i}{\widehat{\pi}_i} - \frac{Y_i (1 - D_i)}{1 - \widehat{\pi}_i} - \widehat{v}_n' S_i \right)^2 ,$$

for a particular $\hat{v}_n'$ and scores $S_i = S(D_i, X_i; \widehat{\beta}_n)$.

- We have $\widehat{V}^{\mathrm{ps}}(D) \xrightarrow{\mathsf{p}} V_{\mathrm{ipw}}$. However,

$$\widehat{V}^{\mathrm{ps}}(\widetilde{D}) \xrightarrow{\mathsf{p}} V_{\mathrm{samp}} - H_\beta' I_\beta^{-1} (H_\beta - 2\widetilde{H}_\beta) \neq V_{\mathrm{rand}} \neq \mathbb{V}\mathrm{ar}[nt^{\mathrm{ipw}}(\widetilde{D}, Y, X; \widehat{\beta}_n)].$$

- As a result, the randomization test **is invalid** under this naive studentization.
- The reason for this discrepancy is that $\widehat{\beta}_n$ depend on $D$ but in our procedure they are treated as fixed.

# Influence function correction

Consider a Fisher-scored test statistic:

$$t^{\text{IFipw}}(d, Y, X) = t^{\text{ipw}}(d, Y, X) - \widehat{v}_n' n^{-1} \sum_{i=1}^{n} S(d_i, X_i; \widehat{\beta}_n)$$

$$:= n^{-1} \sum_{i=1}^{n} \varphi_i(d, Y, X).$$

- Since $\widehat{\beta}_n$ is MLE, $t^{\text{IFipw}} = t^{\text{ipw}}$ under $D$. Sampling distributions match.
- However, $t^{\text{IFipw}} \neq t^{\text{ipw}}$, under $\widetilde{D}$, and the scoring term corrects for the randomization error.

**Theorem 4.** Consider the randomization test, $\phi_\alpha^{\text{IFipw}}(D, Y, X)$, based on $t^{\text{IFipw}}$ studentized by $\widehat{V}^{\text{ps}}(\cdot)$. Then, under the weak null,

$$\lim_{n \to \infty} \mathbb{E}[\phi_\alpha^{\text{IFipw}}(D, Y, X)] = 1 - \Phi(z_{1-\alpha}) = \alpha \ .$$

# DGP2: parametric propensity, influence function corrections

$$X_{i1} \sim \text{Unif}(-1,3), \quad X_{ij} \sim \begin{cases} \text{Gamma}(4,2) & j = 2,3,4,5 \\ \text{logNorm}(-1/2,1) & j = 6,7,8,9 \\ \text{Unif}(-5,5) & j = 10,11,12,13 \end{cases}, X_{ij} \perp\!\!\!\perp X_{ij'} \quad \text{(DGP2-}X\text{)}$$

$$\beta_1^D = 0, \quad \beta_{2,3,4,5}^D = \beta_{6,7,8,9}^D = \beta_{10,11,12,13}^D = (-1,-1/3,1/3,1)/\sqrt{p} \quad \text{(DGP2-}\beta^D\text{)}$$

$$\pi(X_i;\beta^D) = \text{logit}(X_i'\beta^D), \quad D_i \sim \text{Ber}(\pi(X_i,\beta^D)) \quad \text{(DGP2-}D\text{)}$$

$$\beta_1^Y = \tau, \quad \beta_{2,3,4,5}^Y = \beta_{6,7,8,9}^Y = \beta_{10,11,12,13}^Y = (-2,4,6,8)/\sqrt{p} \quad \text{(DGP2-}\beta^Y\text{)}$$

$$Y_i(0) \sim N(X_i'\beta_{\tau=0}^Y, 1), \quad Y_i(1) \sim N(X_i'\beta^Y, 1)\ . \quad \text{(DGP2-}Y\text{)}$$

- Some $X$ are heavy-tailed.
- CATE is $\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i] = X_{i1}\tau$.
- ATE is $\mathbb{E}[Y_i(1) - Y_i(0)] = \tau$.
- Goal: Compare IPW using either normal or randomization critical values.

# Results

| Tests | $n$ | $\tau = 0$ | $\tau = 0.25$ | $\tau = 0.5$ | $\tau = 0.75$ |
|---|---|---|---|---|---|
| IPW (normal test) | 200 | 7.54 | 10.94 | 14.39 | 18.44 |
| | 500 | 9.07 | 14.28 | 20.63 | 27.66 |
| | 1000 | 8.22 | 14.73 | 23.49 | 33.96 |
| Studentized IPW (rand test) | 200 | 1.65 | 1.63 | 1.45 | 1.37 |
| | 500 | 0.55 | 0.48 | 0.74 | 1.32 |
| | 1000 | 0.14 | 0.37 | 0.87 | 1.82 |
| IF-studentized IPW (rand test) | 200 | 0.92 | 2.25 | 4.11 | 7.14 |
| | 500 | 3.84 | 7.67 | 12.28 | 19.09 |
| | 1000 | 4.58 | 9.52 | 16.72 | 26.80 |

- Normal test tends to **over-reject.** Asymptotic normality fails in small samples for IPW, especially under extreme propensity scores.
- Randomization test with naive studentization is miscalibrated, although still valid (open question).
- Influence function correction leads to best finite-sample performance.

