
Review of Causal Inference

Lecture 5 ■ POLSCI 689L

Jiawei Fu

Department of Political Science ■ Duke University

Overview

1. **Randomized experiments:** finite-population and superpopulation inference
2. **Observational studies:** selection bias, ignorability, and overlap
3. **Outcome regression:** standardize fitted potential-outcome means
4. **Propensity-score methods:** regression adjustment, Horvitz–Thompson, and Hájek weighting
5. **AIPW:** combine both models and study weak and strong double robustness
6. **Inference:** sandwich standard errors and the nonparametric bootstrap

Potential outcomes

For each unit i , define two potential outcomes:

$$Y_i(1) \quad \text{under treatment,} \quad Y_i(0) \quad \text{under control.}$$

Consistency and no interference

The observed outcome equals the potential outcome under the treatment received:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) = Y_i(Z_i).$$

This notation also assumes that unit i 's outcome does not depend on other units' assignments.

- $Z_i \in \{0, 1\}$ records treatment assignment.
- Pretreatment covariates X_i will become central in observational studies.

Wager (2026), Chapter 1, pp. 3–5; Ding (2023), Chapter 2, pp. 15–20.

The causal target and the missing counterfactual

Unit-level contrast

$$Y_i(1) - Y_i(0)$$

Only one term can be observed for unit i .

Average treatment effect

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

Its precise meaning depends on whether the units are fixed or sampled from a population.

Fundamental problem

No dataset reveals both potential outcomes for the same unit. Causal inference supplies a credible comparison group for the missing outcomes.

Wager (2026), pp. 3–4; Ding (2023), pp. 16–18.

Two frameworks for randomized experiments

	Finite population	Superpopulation
Units	The n study units are fixed.	Units are IID draws from a population.
Potential outcomes	$\{Y_i(1), Y_i(0)\}_{i=1}^n$ are fixed.	$\{Y_i(1), Y_i(0), X_i\}$ are random.
Assignment	Only $Z = (Z_1, \dots, Z_n)$ is random.	Sampling and randomized assignment are random.
Target	$\tau_F = \frac{1}{n} \sum_i \{Y_i(1) - Y_i(0)\}$	$\tau_S = \mathbb{E}\{Y(1) - Y(0)\}$

Key idea

The same difference in means can estimate either target, but the source and interpretation of uncertainty differ.

Wager (2026), pp. 4–7 and 15–16; Ding (2023), pp. 43–49 and 129–131.

Finite-population randomization inference

Under complete randomization, exactly n_1 units are treated and $n_0 = n - n_1$ are controls. The difference in means is

$$\hat{\tau}_{\text{DM}} = \frac{1}{n_1} \sum_{i:Z_i=1} Y_i - \frac{1}{n_0} \sum_{i:Z_i=0} Y_i.$$

Holding the units and all potential outcomes fixed,

$$\mathbb{E}_Z(\hat{\tau}_{\text{DM}}) = \tau_F.$$

Define $\tau_i = Y_i(1) - Y_i(0)$ and the finite-population variances

$$S_z^2 = \frac{1}{n-1} \sum_i \{Y_i(z) - \bar{Y}(z)\}^2, \quad S_\tau^2 = \frac{1}{n-1} \sum_i \{\tau_i - \tau_F\}^2.$$

Then the exact randomization variance is

$$\text{Var}_Z(\hat{\tau}_{\text{DM}}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n}.$$

Ding (2023), Chapter 4, pp. 43–47; Wager (2026), pp. 4–5.

Estimating the finite-population variance

Let $\hat{S}_z^2$ be the usual sample variance of observed outcomes in arm z . Because both potential outcomes are never observed together, S_τ^2 is not identified.

$$\hat{V}_N = \frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_0^2}{n_0}$$

is therefore the standard Neyman variance estimator. It is conservative:

$$\mathbb{E}_Z(\hat{V}_N) - \text{Var}_Z(\hat{\tau}_{DM}) = \frac{S_\tau^2}{n} \geq 0.$$

A large-sample $(1 - \alpha)$ confidence interval is

$$\hat{\tau}_{DM} \pm z_{1-\alpha/2} \sqrt{\hat{V}_N}.$$

Regression equivalence. The HC2 robust variance from OLS of Y on an intercept and Z equals $\hat{V}_N$; conservativeness disappears under constant individual effects.

