MS&E 228: Applied Causal Inference Powered by ML and AI
Lecture 2: Identification by Conditioning (Outcome Regression)

Vasilis Syrgkanis

Stanford University

Winter 2026

Readings: *Applied Causal Inference Powered by ML and AI*, Ch. 5; (optionally) Hernán & Robins, *What If*, Chs. 2–4.

Goals for Today

1. Move beyond RCTs: identify causal effects when treatment is *as-if randomized conditional on covariates* X.
2. Formalize *conditional ignorability* and *overlap/positivity*.
3. Derive the *g-formula* for ATE and ATT and connect it to outcome regression.
4. Learn why *post-treatment* variables are *bad controls* X (birth-weight paradox).
5. Understand why naive stratification explodes in high dimensions and motivates flexible models (and later ML).

From RCTs to the Next Base Case

- In an RCT:

$$(Y(0), Y(1)) \perp\!\!\!\perp D \Rightarrow \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0].$$

- Next base case: a randomized trial where treatment assignment depends on observed variables X (stratified trials, rule-based assignment).
- Key idea: within levels of X, treatment behaves like random assignment.

Healthcare Example

PrecISE and Conditional Randomization

Healthcare Example: PrecISE (Severe Asthma Precision Trial)

- Goal: rapidly evaluate multiple candidate therapies for *severe and/or exacerbation-prone asthma*.
- Design: conducted under a master protocol; participants can receive multiple interventions over time (with placebo periods).[1]
- Key feature for us: **treatment assignment uses observed baseline information** X (biomarker profiles / phenotypes) that the protocol makes explicit.

[1] Israel et al. (2021, *JACI*) and Ivanova et al. (2020, *J. Biopharm. Stat.*); see PubMed 33667479 and 32941098.

PrecISE: Interventions & Biomarkers

Intervention	A priori best subgroup	Prevalence
Imatinib	Eos $<$ 300 cells/μl	62%
Clazakizumab	IL-6 $>$ 3.1 pg/μl	33%
Itacitinib	Eos $\geq$ 300 cells/μl or FeNO $>$ 20 ppb	57%
Cavosonstat	Genotypes	64%
Broncho-Vaxom	Eos $\geq$ 300 cells/μl	38%
Medium Chain Triglycerides (MCT)	FeNO $\geq$ 15 ppb	64%

Pedagogical Simplification: Conditional Randomization

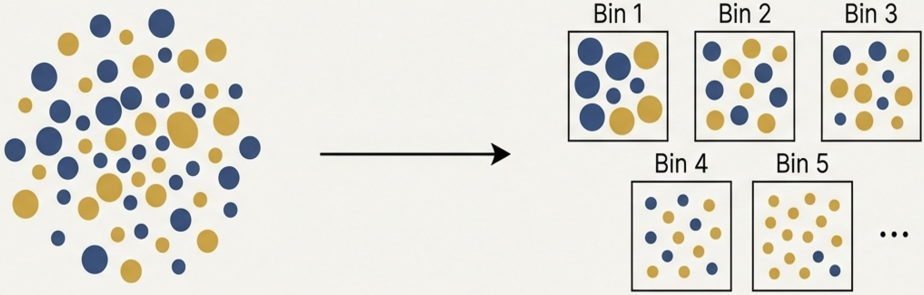
X = Biomarker group (believed responsiveness)	Treatment prob.	Control prob.	Mass
1) (High)	0.70	0.30	35%
2) (Moderate)	0.55	0.45	40%
3) (Low)	0.10	0.90	25%

What goes wrong with a naive two-means estimate

- The treated arm is enriched with a larger fraction of the a-priori high-responder groups.
- A simple *difference in means* between treated and control will therefore be biased upward (it conflates better outcomes due to patient mix with the causal effect of treatment).

The Stratification Solution

Comparing Comparable Units Within Bins



Partition

Divide units into strata based on covariates X.

Compare

Calculate effect within each bin (locally random).

Average

Aggregate bin effects weighted by population size.