It turns out this can generalize in the nonparametric case as well.

# Doubly robust estimator

- Apart from modeling $\pi(\cdot)$ alone, we can also model outcomes $\mu_d(x) = \mathbb{E}[Y_i \mid D_i = d, X_i = x]$ following (Niu et al., 2024) for the sharp null.
- Define the doubly robust AIPW estimator:

$$
\begin{aligned}
&t^{\mathrm{dr}}(D, Y, X; \mu_d, \pi) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{D_i(Y_i - \mu_1(X_i))}{\pi(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - \pi(X_i)} + \mu_1(X_i) - \mu_0(X_i) \right) \\
&\quad := \frac{1}{n} \sum_{i=1}^{n} \varphi_i(D, Y, X; \mu_d, \pi).
\end{aligned}
$$

- Also define the variance estimator

$$
\widehat{V}^{\mathrm{dr}}(D, Y, X; \widehat{\mu}_d, \widehat{\pi}) := \frac{1}{n} \sum_{i=1}^{n} [\varphi_i(D, Y, X; \widehat{\mu}_d, \widehat{\pi}) - t^{\mathrm{dr}}(D, Y, X; \widehat{\mu}_d, \widehat{\pi})]^2
$$

- We estimate the nuisance functions using cross-fitting (Chernozhukov et al., 2018).

# Main result

- Define the error rates: $E_{k,y_d} := (|I_k|^{-1} \sum_{i \in I_k} (\widehat{\mu}_{d,i} - \mu_{d,i})^2)^{1/2}$ and $E_{k,d} := (|I_k|^{-1} \sum_{i \in I_k} (\widehat{\pi}_i - \pi_i)^2)^{1/2}$.

**Theorem 5.** Suppose that regularity assumptions hold, and the DR/DML estimation rates are typical: $E_{k,y_d} = o_P(1)$, $E_{k,d} = o_P(1)$ and $E_{k,y_d} E_{k,d} = o_P(n^{-1/2})$. Consider the randomization test $\phi_\alpha^{\mathrm{dr}}$ based on $t^{\mathrm{dr}}(D)/\widehat{V}^{\mathrm{dr}}(D)$.

Then, under the weak null, for any $\alpha \in (0, 1)$,

$$\lim_{n \to \infty} \mathbb{E}[\phi_\alpha^{\mathrm{dr}}(D, Y, X; \widehat{\mu}_d, \widehat{\pi})] = \alpha \ .$$

- As usual, the novel part of the proof is the convergence of the randomization distribution quantiles.

# Higher-order properties

- In previous results, we have shown validity of randomization tests.
- However, both standard normal and randomization tests are first-order ($\sqrt{n}$) equivalent.
- So, it's natural to ask:

  *Since we lose finite-sample validity anyway, why use randomization inference in observational studies?*

- Next, we study higher-order properties through Edgeworth expansions (Edgeworth, 1894; Hall, 1992).

# Higher-order properties

- Under the weak null, uniformly in $u \in \mathbb{R}$:[1]

$$\mathbb{P}\left( \frac{\sqrt{n}t^{\mathrm{dr}}(D; \widehat{\mu}_d, \widehat{\pi}))}{\widehat{V}^{\mathrm{dr}}(D; \widehat{\mu}_d, \widehat{\pi})} \leqslant u \right) = \Phi(u) + \gamma \frac{(2u^2+1)\varphi(u)}{6\sqrt{n}} + O(\Delta_n) + O(n^{-1})$$

$$\widetilde{\mathbb{P}}\left( \frac{\sqrt{n}t^{\mathrm{dr}}(\widetilde{D}; \widehat{\mu}_d, \widehat{\pi})}{\widehat{V}^{\mathrm{dr}}(\widetilde{D}; \widehat{\mu}_d, \widehat{\pi})} \leqslant u \right) = \Phi(u) + \widetilde{\gamma} \frac{(2u^2+1)\varphi(u)}{6\sqrt{n}} + O_P(n^{-1}) \ .$$

- Here, $\gamma, \widetilde{\gamma}$ are skewness terms of the influence functions.

