Valid and Powerful Hypothesis Testing in Observational Studies

Sensitivity analysis for unmeasured confounding

Elaine K. Chiu (University of Wisconsin–Madison)

Expected duration: 38 minutes

Outline

- Introduction
- Valid Tests for Categorical Treatments
- Design Sensitivity and Power
- Sample Size Under Hidden Bias
- Conclusion

A Note on Notation

s : subject index, $s \in \{1, \ldots, N\}$
i : treatment category, $i \in \{1, \ldots, I\}$

- The **gray box** in the upper-right corner collects the notation.


Introduction
Categorical Tests
Design Sensitivity
Sample Size
Conclusion

Hypothesis Testing: A Well-Rounded Machine in RCTs

We desire to control **two error rates** for testing against H_0 of no effect:

	Reject H_0	Do not reject
H_0 **true**	**Type I** (α)	Correct
H_0 **false**	Correct Power $= 1 - \beta$	**Type II** (β)

In an RCT, the machine is complete

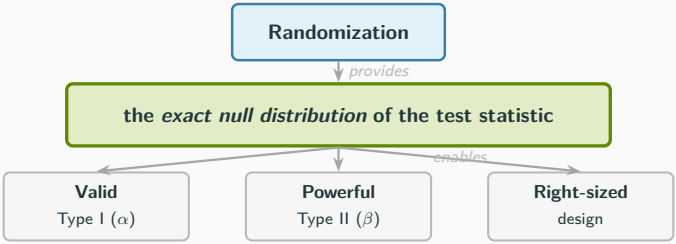
Valid p**-value** — exact null distribution by randomization; Type I controlled (Fisher[1], 1935).

Power & efficiency — asymptotic relative efficiency ranks competing tests (Serfling[2], 1980).

Sample size — closed-form pairs I for target power $1 - \beta$ (Noether[3], 1987; Chow et al.[4], 2017).

[1] Fisher, R. A. (1935). *The Design of Experiments.*
[2] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Ch. 10.
[3] Noether, G. E. (1987). *Sample size determination for some common nonparametric tests.* JASA.
[4] Chow, S.-C., Shao, J., Wang, H. & Lokhnygina, Y. (2017). *Sample Size Calculations in Clinical Research*, 3rd ed.

The Null Distribution of a Test is Available under RCT



Takeaway

No randomization $\Rightarrow$ null distribution unknown $\Rightarrow$ none of the three comes for free.

Observational Data Breaks the Machine

- Without randomization, an **unmeasured confounder** alters the null distribution in a non-estimable way. The RCT machine no longer applies.
- **Sensitivity analysis** rebuilds each piece, accounting for hidden bias.

Component	Previous Literature	Scope
Valid p-value	P-value under hidden bias (Rosenbaum[1]; Rosenbaum & Krieger[2]; Zhang et al.[3])	binary treatment or binary outcome
Power	Design sensitivity $\widetilde{\Gamma}$ (Rosenbaum[4])	binary treatment matched pairs
Sample size	Variance inflation for weighting (Shook-Sa & Hudgens[5]; Liu, Yang & Li[6])	observed, not unmeasured, confounding

[1] Rosenbaum, P. R. (1987). *Sensitivity analysis for permutation inferences in matched studies.* Biometrika.
[2] Rosenbaum, P. R. & Krieger, A. M. (1990). *Sensitivity of two-sample permutation inferences.* JASA.
[3] Zhang, J., Small, D. S. & Heng, S. (2024). *Sensitivity analysis: continuous exposures, binary outcomes.* Biometrika.
[4] Rosenbaum, P. R. (2004). *Design sensitivity in observational studies.* Biometrika.
[5] Shook-Sa, B. E. & Hudgens, M. G. (2022). *Power and sample size for point-exposure observational studies.* Biometrics.
[6] Liu, B., Yang, C. & Li, F. (2026). *Sample size and power calculations for causal inference of observational studies.* Preprint.

Our Papers Advance Hypothesis Testing in Observational Studies via Sensitivity Analysis

We advance hypothesis testing under hidden bias. (★ = covered today)

★ **Paper 1** — valid p-value, categorical $I \times J$ tables — *JASA* (T&M), major revision
"Exact, Nonparametric Sensitivity Analysis for Observational Studies of Contingency Tables"

Paper 2 — power, general & continuous treatments — *JRSS-B*, R&R
"Towards Robust Matched Observational Studies with General Treatment Types"

Paper 3 — new tests & sample size, clustered designs — working paper
"Sensitivity Analysis for Matched-Pair Cluster Designs"

★ **Paper 4** — sample size for hidden bias Γ — working paper
"Sample Size Calculation in Causal Inference for Unmeasured Confounding"

Reminder: Other Sensitivity Analysis Frameworks Exist

Quantity of Interest	**Example Papers**
Regression coefficient	Frank (2000)[1]; Oster (2019)[2]; Cinelli & Hazlett (2020)[3]
ATT / ATE	Imbens (2003)[4]; Huang & Pimentel (2025)[5]
Risk ratio	Ding & VanderWeele (2016)[6]
P-value in hypothesis testing	Rosenbaum & Krieger (1990)[7]

[1] Frank, K. A. (2000). *Impact of a Confounding Variable on a Regression Coefficient.*
[2] Oster, E. (2019). *Unobservable Selection and Coefficient Stability: Theory and Evidence.*
[3] Cinelli, C. & Hazlett, C. (2020). *Making Sense of Sensitivity: Extending Omitted Variable Bias.*
[4] Imbens, G. W. (2003). *Sensitivity to Exogeneity Assumptions in Program Evaluation.*
[5] Huang, M. & Pimentel, S. D. (2025). *Variance-based sensitivity analysis for weighting estimators results in more informative bounds.*
[6] Ding, P. & VanderWeele, T. J. (2016). *Sensitivity Analysis Without Assumptions.*
[7] Rosenbaum, P. R. & Krieger, A. M. (1990). *Sensitivity of Two-Sample Permutation Inferences in Observational Studies.*


Introduction
Categorical Tests
Design Sensitivity
Sample Size
Conclusion

Categorical Data & Classical Tests (1/3)

- Categorical data is common in observational studies.
- It forms an I (treatment) by J (outcome) contingency table.
- Does pre-kindergarten care type affect a child's math proficiency?

Pre-K care	**Math proficiency**		
	Number & shape	Relative size	Ordinality & sequence
No care	12	3	0
Relative	18	12	3
Center-based	17	25	4

3×3 table; girls from low-income, single-parent families. *[P1]*

Categorical Data & Classical Tests (2/3)

s	: subject index, $s \in \{1, \ldots, N\}$
i	: treatment category, $i \in \{1, \ldots, I\}$
j	: outcome category, $j \in \{1, \ldots, J\}$
$r_s^{(i)}$	: potential outcome of s under i

- Does pre-kindergarten care type affect a child's math proficiency?
- Formally, we test Fisher's sharp null of no effect:

Fisher's sharp null

$$H_0 : \; r_s^{(1)} = r_s^{(2)} = \cdots = r_s^{(I)} \quad \text{for every } s,$$

i.e. a child's response is the same regardless of pre-K care type.

Categorical Data & Classical Tests (3/3)

s : subject index, $s \in \{1, \ldots, N\}$
Z_s : treatment received by s
i : treatment category, $i \in \{1, \ldots, I\}$

- Test statistics T: Fisher's exact, χ^2, G^2, ordinal [1].
- Type I error rate control:

  Randomization gives $\mathbb{P}(Z_s{=}i) = \mathbb{P}(Z_{s'}{=}i)$ for every $s \neq s'$

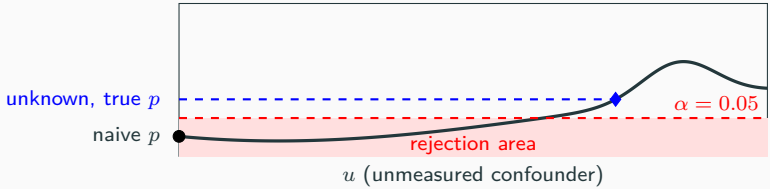
  $\rightarrow$ Permutation[†] yields the (known) dist. of T, thus the p-value $\mathbb{P}(T \geq T_{\text{obs}})$ [2]

  $\rightarrow$ rejecting H_0 when $p \leq \alpha$ controls the Type I error rate below α [3]

[†] Treatment-category totals are fixed by design.
[1] Agresti, A. (2013). *Categorical Data Analysis*, 3rd ed.
[2] Fisher, R. A. (1935). *The Design of Experiments*. (Ch. III, §21.)
[3] Casella & Berger (2002), Thm. 8.3.27.