Ding (2023), pp. 47–49; Wager (2026), pp. 5–7.

Superpopulation inference and the bridge ahead

Now suppose $\{Y_i(1), Y_i(0), X_i\}_{i=1}^n$ are IID draws and treatment is randomized:

$$Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0), X_i\}.$$

The target is $\tau_S = \mathbb{E}\{Y(1) - Y(0)\}$, and

$$\sqrt{n}(\hat{\tau}_{\text{DM}} - \tau_S) \rightsquigarrow N(0, V_{\text{DM}}), \quad V_{\text{DM}} = \frac{\text{Var}\{Y(1)\}}{\pi} + \frac{\text{Var}\{Y(0)\}}{1 - \pi},$$

where $\pi = \mathbb{P}(Z = 1)$. Within-arm sample variances consistently estimate this variance.

Framework for the rest of the lecture

For observational studies, we use the IID superpopulation framework. The challenge is that treatment is no longer randomized.

Wager (2026), pp. 5–7 and 15–17; Ding (2023), pp. 129–131.

The observational-data problem

Without random assignment, treatment status may reveal something about the potential outcomes:

$$Z \not\perp \{Y(1), Y(0)\}.$$

Add and subtract the treated units' untreated potential-outcome mean:

$$\begin{aligned}\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) &= \underbrace{\mathbb{E}\{Y(1) - Y(0) \mid Z = 1\}}_{\text{effect among treated units}} \\ &\quad + \underbrace{\mathbb{E}\{Y(0) \mid Z = 1\} - \mathbb{E}\{Y(0) \mid Z = 0\}}_{\text{selection bias}}.\end{aligned}$$

Key idea

A raw difference in outcomes mixes a causal effect with preexisting differences between groups.

Ding (2023), Chapter 10, pp. 141–142.

Observational studies use a superpopulation

Treat $O_i = (X_i, Z_i, Y_i)$ as IID draws from a population, with target $\tau = \tau_S = \mathbb{E}\{Y(1) - Y(0)\}$.
Write $e(x) = \mathbb{P}(Z = 1 \mid X = x)$.

Consistency

$Y = Y(Z)$ and the treatment levels have well-defined meanings.

Conditional exchangeability

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid X.$$

Given X , assignment is as good as random.

Overlap

$$0 < e(X) < 1.$$

Both treatments occur at relevant covariate values.

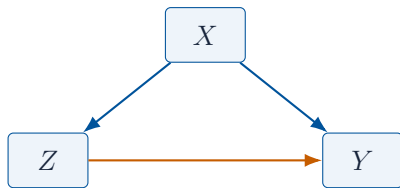
Identification claim

Together, these assumptions connect unobserved potential outcomes to the observed distribution of (X, Z, Y) .

Wager (2026), pp. 20–24; Ding (2023), pp. 142–146 and 158–160.

The role of pretreatment covariates

- X should contain pretreatment variables needed to block common causes of treatment and outcome.
- Scientific knowledge determines whether the covariate set makes ignorability plausible.
- Observed balance cannot establish the absence of unmeasured confounding.
- Variables caused by treatment should not enter the adjustment set for the total effect.



Ding (2023), pp. 145–146 and Chapter 16, pp. 219–228.

Overlap and empirical support

Good overlap

At a given x , both groups occur. The two conditional outcome means can be learned locally.

If $e(x) = 1$, the data do not identify

$$\mathbb{E}\{Y(0) \mid X = x\}.$$

The potential outcome still exists conceptually. The missing comparison cannot be learned from these data.

Key idea

Overlap is an identification condition and a practical diagnostic.

Wager (2026), pp. 19 and 24; Ding (2023), pp. 158–161 and 267–269.