- Term $\Delta_n$ is the absolute difference between ideal and estimated DR:

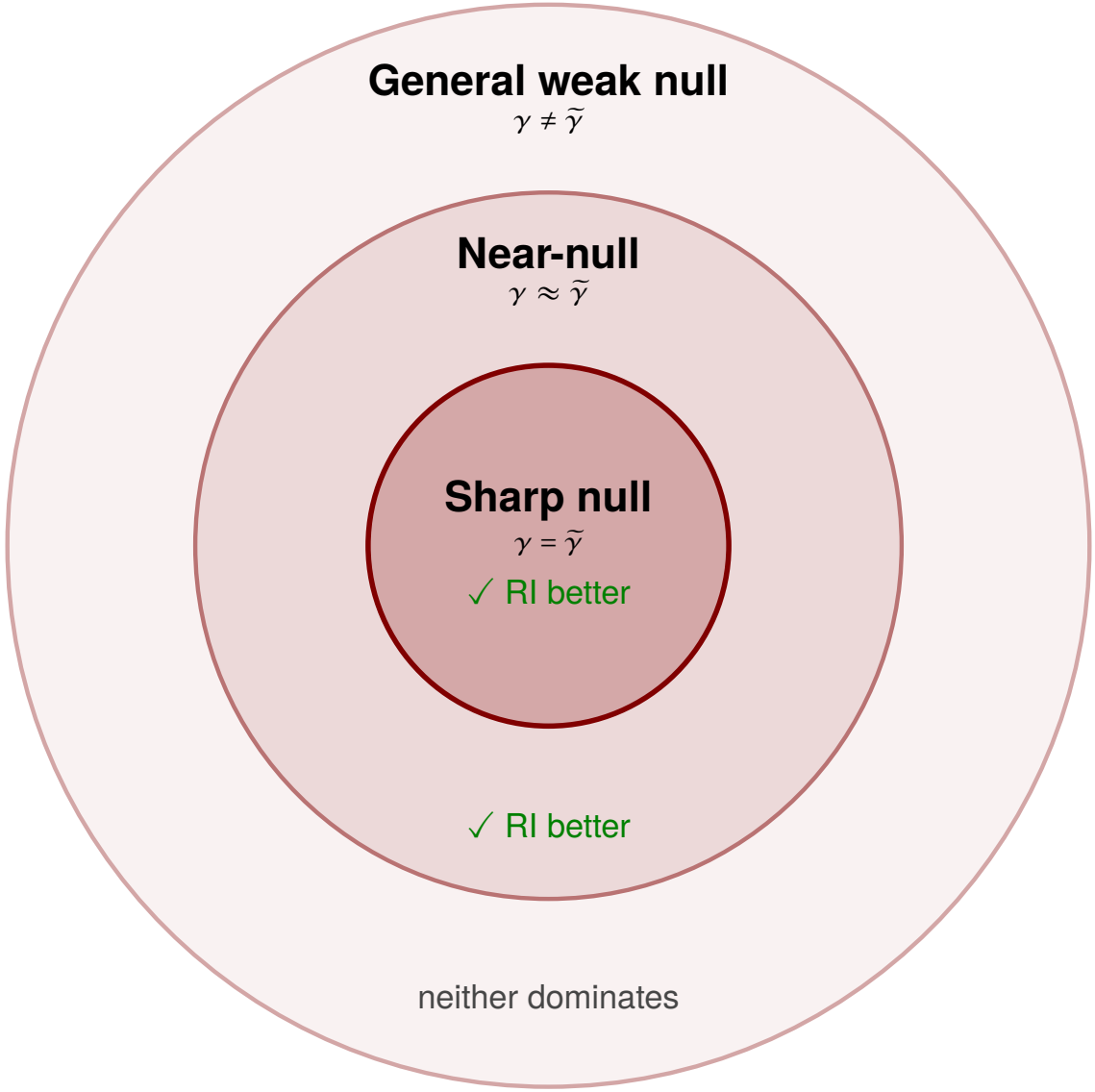
$$\Delta_n := \left| \frac{\sqrt{n}t^{\mathrm{dr}}(D; \widehat{\mu}_d, \widehat{\pi})}{\widehat{V}^{\mathrm{dr}}(D; \widehat{\mu}_d, \widehat{\pi})^{1/2}} - \frac{\sqrt{n}t^{\mathrm{dr}}(D; \mu_d, \pi)}{\widehat{V}^{\mathrm{dr}}(D; \mu_d, \pi)^{1/2}} \right| ,$$

[1] We silence dependence on $D, Y, X$.

# Interpretation

- If $|\gamma - \widetilde{\gamma}| < |\gamma|$, then the randomization test will generally yield a better second-order approximation than standard normal inference.
- Overall, no method dominates the other (we identify Regimes I-VI).
- However, when the weak null holds *and* the effects are very small or sparse, then $\gamma \approx \widetilde{\gamma}$.
- In such "near-sharp null" regimes, randomization inference **should perform better** in finite samples.