Unmeasured Confounding Inflates Type I Error Rate

- **Observational:**
  - Unmeasured confounder u_s makes the assignment probabilities $\mathbb{P}(Z_s{=}i)$ differ across subjects *in some unknown way*.
  - Dist. of trt (X) $\rightarrow$ Dist. of T(trt, out) (X) $\rightarrow$ p-value (X)
- **Issue:**
  - The unknown, true p may exceed the naive p (under randomization)
  - if we still reject H_0 when (the underestimated) $p \leq \alpha$
  - Reject *too often* $\rightarrow$ Type I error rate inflates.



The Map of a Sensitivity Analysis

1. **Model the bias.** Assume a model for how an unmeasured confounder could distort the treatment-assignment probabilities.

2. **Bound the p-value.** Find the largest p-value it could take under that model — the upper bound $\bar{p}$.

3. **Decide.** Reject H_0 only when even this largest p-value stays below α: $\bar{p} < \alpha$.

Validity — Casella & Berger (2002), Thm. 8.3.27

$\bar{p}$ is a *valid* p-value: $\mathbb{P}_{H_0}(\bar{p} \leq \alpha) \leq \alpha$. So Step 3 controls the Type I error rate $\leq \alpha$.

Model the bias: Sensitivity Model (Categorical Treatment) (1/2)

Z_s	: treatment received by s
$u_s \in [0,1]^\dagger$	: unmeasured confounder
$\mathbf{x}_s$	: observed covariates of s
$\delta_i \in \{0,1\}$	: is level i favored?
$\Gamma = \exp(\gamma)$	: max. treatment-odds ratio
$\mathcal{F}$	: $\{r_s^{(i)}, \mathbf{x}_s, u_s\}$, held fixed

Sensitivity model (Rosenbaum, 2006[1])

$$\mathbb{P}(Z_s{=}i \mid \mathcal{F}) = \frac{\exp(\kappa_i(\mathbf{x}_s) + \gamma\, \delta_i\, u_s)}{\sum_{k=1}^{I} \exp(\kappa_k(\mathbf{x}_s) + \gamma\, \delta_k\, u_s)}.$$

Two subjects with the same covariates $\mathbf{x}_s = \mathbf{x}_{s'}$: their odds of i vs. i' obey

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(Z_s{=}i \mid \mathcal{F})/\mathbb{P}(Z_s{=}i' \mid \mathcal{F})}{\mathbb{P}(Z_{s'}{=}i \mid \mathcal{F})/\mathbb{P}(Z_{s'}{=}i' \mid \mathcal{F})} \le \Gamma.$$

$\delta_i{=}\delta_{i'}$: ratio $= 1$ (*generic* bias cancels); $\delta_i{\neq}\delta_{i'}$: up to Γ (*differential* bias).

[1] Rosenbaum (2006), *Biometrika* 93(3):573–586.

[†] $u_s \in [0,1]$ fixes the scale, so $\Gamma = e^\gamma$ is the odds-ratio bound.

Model the bias: Sensitivity Model (Categorical Treatment) (2/2)

Z_s	: treatment received by s
$u_s \in [0,1]^\dagger$	: unmeasured confounder
$\mathbf{x}_s$	: observed covariates of s
$\delta_i \in \{0,1\}$	: is level i favored?
$\Gamma = \exp(\gamma)$	: max. treatment-odds ratio
$\mathcal{F}$	: $\{r_s^{(i)}, \mathbf{x}_s, u_s\}$, held fixed

Sensitivity model (Rosenbaum, 2006[1])

$$\mathbb{P}(Z_s{=}i \mid \mathcal{F}) = \frac{\exp(\kappa_i(\mathbf{x}_s) + \gamma\, \delta_i\, u_s)}{\sum_{k=1}^{I} \exp(\kappa_k(\mathbf{x}_s) + \gamma\, \delta_k\, u_s)}.$$

Two subjects with the same covariates $\mathbf{x}_s = \mathbf{x}_{s'}$: their odds of i vs. i' obey

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(Z_s{=}i \mid \mathcal{F})/\mathbb{P}(Z_s{=}i' \mid \mathcal{F})}{\mathbb{P}(Z_{s'}{=}i \mid \mathcal{F})/\mathbb{P}(Z_{s'}{=}i' \mid \mathcal{F})} \le \Gamma.$$

$\delta_i{=}\delta_{i'}$: ratio $= 1$ (*generic* bias cancels); $\delta_i{\neq}\delta_{i'}$: up to Γ (*differential* bias).

[1] Rosenbaum (2006), *Biometrika* 93(3):573–586.

$^\dagger u_s \in [0,1]$ fixes the scale, so $\Gamma = e^\gamma$ is the odds-ratio bound.

Bound the p-value: The Computational Challenge (1/3)

s	: subject index
$\mathcal{Z}(\mathbf{N}_{I\cdot})$	: treatment set
$u_s \in [0,1]$	: unmeasured confounder
T_{obs}	: observed statistic
$\mathbf{u} \in [0,1]^N$	: confounder vector
$\mathbf{r}$	: outcome vector

- The exact null distribution at an $\mathbf{u} = (u_1, \ldots, u_N)$:

$$\alpha(T, \mathbf{r}, \mathbf{u}) = \frac{\sum_{\mathbf{z} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \mathbb{1}\{T(\mathbf{N}(\mathbf{z}, \mathbf{r})) \geq T_{\text{obs}}\} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{z_s = i\} u_s\right)}{\sum_{\mathbf{b} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{b_s = i\} u_s\right)}$$

Bound the p-value: The Computational Challenge (2/3)

s	: subject index
$\mathcal{Z}(\mathbf{N}_{I\cdot})$	: treatment set
$u_s \in [0,1]$	: unmeasured confounder
T_{obs}	: observed statistic
$\mathbf{u} \in [0,1]^N$	: confounder vector
$\mathbf{r}$	: outcome vector

- The exact null distribution at an $\mathbf{u} = (u_1, \ldots, u_N)$:

$$\alpha(T, \mathbf{r}, \mathbf{u}) = \frac{\sum_{\mathbf{z} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \mathbb{1}\{T(\mathbf{N}(\mathbf{z}, \mathbf{r})) \geq T_{\text{obs}}\} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{z_s = i\} u_s\right)}{\sum_{\mathbf{b} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{b_s = i\} u_s\right)}$$

- The reported p-value is the largest over all possible unmeasured confounders:

$$\max_{\mathbf{u} \in [0,1]^N} \alpha(T, \mathbf{r}, \mathbf{u}) \qquad \mathbf{u}^+ = \arg\max_{\mathbf{u} \in [0,1]^N} \alpha(T, \mathbf{r}, \mathbf{u})$$

Bound the p-value: The Computational Challenge (3/3)

s	: subject index
$\mathcal{Z}(\mathbf{N}_{I\cdot})$	: treatment set
$u_s \in [0,1]$	: unmeasured confounder
T_{obs}	: observed statistic
$\mathbf{u} \in [0,1]^N$	: confounder vector
$\mathbf{r}$	: outcome vector

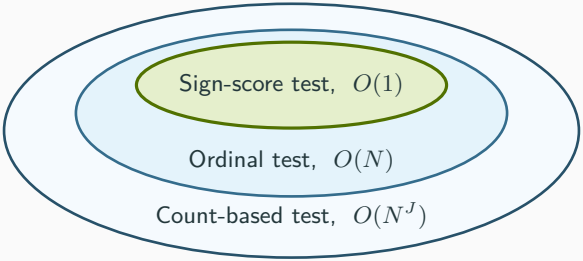
- The exact null distribution at $\mathbf{u}$:

$$\alpha(T, \mathbf{r}, \mathbf{u}) = \frac{\sum_{\mathbf{z} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \mathbb{1}\{T(\mathbf{N}(\mathbf{z}, \mathbf{r})) \geq T_{\text{obs}}\} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{z_s = i\} u_s\right)}{\sum_{\mathbf{b} \in \mathcal{Z}(\mathbf{N}_{I\cdot})} \exp\left(\gamma \sum_{s=1}^{N} \sum_{i=1}^{I} \delta_i \mathbb{1}\{b_s = i\} u_s\right)}$$

- Finding $\mathbf{u}^+ = \arg\max_{\mathbf{u} \in [0,1]^N} \alpha(T, \mathbf{r}, \mathbf{u})$ is difficult:
  - $\alpha(T, \mathbf{r}, \mathbf{u})$ is non-concave in $\mathbf{u}$.
  - The search space $\mathbf{u} \in [0,1]^N$ is large.