Three representations of the same causal target

Under consistency, conditional exchangeability, and overlap:

Outcome regression $\tau = \mathbb{E}\{\mu_1(X) - \mu_0(X)\}$

Propensity weighting $\tau = \mathbb{E}\left\{\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\}$

Augmented weighting $\tau = \mathbb{E}\left[\mu_1(X) - \mu_0(X) + \frac{Z\{Y - \mu_1(X)\}}{e(X)} - \frac{(1-Z)\{Y - \mu_0(X)\}}{1-e(X)}\right]$

Road map

The estimators model different observed-data relationships. Their causal interpretation rests on the same identification assumptions.

Ding (2023), pp. 169–172; Wager (2026), pp. 22–23 and 29–31.

One inference engine for all three estimators

Let $O_i = (X_i, Z_i, Y_i)$ and collect the target and nuisance parameters as $\theta = (\tau, \eta^\top)^\top$. A joint estimator solves a vector of estimating equations:

$$\mathbb{P}_n M(O; \hat{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n M(O_i; \hat{\theta}) = 0.$$

Target moment

The first equation defines $\hat{\tau}$.

Nuisance moments

The remaining equations fit outcome and/or propensity models.

Joint uncertainty

Solving and analyzing the stack propagates every fitted-model contribution.

Key idea

OR, IPW, and AIPW use different moments, but the same large-sample inference machinery.

Tsiatis et al. (2020), Chapter 2, pp. 32–35.

Why the sandwich variance works

Under i.i.d. sampling, finite second moments, smoothness, and a nonsingular Jacobian, let

$$A = -\mathbb{E}\left\{\frac{\partial M(O; \theta_0)}{\partial \theta^\top}\right\}, \quad B = \mathbb{E}\{M(O; \theta_0)M(O; \theta_0)^\top\}.$$

A first-order expansion of $\mathbb{P}_n M(O; \hat{\theta}) = 0$ gives

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n M(O_i; \theta_0) - A\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

Therefore, by the central limit theorem,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N(0, A^{-1}BA^{-\top}).$$

Sample sandwich and a Wald interval

$\hat{A} = -\mathbb{P}_n\{\partial M(O; \hat{\theta})/\partial \theta^\top\}$, $\hat{B} = \mathbb{P}_n\{M(O; \hat{\theta})M(O; \hat{\theta})^\top\}$, and $\hat{\Sigma} = \hat{A}^{-1}\hat{B}\hat{A}^{-\top}$. Hence $\text{SE}(\hat{\tau}) = \sqrt{\hat{\Sigma}_{11}/n}$ and an approximate 95% interval is $\hat{\tau} \pm 1.96 \text{SE}(\hat{\tau})$.

Tsiatis et al. (2020), pp. 32–35, especially equations (2.33)–(2.42).

Nonparametric bootstrap inference

Treat each observed row $O_i = (X_i, Z_i, Y_i)$ as one sampling unit. For $b = 1, \dots, B$:

1. Draw n rows with replacement from $\{O_1, \dots, O_n\}$.
2. Apply the complete estimation procedure to the resampled data and obtain $\hat{\tau}^{*(b)}$.

The spread of the bootstrap replicates estimates sampling uncertainty:

$$\bar{\tau}^* = \frac{1}{B} \sum_{b=1}^B \hat{\tau}^{*(b)}, \quad \hat{V}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^B \{\hat{\tau}^{*(b)} - \bar{\tau}^*\}^2.$$

Thus $\text{SE}_{\text{boot}} = \sqrt{\hat{V}_{\text{boot}}}$ and a Normal-approximation interval is $\hat{\tau} \pm z_{1-\alpha/2} \text{SE}_{\text{boot}}$.

OR Refit the outcome model and restandardize over the resampled rows.

IPW Refit the propensity model and reconstruct all weights.

AIPW Refit both nuisance models and recompute every residual correction.

Critical detail: Resample whole rows and refit every model. The bootstrap measures sampling variation, not model bias.

Ding (2023), pp. 148, 161–162, 172–175, and Appendix A.6, pp. 417–418.

