MS&E 228: Applied Causal Inference Powered by ML and AI
Lecture 3: Identification by Propensity (Inverse Weighting)

Vasilis Syrgkanis

Stanford University

Winter 2026

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

Goals for Today

1. Learn the **reweighting lens**: if treated/control samples are not representative, **reweight** them to mimic the target population.
2. Derive the **Horvitz–Thompson (inverse-propensity) identification** for ATE:

$$\text{ATE} = \mathbb{E}\left[Y\left(\frac{\mathbb{I}\{D{=}1\}}{\Pr(D{=}1|X)} - \frac{\mathbb{I}\{D{=}0\}}{\Pr(D{=}0|X)}\right)\right].$$

3. See why this requires **no outcome modeling** when propensities are **known by design** (stratified trials, digital experiments / bandits).
4. Operationalize in observational data by **estimating propensities** (classification) and **checking overlap**.

Where We Are in the "Base Cases"

- Lecture 1: RCTs: $(Y(0), Y(1)) \perp\!\!\!\perp D \Rightarrow \text{ATE} = \text{difference in means}$.
- Lecture 2: Conditional ignorability + overlap:

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid X, \qquad 0 < \Pr(D = 1 \mid X) < 1.$$

- Lecture 2 lens: **conditioning / outcome regression** (g-formula).
- **Today**: same assumptions, **different lens**: **reweighting by inverse propensities**.

The Mantra (Reweighting Intuition)

Main idea

If the treated population (or control population) is *not representative* of the target population, **reweight** its samples to create a *synthetic population* that *looks like* the target.

- If a group is **over-represented** in the treated sample: **downweight** it.
- If a group is **under-represented**: **upweight** it.

After reweighting: a simple weighted average behaves like the average outcome under treating a random sample (an RCT).

Healthcare Example

PrecISE revisited: from stratification to reweighting

"Stylized" PrecISE Reminder: Randomization by Biomarker Group

X = Biomarker group	$\Pr(D = 1 \mid X)$	$\Pr(D = 0 \mid X)$	Mass $\Pr(X)$
1) (High)	0.70	0.30	35%
2) (Moderate)	0.55	0.45	40%
3) (Low)	0.10	0.90	25%

- The trial *intentionally* oversamples treatment among "High" and "Moderate" groups.
- Goal: ATE in the overall population with masses $\Pr(X)$.

"Stylized" PrecISE Reminder: Stratify → Weighted Average

X Biomarker group	$\Pr(X{=}x)$ Mass	$\bar{Y}_{1,x} = \mathbb{E}[Y \mid D=1, X=x]$ Avg. treated outcome	$\bar{Y}_{0,x} = \mathbb{E}[Y \mid D=0, X=x]$ Avg. control outcome
1) High	0.35	11.0	9.5
2) Moderate	0.40	10.0	9.2
3) Low	0.25	9.1	9.0

$$\text{ATE} = \sum_x \underbrace{\Pr(X{=}x)}_{\substack{\text{Mass of} \\ \text{group } x}} \cdot \underbrace{(\bar{Y}_{1,x} - \bar{Y}_{0,x})}_{\substack{\text{Within-group} \\ \text{treatment effect}}} = 0.35(1.5) + 0.40(0.8) + 0.25(0.1) \;=\; 0.870$$

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

The "Wrong" Estimand: Naive Difference in Means

$$\text{ATE} = \sum_x \underbrace{\Pr(X{=}x)}_{\substack{\text{Mass of} \\ \text{group } x}} \bar{Y}_{1,x} - \underbrace{\Pr(X{=}x)}_{\substack{\text{Mass of} \\ \text{group } x}} \bar{Y}_{0,x}$$

$$\begin{aligned} \text{Naive ATE} &= \mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] \\ &= \sum_x \underbrace{\Pr(X{=}x \mid D=1)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in treated}}} \bar{Y}_{1,x} - \underbrace{\Pr(X{=}x \mid D=0)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in control}}} \bar{Y}_{0,x} \end{aligned}$$

Why it fails here

The treated and control groups have *different biomarker composition*: $\Pr(X \mid D=1) \neq \Pr(X)$ and $\Pr(X \mid D=0) \neq \Pr(X)$.