Our Contribution – Reduction of the Candidate Set of $\mathbf{u}^+$

The largest p-value is attained at a *vertex* of $[0,1]^N$, i.e., $\mathbf{u}^+ \in \{0,1\}^N$, so the search for $\mathbf{u}^+$ is finite — and the size of that candidate set depends on the test statistic:



Finding the upper bound on the p-value → Type I error rate control becomes computationally feasible.

Count-Based Tests and Their Candidate Set $O(N^J)$

Definition 1 — count-based test

A test statistic $T(\mathbf{N})$ is *count-based* if it is a function of the $I \times J$ table $\mathbf{N} = (N_{ij})$, where $N_{ij} = \#\{s : Z_s = i,\, r_s = j\}$.

	j=1	j=2	j=3	$N_{i\cdot}$
i=1	3	2	1	6
i=2	0	2	4	6
i=3	0	1	5	6
$N_{\cdot j}$	3	5	10	18

Theorem 1 — candidate-set cardinality

The maximum search for $\mathbf{u}^+$ is finite, with at most N^J candidates.

Ordinal Tests and Their Candidate Set $N+1$

Definition 2 — ordinal test

A test is *ordinal* if, for monotone scores $w_1 \le \cdots \le w_I$ on treatments and $v_1 \le \cdots \le v_J$ on outcomes, $T(\mathbf{N}) = \sum_{i=1}^{I} \sum_{j=1}^{J} w_i\, v_j\, N_{ij}$, detecting *monotone trend*.

	$j{=}1$	$j{=}2$	$j{=}3$	$N_{i\cdot}$
$i{=}1$	3	2	1	6
$i{=}2$	0	2	4	6
$i{=}3$	0	1	5	6
$N_{\cdot j}$	3	5	10	18

Scores $\mathbf{w} = (0, 1, 2.5)$, $\mathbf{v} = (0, 1, 2)$:

$$\begin{aligned} T &= \textstyle\sum_{i,j} w_i\, v_j\, N_{ij} \\ &= 1{\cdot}1{\cdot}2 + 1{\cdot}2{\cdot}4 + 2.5{\cdot}1{\cdot}1 + 2.5{\cdot}2{\cdot}5 \\ &= 2 + 8 + 2.5 + 25 = 37.5. \end{aligned}$$

Theorem 2 — candidate set

Maximum search for $\mathbf{u}^{+}$: at most $N+1$ candidates.
Binary outcome ($J{=}2$) $\Rightarrow$ *sign-score* test $\Rightarrow$ one closed-form vertex: 1 candidate.


Introduction
Categorical Tests
Design Sensitivity
Sample Size
Conclusion

- **Type I (α)**
  - Validity comes from maximizing the p-value over $\mathbf{u} \in [0,1]^N$ — solved exactly for categorical treatments[1].
- **Type II (β)**
  - Design sensitivity $\widetilde{\Gamma}$ predicts a test's power before seeing the data — established for binary treatments[3]; extended in our work to continuous[2] and clustered[4] treatments.
  - This presentation gives the design sensitivity for binary treatment continuous outcome for intuition.

[1] Chiu, E. K. & Kang, H. *"Exact, Nonparametric Sensitivity Analysis for Observational Studies of Contingency Tables."*
[2] Heng, S.*, Chiu, E. K.*, & Kang, H. *"Towards Robust Matched Observational Studies with General Treatment Types: Consistency, Efficiency, and Adaptivity."*
[3] Rosenbaum, P. R. (2004). *"Design Sensitivity in Observational Studies."* Biometrika, 91(1), 153–164.
[4] Chiu, E. K. & Kang, H. *"Sensitivity Analysis for Matched-Pair Cluster Designs."*

The Matched-Pair Design

Within each pair, one unit is treated and one control, matched on observed covariates ($x_{i1}{=}x_{i2}$) but possibly differing in u_{ij}.

I	: number of matched pairs
$i,\ j$	: pair $i \in \{1,\dots,I\}$, unit $j \in \{1,2\}$
Z_{ij}	: treatment: 1 treated, 0 control
Y_{ij}	: observed outcome
u_{ij}	: unmeasured confounder $\in [0,1]$
D_i	: $(Z_{i1}{-}Z_{i2})(Y_{i1}{-}Y_{i2})$
$\mathcal{F}$	: $\{Y_{ij}(0), Y_{ij}(1), x_{ij}\}$, held fixed
$\mathcal{Z}$	: $\{\mathbf{z} : z_{i1}{+}z_{i2}{=}1\ \forall i\}$

Sensitivity model (Rosenbaum, 2002[1])

Matched on x, two units' odds of treatment differ by at most a factor of $\Gamma = e^{\gamma} \geq 1$:

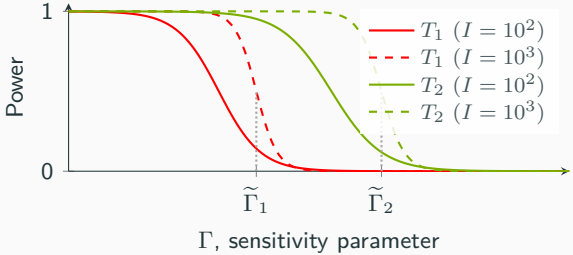
$$\frac{1}{\Gamma} \ \leq\ \frac{\mathbb{P}(Z_{i1}{=}1 \mid \mathcal{F})/\mathbb{P}(Z_{i1}{=}0 \mid \mathcal{F})}{\mathbb{P}(Z_{i2}{=}1 \mid \mathcal{F})/\mathbb{P}(Z_{i2}{=}0 \mid \mathcal{F})} \ \leq\ \Gamma.$$

[1] Rosenbaum, P. R. (2002), *Observational Studies*, 2nd ed., Ch. 4.

What is Design Sensitivity? (1/2)

- **How do we calculate the power in RCT? The stance switching.**
  - Assume H_0 is true, come up with a rule to control Type I error rate
  - Assume H_0 is false, calculate the power of that rule
- **Power of a sensitivity analysis**
  - **Rule:** reject H_0 if the maximal p-value over hidden bias up to Γ is $\leq \alpha$.
  - **Power**$_\Gamma$: if a real effect is present and there is, in truth, no hidden bias, how often does this rule still reject H_0?
- **As Γ — the hidden-bias level we guard against to control the Type I error rate — grows, Power_Γ falls.**

What is Design Sensitivity? (2/2)



Design sensitivity $\widetilde{\Gamma}$

$\widetilde{\Gamma}$ is the largest hidden-bias level Γ the test can withstand before its power collapses: as $I \to \infty$, $\text{Power}_\Gamma \to 1$ for $\Gamma < \widetilde{\Gamma}$, but $\to 0$ once $\Gamma > \widetilde{\Gamma}$.

Design Sensitivity — Sign Test (1/3)

Matched-pair setup (P4)

I pairs, treated−control difference $D_i = (2Z_{i1}-1)(Y_{i1}-Y_{i2})$. *Favorable situation*: a real effect, no hidden bias, $D_i \overset{\text{iid}}{\sim} F$ (continuous, not symmetric about 0); test Fisher's sharp null.

Sign test: $T = \sum_{i=1}^{I} \mathbb{1}\{D_i > 0\}$ — governed by $p_1 = \mathbb{P}(D_i > 0)$.

Design sensitivity (P4, Prop. 2)[1]

$$\widetilde{\Gamma}_{\text{sign}} = \frac{p_1}{1-p_1} = \frac{\mathbb{P}(D_i > 0)}{\mathbb{P}(D_i < 0)}, \qquad p_1 > \tfrac{1}{2}.$$

The *favorable odds ratio*: the sign test uses only the *sign* of each paired difference.

[1] Chiu, E. K. & Kang, H. *"Sample Size Calculation in Causal Inference for Unmeasured Confounding."*

Design Sensitivity — Wilcoxon Signed-Rank Test (2/3)

Matched-pair setup (P4)

Same setup: $D_i = (2Z_{i1}-1)(Y_{i1}-Y_{i2})$, favorable situation, $D_i \overset{\text{iid}}{\sim} F$.

Wilcoxon signed-rank: $W^+ = \sum_{i=1}^{I} R_i \, \mathbb{1}\{D_i > 0\}$, $R_i = \text{rank}(|D_i|)$ — governed by $p_2 = \mathbb{P}(D_i + D_{i'} > 0)$, $D_i \perp D_{i'}$.