Identification through conditional outcome means

Define the observed conditional outcome means

$$\mu_z(x) = \mathbb{E}(Y \mid Z = z, X = x), \quad z \in \{0, 1\}.$$

Then, for either treatment level,

$$\begin{aligned} \mathbb{E}\{Y(z)\} &= \mathbb{E}[\mathbb{E}\{Y(z) \mid X\}] && \text{law of iterated expectations} \\ &= \mathbb{E}[\mathbb{E}\{Y(z) \mid Z = z, X\}] && \text{conditional exchangeability} \\ &= \mathbb{E}[\mathbb{E}(Y \mid Z = z, X)] && \text{consistency} \\ &= \mathbb{E}\{\mu_z(X)\}. \end{aligned}$$

Therefore,

$$\tau = \mathbb{E}\{\mu_1(X) - \mu_0(X)\}.$$

Ding (2023), pp. 143–148; Wager (2026), p. 30.

Outcome regression as standardization

1. Estimate the two conditional mean functions:

$$\hat{\mu}_1(x) \approx \mathbb{E}(Y \mid Z = 1, X = x), \quad \hat{\mu}_0(x) \approx \mathbb{E}(Y \mid Z = 0, X = x).$$

2. For every observed X_i , predict the mean outcome under both treatment levels.
3. Average the predicted contrasts over the empirical covariate distribution:

$$\hat{\tau}_{\text{OR}} = \frac{1}{n} \sum_{i=1}^n \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}.$$

Key idea

Standardization compares treatment conditions at a common distribution of X .

Ding (2023), pp. 147–149; Wager (2026), p. 30.

Outcome-model choices

One working model

$$\mathbb{E}(Y \mid Z, X) = m(Z, X; \beta).$$

Include treatment-covariate interactions when the response surface may differ across treatment groups.

For the interacted linear model

$$\mathbb{E}(Y \mid Z, X) = \beta_0 + \beta_z Z + \beta_x^\top X + Z \beta_{zx}^\top X,$$

standardization gives

$$\hat{\tau}_{\text{OR}} = \hat{\beta}_z + \hat{\beta}_{zx}^\top \overline{X}.$$

Ding (2023), pp. 147–149.

Two response surfaces

Fit $\hat{\mu}_1(x)$ with treated observations and $\hat{\mu}_0(x)$ with control observations. The two fits may use flexible regression methods.

Regression coefficients and standardized effects

Additive linear model

If

$$m(z, x; \beta) = \beta_0 + \beta_z z + \beta_x^\top x$$

is correct, then the fitted contrast is β_z for every x , so $\tau = \beta_z$.

Interactions or nonlinear links

Standardize the fitted contrast over X . For a binary outcome with fitted risks $\hat{p}_z(x)$,

$$\hat{\tau}_{\text{OR}} = \frac{1}{n} \sum_i \{\hat{p}_1(X_i) - \hat{p}_0(X_i)\}.$$

In logistic regression, $\hat{\beta}_z$ is a conditional log odds ratio, not the average risk difference above.

Key idea

Form two treatment-specific predictions and average their difference on the desired outcome scale.

Ding (2023), pp. 147–149; Tsiatis et al. (2020), p. 31.

Consistency of outcome regression

Let $\Delta(x; \beta) = m(1, x; \beta) - m(0, x; \beta)$. Suppose the conditional-mean model is correct at β_0 and $\hat{\beta} \xrightarrow{p} \beta_0$. Then

$$\begin{aligned}\hat{\tau}_{\text{OR}} - \tau &= (\mathbb{P}_n - \mathbb{E})\Delta(X; \beta_0) \\ &\quad + \mathbb{P}_n\{\Delta(X; \hat{\beta}) - \Delta(X; \beta_0)\} \\ &\quad + \underbrace{\mathbb{E}\Delta(X; \beta_0) - \tau}_{0 \text{ by identification}}.\end{aligned}$$

The law of large numbers makes the first term vanish. Consistency of $\hat{\beta}$ and continuity of Δ make the second term vanish.

- Smooth parametric fitting makes both terms contribute at the $n^{-1/2}$ scale, which the joint sandwich captures.
- Under misspecification, the last line need not equal zero. Neither a sandwich variance nor a bootstrap removes that bias.

Ding (2023), pp. 147–149; Tsiatis et al. (2020), pp. 31 and 37–38.