Back to PrecISE: Stratify → Weighted Average

X Biomarker group	$\Pr(X{=}x)$ Mass	$\bar{Y}_{1,x}$ Avg. treated outcome	$\bar{Y}_{0,x}$ Avg. control outcome	$\delta(x)$ Group effect
1) High	0.35	11.0	9.5	1.5
2) Moderate	0.40	10.0	9.2	0.8
3) Low	0.25	9.1	9.0	0.1

$$\text{ATE} = \sum_x \underbrace{\Pr(X{=}x)}_{\substack{\text{Mass of} \\ \text{group } x}} \cdot \underbrace{\delta(x)}_{\substack{\text{Within-group} \\ \text{treatment effect}}} = 0.35(1.5) + 0.40(0.8) + 0.25(0.1) = 0.870$$

$$\text{Naive ATE} = \sum_x \Pr(X{=}x \mid D = 1)\,\bar{Y}_{1,x} - \Pr(X{=}x \mid D = 0)\,\bar{Y}_{0,x} \approx 10.45 - 9.17 = 1.28$$

Punchline: naive pooled comparison is generally wrong

$$\underbrace{\mathbb{E}[Y \mid D{=}1] - \mathbb{E}[Y \mid D{=}0]}_{\text{Naive pooled difference}} \neq \underbrace{\sum_x \Pr(X{=}x)\Big(\mathbb{E}[Y \mid D{=}1, X{=}x] - \mathbb{E}[Y \mid D{=}0, X{=}x]\Big)}_{\text{Correct formula}}.$$

Some Useful Mathematical Formalism

Assignment model (conditional randomization)

$$D = f(X, \varepsilon), \qquad \varepsilon \perp\!\!\!\perp \{(Y(0), Y(1)), X\}.$$

- Here X includes unit level characteristics (e.g. biomarkers, prior beliefs) that guide which interventions are randomized and how heavily the trial *enriches* particular subgroups.
- Consequence: $\Pr(D = 1 \mid X = x)$ can differ across x (and can even be 0 for "ineligible" units).

Tech Example

Batch Thompson Sampling in Online Experiments

Tech Example: Batch Thompson Sampling with User Features

- Many tech platforms run A/B tests using a **batch Thompson sampling** rule.[2]
- Each day, for each user that arrives with characteristics W, the platform maintains a belief about the reward difference (typically updated at the end of the day), e.g.

$$\Delta(W) \equiv Y(A) - Y(B) \mid W \sim \mathcal{N}\big(\mu(W),\ \sigma^2(W)\big),$$

  where $\mu(W)$ is the posterior mean and $\sigma(W)$ is the posterior standard deviation.
- For each arriving user, the platform draws a random sample

$$Z \sim \mathcal{N}\big(\mu(W), \sigma^2(W)\big),$$

  and assigns:

$$D = \begin{cases} 1 & \text{(assign A)) if } Z > 0, \\ 0 & \text{(assign B)) if } Z \leq 0. \end{cases}$$

[2] See, e.g., *Using a Multi-Armed Bandit with Thompson Sampling to Identify Responsive Dashers*

How does this fit the conditional randomization formalism?

Suppose we wanted to analyze the results at the end of the day and calculate ATE.

For belief parameters (μ, σ), we have $Z = \mu + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$,

$$\Pr(D = 1 \mid \mu, \sigma) = \Pr(Z > 0 \mid \mu, \sigma) = \Pr\left(\varepsilon > -\frac{\mu}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma}\right),$$

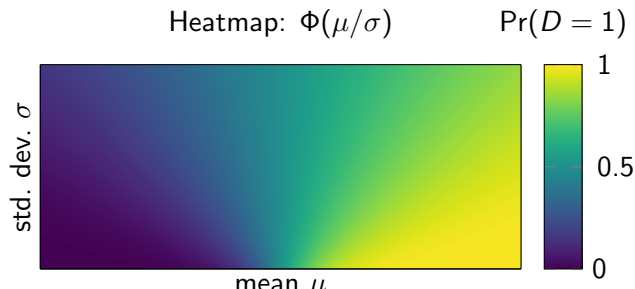
where $\Phi(\cdot)$ is the standard normal CDF.

Conditional on the belief parameters $X = (\mu, \sigma)$ for an arriving user, the only remaining randomness in D is the random seed ε used to draw Z. That seed is independent of the potential outcomes and X:

$$D = f(X, \varepsilon), \qquad \varepsilon \perp\!\!\!\perp \{(Y(0), Y(1)), X\}.$$

Why the abstract mathematical formalism matters?

- Unlike PrecISE, X is **continuous**, so we cannot form a small table of discrete groups.
- At best, we can visualize $\Pr(D = 1 \mid \mu, \sigma) = \Phi(\mu/\sigma)$ as a 2D surface/heatmap.