Design sensitivity (P4, Prop. 5)[1,2]

$$\widetilde{\Gamma}_{\text{Wil}} = \frac{p_2}{1-p_2} = \frac{\mathbb{P}(D_i + D_{i'} > 0)}{\mathbb{P}(D_i + D_{i'} < 0)}, \qquad i \neq i', \quad p_2 > \tfrac{1}{2}.$$

Wilcoxon uses the *sum of two* differences $D_i + D_{i'}$ — more robust than the sign test.

[1] Chiu, E. K. & Kang, H. *"Sample Size Calculation in Causal Inference for Unmeasured Confounding"*

[2] Rosenbaum, P. R. (2010). *Design of Observational Studies*, Ch. 14. Springer.

Design Sensitivity — Sign Test vs. Wilcoxon Signed-Rank Test (3/3)

Matched-pair setup (P4)

Same setup: $D_i = (2Z_{i1}-1)(Y_{i1}-Y_{i2})$, favorable situation, $D_i \overset{\text{iid}}{\sim} F$.

Design sensitivity

$$\widetilde{\Gamma}_{\text{sign}} = \frac{\mathbb{P}(D_i > 0)}{\mathbb{P}(D_i < 0)}; \quad \widetilde{\Gamma}_{\text{Wil}} = \frac{\mathbb{P}(D_i + D_{i'} > 0)}{\mathbb{P}(D_i + D_{i'} < 0)}, \qquad i \neq i'.$$

Wilcoxon is more powerful than the sign test in light-tail cases (e.g., $D_i \sim$ Normal).

Two design sensitivities are the same under Cauchy-distributed D_i.


Introduction
Categorical Tests
Design Sensitivity
Sample Size
Conclusion

Two Error Rates: The Observational Analog

The observational test is a *sensitivity analysis* at a prespecified bias level $\Gamma \geq 1$ — each guarantee then has a direct analog:

	Randomized trial	**Observational study**
Level α (Type I)	reject when the randomization p-value $\leq \alpha$	reject when the upper-bound p-value under bias up to Γ satisfies $\bar{p}(\Gamma) \leq \alpha$
Power (Type II)	how often we reject when a treatment effect is present using randomization p	how often we reject when an effect is present and free of hidden bias using $\overline{p}(\Gamma)$
Design	pairs I for power $1-\beta$ of the randomization test	pairs I for power $1-\beta$ of the Γ-analysis

Required Pairs — Sign Test

Recall: the sign test

$T = \sum_{i=1}^{I} \mathbb{1}\{D_i > 0\}$, governed by $p_1 = \mathbb{P}(D_i > 0)$; design sensitivity $\widetilde{\Gamma}_{\text{sign}} = p_1/(1-p_1)$.

Favorable situation: effect present, no hidden bias, $D_i \overset{\text{iid}}{\sim} F$. Here z_q denotes the q-quantile of $\mathcal{N}(0,1)$: $z_{1-\alpha}$ for level α, $z_{1-\beta}$ for target power $1-\beta$.

Required matched pairs

$$I(\Gamma) \gtrsim \frac{(z_{1-\alpha} + z_{1-\beta})^2 \, \Gamma \left(\widetilde{\Gamma}_{\text{sign}} + 1\right)^2}{\left(\widetilde{\Gamma}_{\text{sign}} - \Gamma\right)^2}, \quad \widetilde{\Gamma}_{\text{sign}} > \Gamma \geq 1$$

Design sensitivity $\widetilde{\Gamma}_{\text{sign}}$: larger under a stronger effect; the larger it is, the fewer pairs needed.

Bias Γ: more bias to defend against costs more pairs; $I \to \infty$ as $\Gamma \to \widetilde{\Gamma}_{\text{sign}}$.

Required Pairs — Wilcoxon Signed-Rank Test

Recall: the Wilcoxon signed-rank test

$W^+ = \sum_{i=1}^{I} R_i \, \mathbb{1}\{D_i > 0\}$, $R_i = \text{rank}(|D_i|)$, governed by $p_2 = \mathbb{P}(D_i + D_{i'} > 0)$; design sensitivity $\widetilde{\Gamma}_{\text{Wil}} = p_2/(1-p_2)$.

Required matched pairs

$$I(\Gamma) \gtrsim \frac{4\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \Gamma \left(\widetilde{\Gamma}_{\text{Wil}} + 1\right)^2}{3\left(\widetilde{\Gamma}_{\text{Wil}} - \Gamma\right)^2}, \quad \widetilde{\Gamma}_{\text{Wil}} > \Gamma \geq 1$$

Design sensitivity $\widetilde{\Gamma}_{\text{Wil}}$: often larger than the sign test's (sum of two differences) thus fewer pairs.

Bias Γ: more bias to defend against costs more pairs; $I \to \infty$ as $\Gamma \to \widetilde{\Gamma}_{\text{Wil}}$.

Sign Test — Formula Meets Simulation

Differences $D_i \overset{\text{iid}}{\sim} N(\tau, 1)$; $\alpha = 0.05$, target power $1 - \beta = 0.90$.

τ	$\widetilde{\Gamma}_{\text{sign}}$	Γ	formula	simulation
0.5	2.241	1.0	55	55
0.5	2.241	1.5	234	234
0.5	2.241	2.0	3041	3045
0.8	3.720	1.0	22	22
0.8	3.720	1.5	50	53
0.8	3.720	2.0	115	115
1.0	5.303	1.0	15	15
1.0	5.303	1.5	28	31
1.0	5.303	2.0	51	50

Takeaway

The formula matches the simulation to within a few pairs.

Required pairs explode as $\Gamma \to \widetilde{\Gamma}_{\text{sign}}$: at $\tau = 0.5$, $55 \to 3041$.

Simulation benchmark. Used package (`mlpwr`; Zimmer et al., 2023), within $\pm 20\%$ of the formula.


Introduction
Categorical Tests
Design Sensitivity
Sample Size
Conclusion

Conclusion

In sum, the tools provided in this thesis allow more rigorous inference from observational data. Thank you — I welcome your questions.

Backup — Q&A Menu I

- Fisher's sharp null vs. the weak null — and why we test the sharp null
- Paper 1's proof strategy: shrinking the search for $\mathbf{u}^+$ (overview)
- The corner solution: why $\mathbf{u}^+$ sits at a vertex (Lemma 1)
- Permutation invariance & equivalence classes
- Arrangement-increasing functions
- Casella–Berger 8.3.27: why the maximal p-value is valid
- Compare different existing sensitivity models
- What are the matching methods we apply?
- What are the advantages of matching as opposed to weighting?
- How is design sensitivity derived? — the bounding distribution & LLN

Backup — Q&A Menu II

- General notion of asymptotic relative efficiency: Pitman, Bahadur, Hodges–Lehmann
- What is Bahadur slope?
- Summary of the E-value: sensitivity analysis *without assumptions* (Ding & VanderWeele)
- Regression-coefficient sensitivity analysis: omitted variable bias (Cinelli & Hazlett)
- ATE estimation under a parametric sensitivity model (Imbens)
- Design-based vs. sampling-based inference: fixed outcomes, Neyman, and the estimand

Backup — Fisher's Sharp Null vs. Weak Null

Q [Hypotheses] Why do we test Fisher's sharp null, and how does it differ from the weak null?

A. Fisher's sharp null, $H_0^{\text{sharp}} : Y_i(1) = Y_i(0)$ for every unit i, says the treatment changes *no* unit's outcome. It fixes *all* potential outcomes — each unit's missing counterfactual equals its observed value — so the full randomization distribution of any test statistic is known exactly: finite-sample, distribution-free p-values. This is exactly what Rosenbaum-style sensitivity analysis perturbs.

Sharp null $\Rightarrow$ every counterfactual is known $\Rightarrow$ exact randomization inference.

▸ weak null & arguments ◂ back to menu

Backup — Fisher's Sharp Null vs. Weak Null

A. (cont.) The weak (Neyman) null constrains only the average effect, H_0^{weak} : $\frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = 0$, allowing nonzero individual effects that cancel.

Some Arguments The sharp null is still valuable: if we *fail* to reject it, we cannot reject the weak null either — since H_0^{sharp} holding implies H_0^{weak} holds.