# Illustration



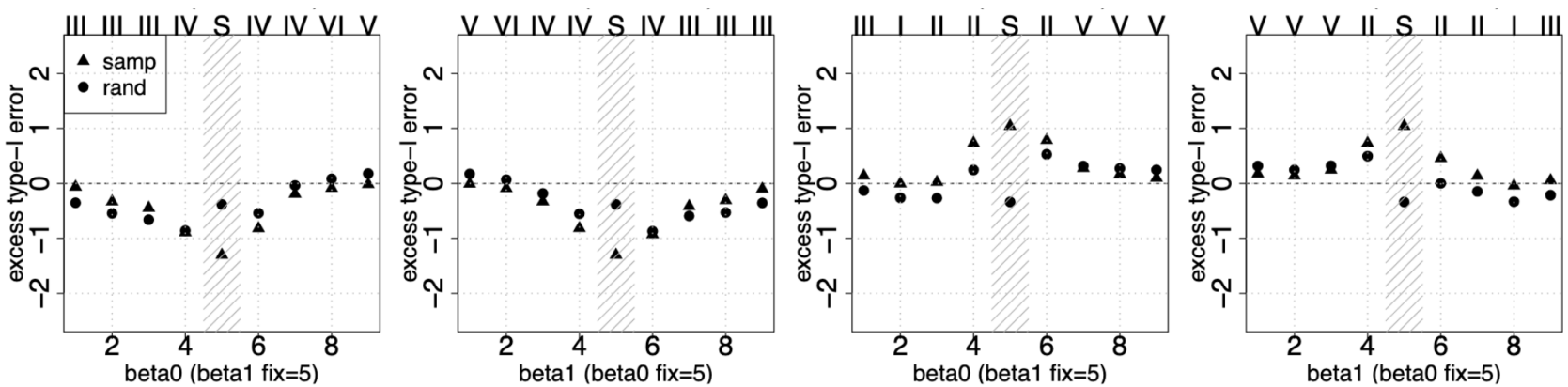

- Randomization inference vs. normal asymptotic inference

# DGP4: higher-order correction

- $\pi_i \sim \text{Beta}(a, b)$.
- $Y_i(1) - Y_i(0) = (\beta_1 - \beta_0)X_i$, where $X_i \sim N(0, 1)$.
- Sample outcome errors $\varepsilon_i$ to control for skewness.
- Use DML estimators (package `DoubleML`) in `R`.
- By construction:
  - Since $\mathbb{E}[Y(1) - Y(0)] = 0$, the weak null always holds.
  - When $\beta_0 = \beta_1$, then the sharp null holds as well.

# Results

**(a,b) = (20,100)** **(a,b) = (100,20)**




- No test dominates the other. Performance depends on particular regime I-VI.
- However, the randomization test **performs better** when $\beta_0 \approx \beta_1$ (shaded areas).
- These are “near-shap null” settings: the ATE is zero but the sharp null does not technically hold. However, individual treatment effects are small.

# Practical takeaways

- Construct the nonparametric DR/DML estimator $t^{\mathrm{dr}}(D)$ for the ATE and its variance estimate $\hat{V}^{\mathrm{dr}}(D)$, as usual.
- For this estimator we can do inference based on $p_{\mathrm{norm}}$, the $p$-value based on the standard normal quantile.
- Or use the randomization $p$-value:

$$p_{\mathrm{rand}} = P\Big(\frac{t^{\mathrm{dr}}(\widetilde{D})}{\hat{V}^{\mathrm{dr}}(\widetilde{D})^{1/2}} > \frac{t^{\mathrm{dr}}(D)}{\hat{V}^{\mathrm{dr}}(D)^{1/2}} \mid X, Y, D\Big).$$

In this paper, we showed that:

1. $p_{\mathrm{rand}}$ is valid asymptotically under roughly the same conditions as $p_{\mathrm{norm}}$.
2. $p_{\mathrm{rand}}$ has better finite-sample performance than $p_{\mathrm{norm}}$ near the sharp null.
3. Otherwise, no method dominates the other in observational studies.
4. In practice, we could use $\max(p_{\mathrm{rand}}, p_{\mathrm{norm}})$ for added robustness.

# References


Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2019). The Conditional Permutation Test for Independence While Controlling for Confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(3):551–577.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484 – 507.

Edgeworth, F. Y. (1894). The asymmetrical probability curve. *Proceedings of the Royal Society of London*, 56(336-339):271–272.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer New York, NY.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Niu, Z., Chakraborty, A., Dukes, O., and Katsevich, E. (2024). Reconciling model-x and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, 52(3):895–921.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Wu, J. and Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536):1898–1913.