- This motivates an **abstract identification recipe** that applies to any type of X.

The Identification Recipe

Conditional Ignorability, Overlap, and the g-Formula

Conditional Randomization $\Rightarrow$ Conditional Ignorability

Conditional on X, the only source of randomness in $D = f(X, \varepsilon)$ is ε, and ε is independent of the potential outcomes, conditional on X.

$$\begin{aligned} & \varepsilon \perp\!\!\!\perp \{(Y(0), Y(1)), X\} \\ \Rightarrow \quad & \varepsilon \perp\!\!\!\perp (Y(0), Y(1)) \mid X \\ \Rightarrow \quad & f(X, \varepsilon) \perp\!\!\!\perp (Y(0), Y(1)) \mid X \\ \Rightarrow \quad & D \perp\!\!\!\perp (Y(0), Y(1)) \mid X. \end{aligned}$$

Conditional ignorability / conditional exogeneity

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid X.$$

"*Within a stratum $X = x$, treated and control units are comparable.*"

Poll Everywhere

The average treatment effect in the whole population can be identified in any conditionally randomized trial setting?

(A) Yes, because it satisfies conditional ignorability

(B) No

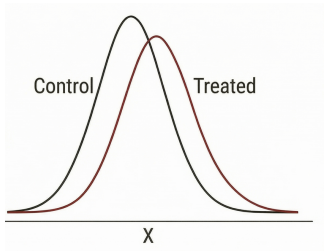
(C) It depends on sample size

Overlap (aka Positivity)

Definition of overlap

$$0 < \Pr(D = 1 \mid X = x) < 1 \qquad \text{for all } x \text{ in the support of } X.$$

- Both treatment arms must be probable at each probable covariate profile.
- This is a *support* condition (not a modeling choice).



Necessity of Overlap

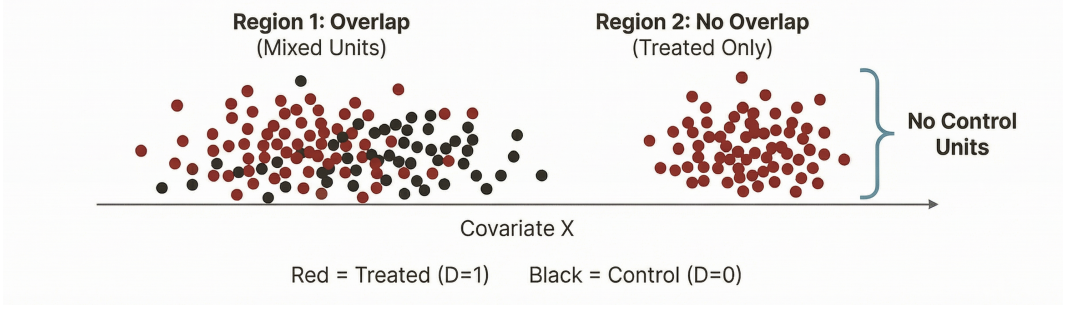
To identify $\mathbb{E}[Y(0) \mid X = x]$ using observables, we need

$$\mathbb{E}[Y(0) \mid X = x] = \mathbb{E}[Y \mid D = 0, X = x],$$

which requires observing some units with $D = 0$ at $X = x$.

If $\Pr(D = 0 \mid X = x) = 0$

Then $\mathbb{E}[Y \mid D = 0, X = x]$ is undefined from data, so $Y(0)$ at that x is not identified without extra assumptions.



Identifying Conditional Mean Counterfactuals

Theorem. Under conditional ignorability and overlap:

$$\mathbb{E}[Y(d) \mid X] = \mathbb{E}[Y \mid D = d, X], \qquad d \in \{0, 1\}.$$

Proof

$$\begin{aligned} \mathbb{E}[Y(d) \mid X] &= \mathbb{E}[Y(d) \mid D = d, X] && \text{(by conditional ignorability + overlap)} \\ &= \mathbb{E}[Y \mid D = d, X] && \text{(by consistency, } Y(d) = Y \text{ when } D = d) \end{aligned}$$

Identification of Conditional Average Treatment Effect (CATE)