Some Arguments Neyman's weak null does not allow us to fill in all values of the missing potential outcomes, and so we cannot derive the exact randomization distribution of statistics[1]

Some Arguments Trade-off: the sharp null yields an *exact* p-value, while weak-null inference often relies on asymptotics.

[1] Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.*

back to menu

Backup — Paper 1's Proof Strategy: Shrinking the Search for $\mathbf{u}^+$

Q [Proof method] How does Paper 1 turn the worst-case search into a *finite, small* problem[1]?

The upper-bound p-value maximizes over the continuous cube, which is uncountably infinite with α non-concave in $\mathbf{u}$:

$$\mathbf{u}^+ = \underset{\mathbf{u}\in[0,1]^N}{\arg\max}\, \alpha(T, \mathbf{r}, \mathbf{u}).$$

Three tools shrink the search, each responsible for one stage:

Search space for $\mathbf{u}$	Cardinality	Tool that shrinks it
$\mathbf{u} \in [0,1]^N$	uncountable	—
$\mathbf{u} \in \{0,1\}^N$ (vertices)	2^N	corner solution
equivalence classes $\mathcal{U}_{\mathrm{C}}$	$\leq N^J$	permutation invariance
ordinal test $\mathcal{U}_{\mathrm{O}}$	$N+1$	arrangement-increasing structure
sign-score test	1	closed-form vertex (lattice + Holley)

Takeaway

Infinite cube $\to 2^N$ vertices $\to \leq N^J$ classes $\to N{+}1 \to$ one closed-form vertex. The corner solution and permutation invariance do the heavy lifting; the ordinal/sign-score structure finishes the job.

[1] Chiu, E. & Kang, H. *"Exact, Nonparametric Sensitivity Analysis for Observational Studies of Contingency Tables."* (P1), Thms. 1–3.

the corner solution · back to menu

Backup — The Corner Solution: Why u^+ Sits at a Vertex (Lemma 1)

Goal. Show the worst case over the cube is attained at a *vertex*, i.e. $\mathbf{u}^+ \in \{0,1\}^N$, collapsing the infinite search to 2^N points.

The coordinate-perturbation vector ε

Pick a coordinate s and let $\boldsymbol{\varepsilon}_s$ be the N-tuple with a *one* in coordinate s and *zeros* in the other $N-1$ coordinates. Then $\mathbf{u} + \delta\,\boldsymbol{\varepsilon}_s$ moves a single entry of $\mathbf{u}$ by δ while holding all the rest fixed.

Coordinatewise monotonicity $\Rightarrow$ vertex

For each fixed $\mathbf{r}, \mathbf{u}$ and s, the objective $\alpha(T, \mathbf{r}, \mathbf{u}+\delta\varepsilon_s)$ is *monotone* in δ: writing it as $\frac{A_0+\exp(\gamma\delta)A_1}{D_0+\exp(\gamma\delta)D_1}$ with $D_0, D_1 > 0$ constant in δ,

$$\frac{\partial\,\alpha(T, \mathbf{r}, \mathbf{u}+\delta\boldsymbol{\varepsilon}_s)}{\partial\delta} = \frac{(D_0A_1 - A_0D_1)\,\gamma\exp(\gamma\delta)}{\{D_0 + D_1\exp(\gamma\delta)\}^2},$$

whose sign never changes with δ. So along each coordinate the maximum sits at an *endpoint* $u_s \in \{0,1\}$; iterating coordinatewise pushes $\mathbf{u}^+$ to a vertex.

[1] Chiu & Kang (P1), Appendix B.2, Corner-Solution Lemma

permutation invariance · back to menu

Backup — Permutation Invariance & Equivalence Classes

Permutation-invariant test

A test statistic is *permutation-invariant* if simultaneously permuting the same pair of indices of $\mathbf{Z}$ and $\mathbf{r}$ leaves its value unchanged. Every count-based statistic qualifies: it sees $(\mathbf{Z}, \mathbf{r})$ only through the cell counts.

Viewed as a function of $(\mathbf{Z}, \mathbf{u}, \mathbf{r})$, the null distribution then *cannot distinguish* confounder vectors that yield the same aggregated outcome counts. So $\{0,1\}^N$ collapses into equivalence classes $\mathcal{U}_C$ with $|\mathcal{U}_C| \leq N^J$ — only one representative per class is evaluated.

Example. Fix $\mathbf{r} = (0,0,0,1,1)$. Then

$$\mathbf{u}_1 = (0,0,0,1,0), \qquad \mathbf{u}_2 = (0,0,0,0,1) \quad \Longrightarrow \quad \mathbf{r}^\intercal \mathbf{u}_1 = \mathbf{r}^\intercal \mathbf{u}_2 = 1,$$

so $\mathbf{u}_1$ and $\mathbf{u}_2$ give the same distribution and belong to one equivalence class.

Takeaway

Permutation invariance is the bridge from the 2^N vertices to the polynomial bound $|\mathcal{U}_C| \leq N^J$ — the reduction that makes the *exact* worst-case p-value tractable.

[1] Chiu & Kang (P1), Thm. 1; Rosenbaum, P. R. & Krieger, A. M. (1990). JASA, 85(410), 493–498, Sec. 3.

arrangement-increasing back to menu

Backup — Arrangement-Increasing Functions

Arrangement-increasing (AI) — Appendix B.5

A permutation-invariant function $T(\mathbf{X}, \mathbf{Y})$ is *arrangement-increasing* (AI), also called *decreasing in transposition* (DT), if

$$T(\mathbf{X},\ {}_{ab}\mathbf{Y}) \ \geq\ T(\mathbf{X}, \mathbf{Y}) \qquad \text{whenever} \qquad (X_a - Y_a)(X_b - Y_b) \ \leq\ 0,$$

where ${}_{ab}\mathbf{Y}$ is $\mathbf{Y}$ with its a-th and b-th entries exchanged: the value cannot decrease when a swap makes $\mathbf{X}$ and $\mathbf{Y}$ *more concordant*.

Example. The inner product $\mathbf{Z}^{\intercal}\mathbf{r}$ is arrangement-increasing. Fix $\mathbf{r} = (0, 0, 0, 1, 1)$; then for

$$\mathbf{Z}_1 = (0, 1, 0, 0, 0), \qquad \mathbf{Z}_2 = (0, 0, 0, 1, 0),$$

we get $\mathbf{Z}_2^{\intercal}\mathbf{r} = 1 \ \geq\ \mathbf{Z}_1^{\intercal}\mathbf{r} = 0$, because $\mathbf{Z}_2$ puts its mass where $\mathbf{r}$ is large — a concordant ordering.

Why it matters

For the ordinal test, $T(\mathbf{Z}, \mathbf{r})$ — and hence the significance level — is AI in $\mathbf{u}$ and $\mathbf{r}$. Combined with the corner solution ($\mathbf{u}^{+} \in \{0,1\}^N$), AI forces $\mathbf{u}^{+}$ to *share the ordering of* $\mathbf{r}$, collapsing the search to $\mathcal{U}_{\mathrm{o}} = \{\mathbf{u} \in \{0,1\}^N : u_1 \leq \cdots \leq u_N\}$ — just $N{+}1$ candidates.

[1] Chiu & Kang (P1), Appendix B.5 (candidate set $\mathcal{U}_{\mathrm{o}}$ for ordinal tests). Definition per Rosenbaum, P. R. (2002), *Observational Studies*, 2nd ed., Sec. 2.3, Def. 3, and Hollander, Proschan & Sethuraman (1977), *Ann. Statist.* 5(4), 722–733, Sec. 2.

back to menu

Backup — Why the maximal p-value Is Valid

Q [Validity] Why does taking the supremum over the null give a *valid* p-value?

A. Casella & Berger (2002), Thm. 8.3.27. For a test statistic W whose large values favor H_1,

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(W(\mathbf{X}) \geq W(\mathbf{x})\big)$$

is a **valid** p-value: $\mathbb{P}_\theta\big(p(\mathbf{X}) \leq \alpha\big) \leq \alpha$ for every $\theta \in \Theta_0$ and every $\alpha \in [0, 1]$.

The sensitivity-analysis maximal p-value is exactly this sup — over the bias family Θ_0 — so it is valid by the same theorem.

back to menu

Backup — Compare Different Existing Sensitivity Models

Q [Sensitivity models] Which sensitivity models do the papers use, and how do they compare?