Translation: selection into treatment changes the distribution of X.

Key Reweighting Idea

Can we weight the observed outcomes such that the naive ATE recovers the correct ATE **as simple pooled weighted averages over individuals**?

Key Question: What Weights Make the Pooled Average Correct?

If we weight the treated outcomes by $w_1(X)$, then simple the pooled average of weighted outcomes among treated becomes

$$\mathbb{E}[w_1(X)Y \mid D = 1] = \sum_x \bar{Y}_{1,x}\, w_1(x)\, \Pr(X = x \mid D = 1).$$

For this to recover the correct expression, we want weights $w_1(X)$ such that:

$$\Pr(X = x) \;=\; w_1(X) \cdot \Pr(X = x \mid D = 1).$$

Intuition

$w_1(X)$ should make the treated sample *look like* a random sample from $\Pr(X)$.

Derive Treated Weights: Population Mass / Treated Mass

To match the g-formula term-by-term, choose:

$$w_1(x) = \frac{\Pr(X = x)}{\Pr(X = x \mid D = 1)}.$$

Interpretation: upweight strata underrepresented among treated; downweight overrepresented.

Poll

Poll Everywhere

In a stratum where $\Pr(X = x \mid D = 1)$ is *larger* than $\Pr(X = x)$ (overrepresented among treated), the treated-unit weight $w_1(x)$ should be:

1. larger than 1
2. smaller than 1
3. exactly 1

From $\Pr(X)/\Pr(X \mid D=1)$ to Inverse Propensity

By Bayes' rule:

$$\frac{\Pr(X=x)}{\Pr(X=x \mid D=1)} = \frac{\Pr(X=x)}{\frac{\Pr(X=x)\Pr(D=1|X=x)}{\Pr(D=1)}} = \frac{\Pr(D=1)}{\Pr(D=1 \mid X=x)}.$$

Define:

$$p(X) = \Pr(D=1 \mid X) \qquad \text{(propensity score)}$$

So treated weights can be written as:

$$w_1(X) = \frac{\Pr(D=1)}{p(X)}.$$

Same story for controls: $w_0(X) = \frac{\Pr(D=0)}{\Pr(D=0|X)} = \frac{\Pr(D=0)}{(1-p(X))}$.

The "Naive" Difference Becomes Correct After Weighting

Define the reweighted treated and control means:

$$\mu_1^{\text{IPW}} = \mathbb{E}[w_1(X)Y \mid D = 1],$$
$$\mu_0^{\text{IPW}} = \mathbb{E}[w_0(X)Y \mid D = 0].$$

Then

$$\text{ATE} = \mu_1^{\text{IPW}} - \mu_0^{\text{IPW}}.$$

Interpretation

We are doing a **pooled difference in means**, but on **weighted outcomes** that correct the X-imbalance.

Tech Example

Thompson sampling: known propensities from the assignment rule

Thompson Sampling Revisited: Can We Still Use IPW?

- In bandit experiments, assignment depends on history and context X.
- But for each decision, the system assigns treatment *stochastically* with a well-defined propensity $p(X) = \Phi(\mu/\sigma)$, where $X = (\mu, \sigma)$.
- If the logging system records $p(X)$, then we can treat this as a **conditionally randomized experiment**.

Question

Does the same formula:

$$\text{ATE} = \mathbb{E}[w_1(X)Y \mid D = 1] - \mathbb{E}[w_0(X)Y \mid D = 0] \qquad w_d(X) = \frac{\Pr(D = d)}{\Pr(D = d \mid X)}$$

still identify ATE?

Answer: yes (it only needs conditional ignorability + overlap).