Outcome-regression inference

For outcome model $m(z, x; \beta)$, let

$$s_\beta(O; \beta) = \dot{m}_\beta(Z, X; \beta)\{Y - m(Z, X; \beta)\}$$

be the least-squares estimating function. Then $\hat{\tau}_{\text{OR}}$ and $\hat{\beta}$ solve

$$\mathbb{P}_n \begin{pmatrix} m(1, X; \beta) - m(0, X; \beta) - \tau \\ s_\beta(O; \beta) \end{pmatrix} = 0.$$

- If the outcome model is correct, $\mathbb{E}\{s_\beta(O; \beta_0) \mid X, Z\} = 0$.
- Identification through conditional outcome means makes the first moment mean zero at τ .
- The sandwich propagates regression-fit uncertainty and empirical averaging over X .

Common mistake

With interactions, the treatment coefficient's standard error alone is not the standard error of $\hat{\tau}_{\text{OR}}$.

Tsiatis et al. (2020), pp. 37–38, equations (2.50)–(2.53); Ding (2023), pp. 147–149.

Where outcome-regression uncertainty comes from

Let $\Delta_i(\beta) = m(1, X_i; \beta) - m(0, X_i; \beta)$ and define

$$\widehat{V}_X = \mathbb{P}_n \{ \Delta_i(\widehat{\beta}) - \widehat{\tau}_{\text{OR}} \}^2,$$

$$\widehat{C} = \mathbb{P}_n \{ \dot{m}_\beta(1, X_i; \widehat{\beta}) - \dot{m}_\beta(0, X_i; \widehat{\beta}) \},$$

$$\widehat{D} = \mathbb{P}_n \{ \dot{m}_\beta(Z_i, X_i; \widehat{\beta}) \dot{m}_\beta(Z_i, X_i; \widehat{\beta})^\top \}.$$

Under homoskedastic errors, block-matrix algebra gives

$$\widehat{\text{Var}}(\widehat{\tau}_{\text{OR}}) = \frac{1}{n} \left\{ \widehat{V}_X + \widehat{\sigma}^2 \widehat{C}^\top \widehat{D}^{-1} \widehat{C} \right\}.$$

Across-unit component

$\widehat{V}_X$ measures variation in fitted treatment contrasts across the target covariate distribution.

Tsiatis et al. (2020), p. 38; the vanishing cross term follows from $\mathbb{E}(s_\beta \mid X, Z) = 0$.

Model-fit component

The second term propagates uncertainty in $\widehat{\beta}$.
Use the full sandwich under heteroskedasticity.

Outcome-regression vulnerabilities

Strength

Prognostic covariates can improve precision and produce direct estimates of both conditional means.

Main vulnerabilities

- Misspecification can bias counterfactual predictions.
- Limited support forces extrapolation.

Diagnostic question

Do data support both fitted response surfaces over the covariate distribution used for averaging? Good prediction of observed outcomes alone is insufficient.

Wager (2026), pp. 26–27 and 30; Ding (2023), pp. 147–149.

Definition 11.1: the propensity score

Ding begins with the general treatment-assignment probability

$$e\{X, Y(1), Y(0)\} = \mathbb{P}\{Z = 1 \mid X, Y(1), Y(0)\}.$$

Without ignorability, this probability may depend on both potential outcomes.

Under strong ignorability,

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X,$$

so

$$\begin{aligned} e\{X, Y(1), Y(0)\} &= \mathbb{P}\{Z = 1 \mid X, Y(1), Y(0)\} \\ &= \mathbb{P}(Z = 1 \mid X) \equiv e(X). \end{aligned}$$

Key idea

Under strong ignorability, the propensity score is the conditional probability of treatment given the observed covariates.

Ding (2023), Definition 11.1, pp. 153–154.

Theorem 11.1: propensity-score reduction

Theorem 11.1. If treatment is ignorable given X , then it is also ignorable given the propensity score:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X \implies \boxed{Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X).}$$

- Ignorability may require conditioning on a high-dimensional vector X .
- The theorem replaces X with the scalar $e(X) \in [0, 1]$ while preserving ignorability.
- The result provides dimension reduction. It does not make ignorability or overlap testable from the observed data.