Paper 1 — the δ model (categorical treatment)[1]

$$\mathbb{P}(Z_s{=}i \mid \mathcal{F}) \;\propto\; \exp\big(\kappa_i(\mathbf{x}_s)\big) \cdot \exp\big(\gamma\, \delta_i\, u_s\big), \qquad \delta_i \in \{0,1\},\ u_s \in [0,1],$$

with $\exp(\kappa_i(\mathbf{x}_s))$ the covariate term and δ_i marking whether level i is favored: the treatment-odds ratio between two subjects with the same $\mathbf{x}$ is bounded by $\Gamma = e^{\gamma}$.

Paper 2 — the $\phi(z)$ model (general treatment)[2]

$$\mathbb{P}(Z_s{=}z \mid \mathbf{x}_s, u_s) \;\propto\; \eta(z, \mathbf{x}_s) \cdot \exp\big(\gamma\, \phi(z)\, u_s\big), \qquad \phi \text{ bijective, monotone, } u_s \in [0,1],$$

with $\eta(z, \mathbf{x}_s)$ an arbitrary unknown covariate term and ϕ a dose-link function prespecified by the researcher (the methods and theory are not tied to any specific choice of ϕ).

The only difference

Same skeleton: assignment probability $\propto$ (covariate term) $\times \exp(\gamma \cdot \text{link} \cdot u)$. The link is $\delta_i \in \{0,1\}$ for categories, $\phi(z)$ for general doses; for binary treatment both reduce to Rosenbaum's model.

[1] Rosenbaum (2006), *Biometrika* 93(3):573–586.
[2] Rosenbaum (1987, 1989, 2002); Gastwirth, Krieger & Rosenbaum (1998).

back to menu

Backup — What Matching Methods Do We Apply?

Q [Matching methods] The matching literature uses many related terms — what do they mean[1,2]?

- **Optimal matching**[1]: Forming matching sample by minimizing total pair distance.
- **Nonbipartite matching**: Pool partitioned into pairs, used with dose.
- **Cardinality matching**: Find the largest subset of units given balance constraints.
- **Fine balance**: marginal distribution of covariates among the trt/ctrl comparable.
- **Refined balance**: nested near-fine-balance constraints, enforced in priority order.
- **Multilevel matching** — ensure cluster- and individual-covariates balance.

[1] Rosenbaum, P. R. (2020). *Design of Observational Studies*, 2nd ed. Springer.
[2] Zubizarreta, J. R., Stuart, E. A., Small, D. S. & Rosenbaum, P. R. (Eds.) (2023). *Handbook of Matching and Weighting Adjustments for Causal Inference*. CRC Press.

back to menu

Backup — Matching vs. Weighting: Trade-offs

Q [Alternatives] Matching, weighting, stratifying, and regression adjustment.

A. All four adjust for observed confounding, but differ along three dimensions:

	Matching	Weighting	Stratification	Regression
Needs outcome[1,3]	×	×	×	✓
Removes bias[2]	✓	✓	×	×
Mimics RCT[1]	✓	×	×	×

[1] Austin, P. C. (2011). *Intro. to propensity score methods for confounding*. Multiv. Behav. Res. 46(3), 399–424.

[2] Austin, P. C. (2009). *Relative ability of PS methods to balance covariates*. Med. Decis. Making, 29(6), 661–677.

[3] Rubin, D. B. (2001), *Using propensity scores to help design observational studies: application to the tobacco litigation*. Health Services and Outcomes Research Methodology, 2(3–4), 169–188.

Only matching literally builds comparable treated/control sets, the way a paired RCT would.

back to menu

Backup — Design Sensitivity: The Sensitivity Model (1/4)

Q [Design sensitivity] How is design sensitivity derived, and what does the derivation require[1]?

The sensitivity model

With $\mathcal{F} = \{Y_{ij}(0), Y_{ij}(1), x_{ij}\}$ held fixed, $\mathcal{Z} = \{\mathbf{z} : z_{i1}+z_{i2}=1 \ \forall i\}$, and $\pi_{ij} = \mathbb{P}(Z_{ij}=1 \mid \mathcal{F})$, the model allows that

$$\frac{1}{\Gamma} \ \leq \ \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \ \leq \ \Gamma,$$

or equivalently, conditional on the matched-pair design,

$$\mathbb{P}(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod_{i=1}^{I} \frac{\exp\big(\gamma z_{i1} u_{i1} + \gamma z_{i2} u_{i2}\big)}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})}, \qquad \mathbf{u} \in [0,1]^{2I}.$$

The test statistic is $T = \sum_{i=1}^{I} \operatorname{sign}(Y_i)\, \varphi(q_i)$.

[1] Rosenbaum, P. R. (2010). *"Design Sensitivity and Efficiency in Observational Studies."* Journal of the American Statistical Association, 105(490), 692–702.

the requirement | back to menu

Backup — The Requirement: A Bounding Distribution (2/4)

One important result — and also the requirement — is that we are able to find the **bounding distribution** of T[1]: a random variable whose distribution is stochastically larger than that of T.

Bounding variable

Let $\overline{T}_\Gamma$ be the sum of I independent random variables taking the value $\varphi(q_i)$ with probability $\Gamma/(1+\Gamma)$ and the value 0 with probability $1/(1+\Gamma)$. Then, under Fisher's sharp null of no effect,

$$\mathbb{P}\big(T \geq k \mid \mathcal{F}, \mathcal{Z}\big) \;\leq\; \mathbb{P}\big(\overline{T}_\Gamma \geq k \mid \mathcal{F}, \mathcal{Z}\big) \qquad \text{for all } \mathbf{u} \in [0,1]^{2I}.$$

Mean and variance of the bounding variable

$$\mathbb{E}\big(\overline{T}_\Gamma \mid \mathcal{F}, \mathcal{Z}\big) = \frac{\Gamma}{1+\Gamma} \sum_{i=1}^{I} s_i\, \varphi(q_i), \qquad \mathrm{Var}\big(\overline{T}_\Gamma \mid \mathcal{F}, \mathcal{Z}\big) = \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} s_i\, \varphi(q_i)^2.$$

[1] Rosenbaum, P. R. (2010). *"Design Sensitivity and Efficiency in Observational Studies."* Journal of the American Statistical Association, 105(490), 692–702.

the rejection rule · back to menu

Backup — The Rejection Rule (3/4)

Since we have the mean and the variance of the bounding variable, under the sensitivity analysis we reject the null hypothesis if[1]

$$\frac{T - \mathbb{E}(\overline{T}_\Gamma)}{\sqrt{\mathrm{Var}(\overline{T}_\Gamma)}} \geq \Phi^{-1}(1-\alpha)$$

$$\Longrightarrow \quad \frac{\frac{T}{I} - \frac{\Gamma}{I(1+\Gamma)} \sum_{i=1}^{I} s_i \, \varphi(q_i)}{\sqrt{\frac{\Gamma}{I^2(1+\Gamma)^2} \sum_{i=1}^{I} s_i \, \varphi(q_i)^2}} \geq \Phi^{-1}(1-\alpha).$$

Orders

T and $\mathbb{E}(\overline{T}_\Gamma)$ are both $O(I)$, whereas the denominator is of the lower order $O(1)$ after this rescaling — essentially the consistency notion.

[1] Rosenbaum, P. R. (2010). *"Design Sensitivity and Efficiency in Observational Studies."* Journal of the American Statistical Association, 105(490), 692–702.

the law of large numbers | back to menu

Backup — The Law of Large Numbers Gives $\widetilde{\Gamma}$ (4/4)

By the law of large numbers, T/I converges to the mean of T/I, which usually has an expression in terms of T and the data-generating process[1].

If there is a true treatment effect

The true mean of the test statistic converges to that mean in probability, so in the limit as $I \to \infty$ the power goes to one if the mean under the treatment effect is larger than the mean under the sensitivity analysis: **a larger sample helps us distinguish the treatment effect from the assumed hidden bias.**

Design sensitivity

Solving

$$\mathbb{E}(T) - \mathbb{E}(\overline{T}_\Gamma) > 0$$

makes the whole term in the rejection rule exceed $\Phi^{-1}(1-\alpha)$, and gives a probability going to one as the denominator shrinks in I. The Γ at which $\mathbb{E}(T) - \mathbb{E}(\overline{T}_\Gamma) = 0$ is the design sensitivity $\widetilde{\Gamma}$.

[1] Rosenbaum, P. R. (2010). *"Design Sensitivity and Efficiency in Observational Studies."* Journal of the American Statistical Association, 105(490), 692–702.

back to menu

Backup — Sign-Test Sample Size Under Bias: Recipe & Rejection Region (1/2)

Q [Sample size] How is the required pair size for the sign test derived under bias Γ[1]?