CATE

$$\delta(X) \equiv \mathbb{E}[Y(1) - Y(0) \mid X].$$

Under conditional ignorability + overlap, CATE is identified as

$$\delta(X) = \mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X].$$

The g-Formula (ATE Identification by Conditioning)

1. For each covariate profile x, compute the conditional effect $\delta(x)$.
2. Average across x using the distribution of X.

$$\begin{aligned} \text{ATE} = \mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X]] && \text{(by tower rule)} \\ &= \mathbb{E}[\delta(X)] \\ &= \mathbb{E}[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]]. \end{aligned}$$

g-formula for the ATE

Under conditional ignorability + overlap, ATE is identified as

$$\text{ATE} = \mathbb{E}[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]]\,.$$

ATT: Effect on the Treated

Average Treatment effect on the Treated (ATT)

$$\text{ATT} \equiv \mathbb{E}[Y(1) - Y(0) \mid D = 1].$$

- Often the most relevant estimand for policy evaluation among adopters.
- Identification requires weaker ignorability and overlap assumptions than the ATE.

One-Sided Assumptions for ATT

Key simplification

We only need to model outcomes under *control* (predict untreated outcomes for the treated).

One-sided conditional ignorability

$$Y(0) \perp\!\!\!\perp D \mid X.$$

One-sided overlap (support for controls among treated)

$$\Pr(D = 0 \mid X = x) > 0 \quad \text{for all } x \text{ with } \Pr(X = x \mid D = 1) > 0.$$

The ATT g-Formula (ATT Identification by Conditioning)

Theorem. Under one-sided ignorability + one-sided overlap,

$$\mathrm{ATT} = \mathbb{E}[Y \mid D = 1] - \mathbb{E}\Big[\mathbb{E}[Y \mid D = 0, X] \;\Big|\; D = 1\Big].$$

Proof

$$\mathrm{ATT} = \mathbb{E}[Y(1) - Y(0) \mid D = 1] = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y(0) \mid D = 1]$$

$$\begin{aligned}
\mathbb{E}[Y(0) \mid D = 1] &= \mathbb{E}[\mathbb{E}[Y(0) \mid X, D = 1] \mid D = 1] && \text{(by tower rule)} \\
&= \mathbb{E}[\mathbb{E}[Y(0) \mid X] \mid D = 1] && \text{(by one-sided ignorability)} \\
&= \mathbb{E}[\mathbb{E}[Y(0) \mid X, D = 0] \mid D = 1] \\
& && \text{(by one-sided (ignorability + overlap))} \\
&= \mathbb{E}[\mathbb{E}[Y \mid D = 0, X] \mid D = 1] && \text{(by consistency)}
\end{aligned}$$

Observational Studies

No Unmeasured Confounding and Overlap

Observational Studies

In observational studies, conditioning can still identify causal effects *if*:

1. **Conditional ignorability holds:** all confounders are in X.
2. **Overlap:** both treatments occur for the X values we care about.

Key difference from conditionally randomized trials

These assumptions are *not enforced by design*. They are derived from domain expertise and typically untestable.

Teaser from future

In a few weeks we will learn a very useful visual tool (Causal Directed Acyclic Graphs) that help us deduce conditional ignorability from human interpretable domain assumptions.

Homework Example: NHEFS (Smoking Cessation $\rightarrow$ Weight Change)

- D: indicator for quitting smoking (vs continuing).
- Y: weight change over follow-up.
- X: baseline variables (age, sex, baseline weight, smoking intensity, health measures, etc.).

Conditional ignorability here means

After conditioning on baseline X, whether someone quits is unrelated to their potential weight changes.

What Conditional Ignorability Really Says (and Why It Is Untestable)

No unmeasured confounding

There is no unobserved U such that U affects D and predicts $Y(0)$ or $Y(1)$, once we condition on X.

- Data alone cannot verify this: you must rely on domain knowledge and study design.
- Diagnostics (e.g., overlap) are helpful but do not prove conditional ignorability.

Poll Everywhere

When given a dataset that includes D, Y and many other variables related to a unit, which variables should I include as X?

(A) All variables that you have

(B) All variables that you have, other than D, Y

(C) Never a variable that occurs after treatment

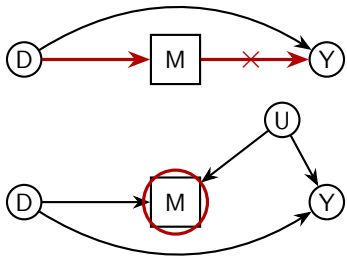
(D) All the variables that occur before treatment

(E) It depends

Bad Controls: Why Post-Treatment Adjustment Can Create Bias

If M is affected by treatment, conditioning on M can:

1. block part of the causal effect (if M is a "mediator"), and/or
2. induce collider bias (if M is a "collider").