General Result

ATE via Inverse Propensity / Horvitz-Thompson (HT) Transform

The "Naive" Difference Becomes Correct After Weighting

We want to prove that under conditional ignorability + overlap it is always the case that

$$\mu_d^{\text{IPW}} = \mathbb{E}[w_d(X)Y \mid D = d], \qquad w_d(X) = \frac{\Pr(D = d)}{\Pr(D = d \mid X)}$$

$$\text{ATE} = \mu_1^{\text{IPW}} - \mu_0^{\text{IPW}}.$$

For technical reasons,[1] it is convenient to re-express this as a single expectation:

$$\mu_d^{\text{IPW}} = \mathbb{E}[w_1(X)Y \mid D = d] = \mathbb{E}\left[w_d(X)Y\frac{\mathbb{I}\{D = d\}}{\Pr(D = d)}\right] = \mathbb{E}\left[Y\frac{\mathbb{I}\{D = d\}}{\Pr(D = d \mid X)}\right]$$

[1] which will be useful later when constructing confidence intervals

Main Theorem

Define the *propensity* $p(X) = \Pr(D = 1 \mid X)$ and the *Horvitz-Thompson (HT) weight*

$$H(D, X) = \frac{\mathbb{I}\{D = 1\}}{p(X)} - \frac{\mathbb{I}\{D = 0\}}{1 - p(X)}.$$

Under conditional ignorability + overlap:

$$\text{ATE} = \mathbb{E}\big[Y \cdot H(D, X)\big].$$

Moreover:

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[Y \frac{\mathbb{I}\{D = 1\}}{p(X)}\right], \qquad \mathbb{E}[Y(0)] = \mathbb{E}\left[Y \frac{\mathbb{I}\{D = 0\}}{1 - p(X)}\right].$$

Proof

Fix $d \in \{0, 1\}$. Consider the conditional expectation given X:

$$\begin{aligned}
\mathbb{E}\left[Y \frac{\mathbb{I}\{D = d\}}{\Pr(D = d \mid X)} \,\middle|\, X \right] &= \frac{\mathbb{E}[Y \cdot \mathbb{I}\{D = d\} \mid X]}{\Pr(D = d \mid X)} \\
&= \frac{\mathbb{E}[Y(d) \cdot \mathbb{I}\{D = d\} \mid X]}{\Pr(D = d \mid X)} \qquad \text{(consistency: } Y = Y(D)\text{)} \\
&= \frac{\mathbb{E}[Y(d) \mid X] \cdot \mathbb{E}[\mathbb{I}\{D = d\} \mid X]}{\Pr(D = d \mid X)} \quad \text{(since } Y(d) \perp\!\!\!\perp D \mid X\text{)} \\
&= \mathbb{E}[Y(d) \mid X].
\end{aligned}$$

Take expectation over X to get: $\mathbb{E}\left[Y \frac{\mathbb{I}\{D=d\}}{\Pr(D=d|X)} \right] = \mathbb{E}[Y(d)]$.

Operationalizing IPW When $p(X)$ Is Known

- Stratified trial: $p(X)$ comes from the trial protocol table.
- Thompson sampling: $p(X)$ comes from the assignment rule (and is logged).

Key robustness point

If $p(X)$ is known, then the ATE formula requires **no modeling of** $\mathbb{E}[Y \mid D, X]$.

We only need to average weighted outcomes.

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{\mathbb{I}\{D_i = 1\}}{p(X_i)} - \frac{\mathbb{I}\{D_i = 0\}}{1 - p(X_i)} \right).$$

Generalized Balance Check for known $p(X)$: $\mathbb{E}[H \mid X] = 0$

If the design / logging is correct, then

$$\mathbb{E}[H \mid X] = 0 \qquad\qquad H = \frac{D}{p(X)} - \frac{1-D}{1-p(X)}$$

- In an RCT: this reduces to "covariates should not predict treatment."
- Here: covariates should not predict the *HT weight* H.