Three steps, run in the two-stance order: **(1)** rejection region *under the null*; **(2)** power *under the alternative*; **(3)** solve power $\geq 1-\beta$ for I.

Step 1 — rejection region (null stance, worst case over bias)

Under randomization the signs are fair coin flips, so $T = \sum_{i=1}^{I} \mathbb{1}\{D_i > 0\} \sim \mathrm{Bin}\big(I, \frac{1}{2}\big)$. Hidden bias Γ raises the worst-case null probability $\mathbb{P}(D_i > 0)$ from $\frac{1}{2}$ to $\frac{\Gamma}{1+\Gamma}$, so the bounding null law is $\mathrm{Bin}\big(I, \frac{\Gamma}{1+\Gamma}\big)$. A normal approximation rejects when

$$T \;\geq\; I \cdot \tfrac{\Gamma}{1+\Gamma} \;+\; z_{1-\alpha}\sqrt{I \cdot \tfrac{\Gamma}{(1+\Gamma)^2}}.$$

[1] Chiu & Kang (P4), *"Sample Size Calculation in Causal Inference for Unmeasured Confounding,"* §Sign test for the matched-pair design (Prop. "Sign test, sensitivity analysis"); full proof in the Supplementary Material.

power & solving for I · back to menu

Backup — Sign-Test Sample Size Under Bias: Power & Solving for I (2/2)

Step 2 — power (alternative stance, favorable situation)

With a real effect and no actual bias, $T \sim \mathrm{Bin}(I, p_1)$ (mean Ip_1, variance $Ip_1(1-p_1)$). Standardizing against the Step-1 threshold,

$$\text{power} = 1 - \Phi\left(\frac{I\left(\frac{\Gamma}{1+\Gamma} - p_1\right) + z_{1-\alpha}\sqrt{I \cdot \frac{\Gamma}{(1+\Gamma)^2}}}{\sqrt{I\,p_1(1-p_1)}}\right).$$

Step 3 — set power $\geq 1-\beta$, solve for I

Rearranging power $\geq 1-\beta$ gives

$$\sqrt{I}\left(p_1 - \tfrac{\Gamma}{1+\Gamma}\right) \;\geq\; z_{1-\alpha}\sqrt{\tfrac{\Gamma}{(1+\Gamma)^2}} + z_{1-\beta}\sqrt{p_1(1-p_1)}.$$

The identity $p_1 - \frac{\Gamma}{1+\Gamma} = \dfrac{\widetilde{\Gamma}_{\text{sign}} - \Gamma}{(\widetilde{\Gamma}_{\text{sign}} + 1)(\Gamma + 1)}$ (from $\widetilde{\Gamma}_{\text{sign}} = p_1/(1-p_1)$) turns the solution into the required-pairs formula on the main slide.

The two-stance payoff

At $\Gamma = 1$ the worst-case probability $\frac{\Gamma}{1+\Gamma}$ is exactly $\frac{1}{2}$: the rejection region collapses to the randomization one and the formula reduces to the RCT sample size. The Γ-analysis is the RCT calculation run against a bias-guarded threshold.

[1] Chiu & Kang (P4), §Sign test for the matched-pair design; full proof in the Supplementary Material.

back to step 1 · back to menu

Backup — Asymptotic Relative Efficiency: The General Notion

A. Let $\{T_n\}$ and $\{V_n\}$ be two sequences of statistics, based on n observations. Let $N_T(\alpha, \beta, \theta)$ be the sample size $\{T_n\}$ needs to have Type II error rate below β at with Type I error rate α under alternative value θ against null θ_0. The relative efficiency of $\{T_n\}$ w.r.t. $\{V_n\}$ is

$$e_{T,V}(\alpha, \beta, \theta) \;=\; \frac{N_V(\alpha, \beta, \theta)}{N_T(\alpha, \beta, \theta)}.$$

Limiting ARE		$\theta \to \theta_0$	$\beta \to 0$	$\alpha \to 0$
Pitman	$e^{P}_{T,V}$	✓	×	×
Bahadur	$e^{B}_{T,V}$	×	×	✓
Hodges–Lehmann	$e^{HL}_{T,V}$	×	✓	×

✓ = parameter driven to its limit; × = held fixed.

[1] Nikitin, Y. Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press.

back to menu

Backup — What Is the Bahadur Slope?

- Define T_n^{θ},(drawn under the alternative θ), and T_n^{0}, the same statistic under the null. With the null CDF $F_n(t) := \mathbb{P}(T_n^0 \leq t)$, the reported p-value under alternative is $p_n := 1 - F_n(T_n^{\theta})$.
- Under the alternative, $p_n \to 0$ a.s. for *every* consistent test; the informative question is how fast. The decay is exponential[1]:

$$-\tfrac{2}{n} \log p_n \xrightarrow{\text{a.s.}} c(\theta).$$

- Bahadur relative efficiency = the ratio of slopes:

$$e^{B}_{T,V} \;=\; \lim_{\alpha \to 0} \frac{N_V(\alpha, \beta, \theta)}{N_T(\alpha, \beta, \theta)} \;=\; \frac{c_T(\theta)}{c_V(\theta)}$$

- Closed form[2,3]: $c(\theta)$ is computable given the MGF of the null.

[1] Bahadur, R. R. (1960). *Stochastic comparison of tests.* Ann. Math. Statist. 31(2), 276–295.
[2] Sievers, G. L. (1969). *On the probability of large deviations and exact slopes.* Ann. Math. Statist. 40(6), 1908–1921.
[3] Plachky, D. (1971). *On a theorem of G. L. Sievers.* Ann. Math. Statist. 42(4), 1442–1443.

back to menu

Backup — Summary of the E-value: Setup & Notation (1/2)

Q [E-value] What is the E-value[1]?

A. It bounds how far an unmeasured confounder U can pull the risk ratio between treatment and outcome while allowing U to be binary, categorical or continuous.

Notation

E: exposure (binary; or any two levels of a categorical/continuous treatment);
D: **binary** outcome;
U: unmeasured confounder(s), *general type*.

$$RR_{ED}^{obs} = \frac{\mathbb{P}(D{=}1 \mid E{=}1)}{\mathbb{P}(D{=}1 \mid E{=}0)},$$

$$RR_{ED}^{true} = \frac{\sum_k \mathbb{P}(D{=}1 \mid E{=}1,\, U{=}k)\,\mathbb{P}(U{=}k)}{\sum_k \mathbb{P}(D{=}1 \mid E{=}0,\, U{=}k)\,\mathbb{P}(U{=}k)},$$

$$RR_{EU} = \max_k \frac{\mathbb{P}(U{=}k \mid E{=}1)}{\mathbb{P}(U{=}k \mid E{=}0)},$$

$$RR_{UD} = \max_{e \in \{0,1\}} \frac{\max_k \mathbb{P}(D{=}1 \mid E{=}e,\, U{=}k)}{\min_k \mathbb{P}(D{=}1 \mid E{=}e,\, U{=}k)}.$$

[1] Ding, P. & VanderWeele, T. J. (2016). *"Sensitivity Analysis Without Assumptions."* Epidemiology, 27(3), 368–377.

the bounding factor · back to menu

Backup — Summary of the E-value: The Bounding Factor (2/2)

Everything runs through one object, the **bounding factor**

$$BF_U \;=\; \frac{RR_{EU}\,RR_{UD}}{RR_{EU}+RR_{UD}-1},$$

a function of just *two* sensitivity parameters.

Result 1 — the sharp bound

Without any assumption on U,

$$RR_{ED}^{true} \;\geq\; \frac{RR_{ED}^{obs}}{BF_U} \qquad \text{i.e.} \qquad \frac{RR_{EU}\,RR_{UD}}{RR_{EU}+RR_{UD}-1} \;\geq\; \frac{RR_{ED}^{obs}}{RR_{ED}^{true}}.$$

[1] Ding, P. & VanderWeele, T. J. (2016). *"Sensitivity Analysis Without Assumptions."* Epidemiology, 27(3), 368–377.

back to menu

Backup — Regression-Coefficient Sensitivity (OVB): Setup & Notation (1/2)

Q [OVB] How does the omitted variable bias framework measure sensitivity[1]?

A. It asks how much an omitted confounder U shifts the OLS coefficient $\hat{\tau}$, indexing U's strength by *variance explained* (partial R^2) rather than by an odds bound Γ.