Big picture

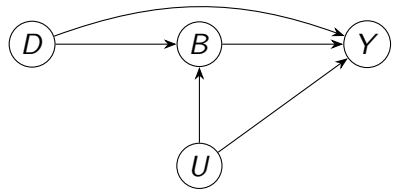
"Control for everything" is *not* a valid strategy in causal inference.

Birth-weight paradox: the empirical findings

- Empirically: infants born to smokers have higher infant mortality than non-smokers.
- But *conditioning on low birth weight* can make the association appear reversed.
- This is a classic illustration that **conditioning/adjusting can go wrong** when we condition on an **endogenous** (post-treatment) variable.
- Has created many controversies in epidemiology [Hernandez et al, Am. J. of Epidemiology, 06]

Birth-weight paradox: let's visualize the problem

- D: smoking during pregnancy
- B: low birth weight
- Y: infant mortality
- U: other unmeasured causes (aka *competing risks*) of low birth weight and mortality (e.g., poor nutrition, infections)



What goes wrong (technically)

- After conditioning on B, the variables D and U become statistically associated.
- The conditional effect we measure is a mixture of the direct effect of D and partially the effect of the "competing risks" U.

Birth-weight paradox: intuition (why the reversal happens)

- Focus on the subpopulation with **low birth weight** ($B = \text{low}$).
- If $D = 1$ (smoking) and B is low, then low B can be "explained" by smoking, so U is *less likely*.
- If $D = 0$ (non-smoking) and B is low, low B must be "explained" by other risks, so U is *more likely*.

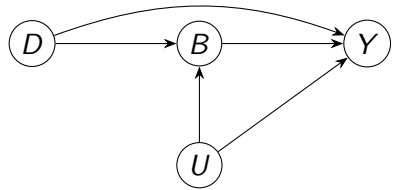
So conditioning on B and comparing treatment vs. control, we compare:

- smokers with relatively *fewer* competing risks (U low)
- to non-smokers with relatively *more* competing risks (U high),

which can make smoking appear "protective" within the low-B group.

Birth-weight paradox: an illustrative linear model

$$
\begin{aligned}
Y &:= D + B + \kappa U + \varepsilon_Y, \\
B &:= D + U + \varepsilon_B, \\
D &:= \varepsilon_D,
\end{aligned}
$$



with $\varepsilon_Y, \varepsilon_B, \varepsilon_D, U \sim \mathcal{N}(0,1)$ independent.

Conditioning on (B, D):

$$\mathbb{E}[Y \mid B, D] = D + B + \kappa\, \mathbb{E}[U \mid B, D].$$

From the B equation, $(B - D) = U + \varepsilon_B$ with independent $U, \varepsilon_B \sim N(0,1)$, so

$$\mathbb{E}[U \mid B, D] = \frac{B - D}{2} \implies \mathbb{E}[Y \mid B, D] = \left(1 - \frac{\kappa}{2}\right) D + \left(1 + \frac{\kappa}{2}\right) B$$

If $\kappa > 2$, the conditional coefficient on D becomes negative (apparent reversal).

Operationalizing the g-Formula

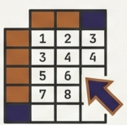
Outcome Regression and Plug-in Estimation

Fit Models

Estimate $m_d(x) = E[Y \mid D=d, X=x]$

Approaches: Two separate models (T/C) or one model with interactions.

Predict Counterfactuals



Predict Y(1) and Y(0) for *every* unit in the dataset, regardless of what they actually received.

Average

Compute mean difference:

$$\frac{1}{n}\sum(\hat{m}_1(x_i) - \hat{m}_0(x_i))$$

Operationalizing the g-Formula: Outcome Regression

Define the outcome regression function:

$$g(d, x) \equiv \mathbb{E}[Y \mid D = d, X = x].$$

Many times we write this as the composition of two regression functions:

$$g(d, x) = d \cdot m_1(x) + (1 - d) \cdot m_0(x)$$

$m_d(x)$ can be thought as predicting Y from X, using samples with $D = d$.