Practical check: linear regression of H on functions of X and test for significance. Equivalent, to post-weight "balance" check:

$$0 = \mathbb{E}[H\,\phi(X)] = \underbrace{\mathbb{E}[\phi(X)w_1(X) \mid D=1] - \mathbb{E}[\phi(X)w_0(X) \mid D=0]}_{\text{A weighted balance check of characteristic } \phi(X)} \tag{1}$$

Observational Data

Estimate propensities + diagnose overlap

Observational Settings: Same Formula, Now $p(X)$ Must Be Estimated

Identification (still true)

$$\mathsf{ATE} = \mathbb{E}\left[Y\left(\frac{\mathbb{I}\{D=1\}}{p(X)} - \frac{\mathbb{I}\{D=0\}}{1-p(X)}\right)\right].$$

Operationalization

Estimate $p(X) = \Pr(D = 1 \mid X)$ from data, then plug-in.

Key point: estimating $p(X)$ is a **classification problem**.

Propensity Estimation = Classification

- Classical: logistic regression (interpretable; strong modeling assumptions).
- Flexible ML: random forests, gradient boosting, neural nets, etc.
- Practical note: calibration matters (probabilities should be meaningful).

Why errors matter

Because the weights divide by $\hat{p}(X)$ and $1 - \hat{p}(X)$: small errors can blow up when propensities are extreme.

Overlap Diagnostics: What Can Go Wrong?

- If $\hat{p}(X)$ is near 0 or 1 for many units, weights explode.
- Then the ATE estimate can have **high variance** (unstable).

Common checks

- Histogram / density of $\hat{p}(X)$ by D.
- Distribution of weights (max weight, tail behavior).

Two Common Fixes: Clipping vs Trimming

1) Clip propensities (stabilize denominators)

Replace $\hat{p}(X)$ with

$$\text{clip}(\hat{p}(X), \varepsilon, 1-\varepsilon) = \max\{\varepsilon, \min\{1-\varepsilon, \hat{p}(X)\}\}.$$

Tradeoff: introduces bias, reduces variance.

2) Trim the sample (change the estimand)

Drop units with extreme $\hat{p}(X)$ (e.g., outside $[\varepsilon, 1-\varepsilon]$).
Interpretation: estimates ATE for a restricted subpopulation where overlap holds.

ATT

reweight controls to look like treated (one-sided overlap)

ATT: Average Treatment Effect on the Treated

Definition

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

- In many applications, "treated" are participants/adopters; ATT is the effect on those who actually received the intervention.
- Today: express ATT as **treated mean minus a reweighted control mean**.

PrecISE Intuition for ATT: Only Reweight Controls

- ATT conditions on $D = 1$, so treated units already have the correct target distribution: $\Pr(X \mid D = 1)$.
- We need to reweight controls so they mimic $\Pr(X \mid D = 1)$.

$$\text{ATT (g-formula)} = \sum_x \underbrace{\Pr(X{=}x \mid D = 1)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in treated}}} \bar{Y}_{1,x} - \underbrace{\Pr(X{=}x \mid D = 1)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in treated}}} \bar{Y}_{0,x}$$

$$\text{Naive Difference in Means} = \sum_x \underbrace{\Pr(X{=}x \mid D = 1)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in treated}}} \bar{Y}_{1,x} - \underbrace{\Pr(X{=}x \mid D = 0)}_{\substack{\text{Cond. Mass} \\ \text{of } x \text{ in control}}} \bar{Y}_{0,x}$$

Desired weights for controls to make the latter emulate the former:

$$w_0(X) = \frac{\Pr(X \mid D = 1)}{\Pr(X \mid D = 0)}.$$

PrecISE Intuition for ATT: Only Reweight Controls

- ATT conditions on $D = 1$, so treated units already have the correct target distribution: $\Pr(X \mid D = 1)$.
- We need to reweight controls so they mimic $\Pr(X \mid D = 1)$.