Notation

Y: outcome; Z: treatment (continuous or binary); $\boldsymbol{X}$: observed covariates; U: single unobserved covariate. The model we want, and the restricted one we can run:

$$Y = \hat{\tau} Z + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\gamma}\, U + \hat{\varepsilon}_{\text{full}},$$

$$Y = \hat{\tau}_{\text{res}}\, Z + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{res}} + \hat{\varepsilon}_{\text{res}}.$$

With $Y^{\perp \boldsymbol{X}}, Z^{\perp \boldsymbol{X}}, U^{\perp \boldsymbol{X}}$ the residuals after removing $\boldsymbol{X}$, the imbalance is $\hat{\delta} = \text{cov}(Z^{\perp \boldsymbol{X}}, U^{\perp \boldsymbol{X}})/\text{var}(Z^{\perp \boldsymbol{X}})$.

[1] Cinelli, C. & Hazlett, C. (2020). *"Making Sense of Sensitivity: Extending Omitted Variable Bias."* JRSS-B, 82(1), 39–67.
See also, on adjusting a regression coefficient for hidden confounding: Frank (2000), "Impact of a Confounding Variable on a Regression Coefficient"; and Oster (2019), "Unobservable Selection and Coefficient Stability: Theory and Evidence."

▸ the bias & its R^2 form ◂ back to menu

Backup — Regression-Coefficient Sensitivity (OVB): The Bias (2/2)

Bias in regression coefficient form

The bias from omitting U is its **impact** times its **imbalance**,

$$\hat{\text{bias}} = \hat{\tau}_{\text{res}} - \hat{\tau} = \hat{\gamma}\,\hat{\delta}.$$

Bias in partial-R^2 form

$$|\hat{\text{bias}}| = \sqrt{\frac{R^2_{Y\sim U|Z,\boldsymbol{X}}\, R^2_{Z\sim U|\boldsymbol{X}}}{1 - R^2_{Z\sim U|\boldsymbol{X}}}}\ \frac{\text{sd}(Y^{\perp \boldsymbol{X},Z})}{\text{sd}(Z^{\perp \boldsymbol{X}})},$$

a function of just two sensitivity parameters — the confounder's partial R^2 with the outcome, $R^2_{Y\sim U|Z,\boldsymbol{X}}$, and with the treatment, $R^2_{Z\sim U|\boldsymbol{X}}$.

[1] Cinelli, C. & Hazlett, C. (2020). *"Making Sense of Sensitivity: Extending Omitted Variable Bias."* JRSS-B, 82(1), 39–67.
See also, on adjusting a regression coefficient for hidden confounding: Frank (2000), "Impact of a Confounding Variable on a Regression Coefficient"; and Oster (2019), "Unobservable Selection and Coefficient Stability: Theory and Evidence."

back to menu

Backup — Imbens (2003): Model & Notation (1/2)

Notation & modeling assumptions

$W \in \{0,1\}$: treatment; $Y(w)$: potential outcome; $\boldsymbol{X}$: covariates; U: single unobserved confounder; τ: constant treatment effect (ATE); α, δ: strength of U on assignment / outcome. Weakened unconfoundedness $Y(0), Y(1) \perp W \mid \boldsymbol{X}, U$, with

$$U \sim \text{Bernoulli}(1/2),$$

$$\mathbb{P}(W{=}1 \mid \boldsymbol{X}, U) = \frac{\exp(\boldsymbol{\gamma}'\boldsymbol{X} + \alpha U)}{1 + \exp(\boldsymbol{\gamma}'\boldsymbol{X} + \alpha U)},$$

$$Y(w) \mid \boldsymbol{X}, U \sim \mathcal{N}\big(\tau w + \boldsymbol{\beta}'\boldsymbol{X} + \delta U,\ \sigma^2\big).$$

[1] Imbens, G. W. (2003). *"Sensitivity to Exogeneity Assumptions in Program Evaluation."* American Economic Review (P&P), 93(2), 126–132.

▸ likelihood & the R^2 map ◂ back to menu

Backup — Imbens (2003): The R^2 Map (2/2)

- Once the model is set, the log-likelihood is explicit in the sensitivity parameters (α, δ), the effect τ, and the nuisances.
- Fix (α, δ), maximize over the rest $\Rightarrow$ bias-adjusted effect $\hat{\tau}(\alpha, \delta)$.
- (α, δ) have equivalent partial-R^2 expressions[1,2].

[1] Imbens, G. W. (2003). *"Sensitivity to Exogeneity Assumptions in Program Evaluation."* AER (P&P), 93(2), 126–132, Sec. II:
$R^2_{W,\mathrm{par}}(\alpha, \delta) = \Big(R^2_W(\alpha, \delta) - R^2_W(0,0)\Big) \Big/ \Big(1 - R^2_W(0,0)\Big)$, with
$R^2_W(\alpha, \delta) = \Big(\hat{\boldsymbol{\gamma}}' \Sigma_{\boldsymbol{X}} \hat{\boldsymbol{\gamma}} + \alpha^2/4\Big) \Big/ \Big(\hat{\boldsymbol{\gamma}}' \Sigma_{\boldsymbol{X}} \hat{\boldsymbol{\gamma}} + \alpha^2/4 + \pi^2/3\Big)$; $R^2_{Y,\mathrm{par}}$ analogously from the outcome regression.

[2] Oster, E. (2019). *"Unobservable Selection and Coefficient Stability."* JBES, 37(2), 187–204, Prop. 1:
$\beta^* = \tilde{\beta} - (\mathring{\beta} - \tilde{\beta}) \Big(R_{\max} - \tilde{R}\Big) \Big/ \Big(\tilde{R} - \mathring{R}\Big)$.

back to menu

Backup — Design-based vs. Sampling-based Inference (1/3)

Q [Frameworks] What is the difference between design-based and sampling-based inference[1]?

A. **Design-based (finite population, Neyman):** the potential outcomes of the finite population are *fixed*; the only source of randomness is the treatment assignment. **Sampling-based (super population):** the subjects are i.i.d. draws from a hypothetical infinite population.

Design-based fixes the units and randomizes assignment; sampling-based re-draws the units from a super-population model.

[1] Ding, P., Li, X. & Miratrix, L.W. (2017). *"Bridging Finite and Super Population Causal Inference."* Journal of Causal Inference, 5(2).

▸ the two tables ◂ back to menu

Backup — Two Kinds of Uncertainty (2/3)

Abadie et al. separate the two sources of uncertainty by *which entries are missing*[1] (✓ observed, ? missing).

Two indicators

$R_i \in \{0,1\}$: *sampling* indicator — whether unit i is drawn into the sample.
$X_i \in \{0,1\}$: *assignment* indicator — which potential outcome is revealed, $Y_i = Y_i^*(X_i)$.

Sampling-based — randomness in R

	Actual			Alt. I		
Unit	Y_i	Z_i	R_i	Y_i	Z_i	R_i
1	✓	✓	1	?	?	0
2	?	?	0	?	?	0
3	?	?	0	✓	✓	1
⋮						

which units are sampled

Design-based — randomness in X

	Actual			Alt. I		
Unit	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i
1	✓	?	1	✓	?	1
2	?	✓	0	?	✓	0
3	?	✓	0	✓	?	1
⋮						

which potential outcome we see

Sampling-based: which units enter the sample (R). Design-based: which potential outcome is revealed (X).

[1] Abadie, A., Athey, S., Imbens, G. W. & Wooldridge, J. M. (2020). *"Sampling-based versus Design-based Uncertainty in Regression Analysis."* Econometrica, 88(1), 265–296.

▸ the estimands ◂ back to menu

Backup — Same Estimator, Two Estimands (3/3)

Under the two frameworks the average treatment effect is defined differently[1,2].

Two estimands

Finite / sample ATE (design-based):

$$\tau_{\mathcal{S}} \;=\; \bar{Y}_1 - \bar{Y}_0 \;=\; \frac{1}{n}\sum_{i=1}^{n}\big\{Y_i(1) - Y_i(0)\big\}.$$

Super-population ATE (sampling-based):

$$\tau \;=\; \mathbb{E}\big\{Y(1) - Y(0)\big\}.$$

Takeaway

The difference-in-means $\hat{\tau} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}}$ is unbiased for *both*: $\tau_{\mathcal{S}}$ conditions on the units in hand, while τ averages over the hypothetical super-population.

[1] Ding, Li & Miratrix (2017), *J. Causal Inference* 5(2). [2] Abadie, Athey, Imbens & Wooldridge (2020), *Econometrica* 88(1), 265–296.

back to menu