Plug-in estimator (empirical analogue)

Given n i.i.d. samples $\{(Y_i, D_i, X_i)\}_{i=1}^n$: fit regression models $(\hat{m}_0, \hat{m}_1)$ or $\hat{g}$ and compute

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{g}(1, X_i) - \hat{g}(0, X_i)) = \frac{1}{n} \sum_{i=1}^{n} (\hat{m}_1(X_i) - \hat{m}_0(X_i))$$

$$\widehat{\text{ATT}} = \frac{1}{n_1} \sum_{i:D_i=1} (Y_i - \hat{m}_0(X_i)).$$

Two Styles of Implementation

1. **Two-model approach (T-Learner):** fit $\hat{m}_1$ on treated units, $\hat{m}_0$ on controls.
2. **Single interacted model (S-Learner):** fit one model $\hat{g}$ predicting the outcome from treatment D and controls X (if model is linear, advisably include interactions between treatment and features of X to allow for effect heterogeneity).

Linear Regression Adjustment (No Interactions)

Suppose we model

$$\mathbb{E}[Y \mid D, X] = \beta_0 + \alpha D + \gamma^\top \phi(X).$$

- $\phi(X)$: engineered features (polynomials, splines, bins, interactions among covariates).

Then

$$m_1(x) - m_0(x) = \alpha, \qquad \Rightarrow \qquad \text{ATE} = \alpha.$$

Interpretation

Under correct linear specification, the OLS coefficient on D is exactly the g-formula plug-in estimate.

Linear Model with Treatment Effect Heterogeneity

Consider

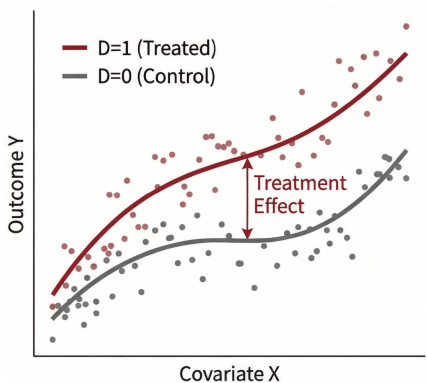
$$\mathbb{E}[Y \mid D, X] = \beta_0 + \beta^\top \phi(X) \; + \; D \cdot \big(\alpha + \theta^\top \psi(X)\big).$$

Then the CATE is

$$\delta(X) = \alpha + \theta^\top \psi(X),$$

and the ATE is

$$\text{ATE} = \mathbb{E}[\delta(X)] = \alpha + \theta^\top \mathbb{E}[\psi(X)].$$



Nonparametric Stratification as a Special Case

If X is discrete, we can stratify exactly:

$$\text{ATE} = \sum_{x} \Big(\mathbb{E}[Y \mid D = 1, X = x] - \mathbb{E}[Y \mid D = 0, X = x] \Big) \Pr(X = x).$$

- Estimation: replace conditional expectations by within-cell means.
- This corresponds to using indicator features for each stratum as both ψ and ϕ in the interactive linear specification.

Continuous Covariates: Discretization Tradeoff

When X includes continuous variables:

- Bin into strata: coarse bins reduce variance but can leave residual confounding.
- Fine bins reduce confounding but create sparse cells with few (if any) samples.

Bias–variance and overlap are intertwined

The more granular the stratification, the harder it is to have data in both treatment arms.

Combinatorial Explosion (Curse of Dimensionality)

If we have d binary covariates, then there are 2^d strata.

- Even with moderate d, most strata are empty.
- This makes purely nonparametric stratification infeasible in observational settings.

Bridge to later lectures

We will use flexible models (and ML) to estimate $m_d(x)$, helping us bypass the curse of dimensionality.

Summary: Identification by Conditioning

Assumptions $\Rightarrow$ Identification

- Conditional ignorability + overlap $\Rightarrow$ g-formula identifies ATE.
- One-sided versions $\Rightarrow$ g-formula identifies ATT.

Identification $\Rightarrow$ Estimation

Outcome regression: estimate conditional expectations and plug them into the g-formula.

Looking Ahead

Next steps in the course:

- Propensity scores and reweighting as an alternative route to the same estimand.
- Doubly robust estimators that combine outcome regression + propensity modeling.
- ML methods for flexible $m_d(x)$ estimation + confidence intervals.