Desired weights for controls:

$$w_0(X) = \frac{\Pr(X \mid D = 1)}{\Pr(X \mid D = 0)}.$$

Using Bayes' rule:

$$w_0(X) = \frac{p(X)\Pr(D = 0)}{(1 - p(X))\Pr(D = 1)}.$$

Interpretation: an (normalized) odds-ratio weight.

ATT Identification Formula (One-Sided Ignorability + One-Sided Overlap)

Theorem. Under one-sided ignorability + one-sided overlap:[2]

$$\text{ATT} = \mathbb{E}[Y \mid D = 1] \; - \; \mathbb{E}\big[w_0(X)\, Y \mid D = 0\big],$$

where

$$w_0(X) = \frac{p(X)\Pr(D = 0)}{(1 - p(X))\Pr(D = 1)}.$$

Note: we only ever divide by $1 - p(X)$, not by $p(X)$ (hence only one-sided overlap required; $p(X)$ bounded away from 1).

Note: Can be written as a single expectation:

$$\text{ATT} = \mathbb{E}[Y \cdot H_1(D, X)], \qquad H_1(D, X) = \frac{\mathbb{I}\{D = 1\}}{\Pr(D = 1)} - \frac{\mathbb{I}\{D = 0\}}{\Pr(D = 1)} \frac{p(X)}{(1 - p(X))}$$

[2] See Appendix 5.C in Causal ML book for proof; similar to ATE proof.

Operationalizing ATT (Known or Estimated Propensities)

- If $p(X)$ is known (conditionallly randomized design): plug into $w_0(X)$ and use estimate $\hat{\pi} = \frac{1}{n}\sum_i D_i$ for $\pi = \Pr(D = 1)$.
- If observational: estimate $\hat{p}(X)$ by classification, estimate $\hat{\pi} = \frac{1}{n}\sum_i D_i$.

Sample implementation:

$$\widehat{\mathrm{ATT}} = \underbrace{\frac{1}{n_1}\sum_{i:D_i=1} Y_i}_{\text{treated mean}} - \underbrace{\frac{1}{n_0}\sum_{i:D_i=0} \hat{w}_0(X_i)\, Y_i}_{\text{reweighted controls}}, \qquad \hat{w}_0(X) = \underbrace{\frac{\hat{p}(X)}{1-\hat{p}(X)}}_{\substack{\text{conditional} \\ \text{odds ratio}}} \cdot \frac{1-\hat{\pi}}{\hat{\pi}}$$

Note: Only extreme propensities near 1 create problems (one-sided overlap). Either:

- perform one-sided clipping $\mathrm{clip}(\hat{p}(X), 0, 1-\epsilon)$
- trim by throwing away samples for which $\hat{p}(X) \in [1-\epsilon, 1]$.

Wrap-up: Four Takeaways

1. **If treated/control are not representative, reweight.**
2. Under ignorability + overlap, **ATE** $= \mathbb{E}[Y \cdot H]$ (Horvitz–Thompson).
3. If propensities are **known by design**, we can estimate effects **without outcome modeling**.
4. In observational data, we must **estimate propensities** and **check overlap** (clipping/trimming if needed).

Bridge: next lectures will combine both lenses (conditioning + propensity) for stability and efficiency.

In-class Activity

Lalonde: get the best ATT you can

In-Class Activity: The Lalonde ATT Challenge

Estimate the ATT of job training in the observational Lalonde dataset using *identification by conditioning* or *by propensity*.

- Work in groups of 3–4.
- At the end report out your final estimate + two-three sentences on how you trained your model.
- **Closest "valid" point estimate to experimental benchmark wins! Choose a song to play at beginning of next class.**

You may use

- Lecture 2: identification by conditioning (outcome regression / g-formula),
- Lecture 3: identification by propensity (IPW / reweighting),
- any modeling choices you like for the outcome regression(s) or for the propensity classification (linear, trees, boosting, etc.).

Lalonde Notebook

Results Report