Ding (2023), Theorem 11.1, p. 154.

Proof of Theorem 11.1

Let $U = \{Y(1), Y(0)\}$ and $S = e(X)$. Conditional independence requires $\mathbb{P}(Z = 1 \mid U, S) = \mathbb{P}(Z = 1 \mid S)$.

Condition on U and S

$$\begin{aligned}\mathbb{P}(Z = 1 \mid U, S) &= \mathbb{E}(Z \mid U, S) \\ &= \mathbb{E}\{\mathbb{E}(Z \mid U, S, X) \mid U, S\} \\ &= \mathbb{E}\{\mathbb{E}(Z \mid U, X) \mid U, S\} \\ &= \mathbb{E}\{\mathbb{E}(Z \mid X) \mid U, S\} \\ &= \mathbb{E}\{e(X) \mid U, S\} = S.\end{aligned}$$

Condition on S

$$\begin{aligned}\mathbb{P}(Z = 1 \mid S) &= \mathbb{E}(Z \mid S) \\ &= \mathbb{E}\{\mathbb{E}(Z \mid S, X) \mid S\} \\ &= \mathbb{E}\{\mathbb{E}(Z \mid X) \mid S\} \\ &= \mathbb{E}\{e(X) \mid S\} = S.\end{aligned}$$

The fourth line uses strong ignorability.

Thus both conditional probabilities equal S . Since Z is binary,

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid e(X).$$

Ding (2023), proof of Theorem 11.1, pp. 154–155.

Theorem 11.3: the balancing property

The propensity score also satisfies

$$Z \perp\!\!\!\perp X \mid e(X).$$

For binary treatment,

$$\begin{aligned}\mathbb{P}\{Z = 1 \mid X, e(X)\} &= \mathbb{P}(Z = 1 \mid X) = e(X), \\ \mathbb{P}\{Z = 1 \mid e(X)\} &= \mathbb{E}\{e(X) \mid e(X)\} = e(X).\end{aligned}$$

- **Theorem 11.1:** treatment is independent of the potential outcomes given $e(X)$. It requires ignorability given X .
- **Theorem 11.3:** treatment is independent of the covariates given $e(X)$. This balancing result does not require ignorability.

Ding (2023), Theorem 11.3 and proof, pp. 163–164.

Estimating the propensity score

A common working model is logistic regression:

$$\text{logit}\{e(x; \gamma)\} = \gamma_0 + \gamma^\top x, \quad \hat{e}_i = e(X_i; \hat{\gamma}).$$

1. Choose pretreatment covariates using subject-matter knowledge.
2. Fit the treatment model without using the outcome.
3. Use $\hat{e}_i$ for score regression, stratification, matching, or inverse weighting.
4. Examine overlap and covariate balance after adjustment.

What to optimize

The goal is covariate balance and credible adjustment, not maximal prediction of treatment. For inference, treat $\hat{e}$ as estimated: stack its score equations or refit it in every bootstrap sample.

Ding (2023), Chapter 11, pp. 153–164; Tsiatis et al. (2020), pp. 39–44; Wager (2026), pp. 20–25.

The inverse-weighting identity

For the treated potential-outcome mean,

$$\begin{aligned}\mathbb{E}\left\{\frac{ZY}{e(X)}\right\} &= \mathbb{E}\left[\mathbb{E}\left\{\frac{ZY(1)}{e(X)} \mid X\right\}\right] && \text{consistency} \\ &= \mathbb{E}\left[\frac{\mathbb{E}(Z \mid X)\mathbb{E}\{Y(1) \mid X\}}{e(X)}\right] && \text{exchangeability} \\ &= \mathbb{E}\{Y(1)\}. && \text{definition of } e(X)\end{aligned}$$

Similarly,

$$\mathbb{E}\left\{\frac{(1-Z)Y}{1-e(X)}\right\} = \mathbb{E}\{Y(0)\}.$$

Each observed outcome represents units with the same X whose treatment realization differs.

Wager (2026), pp. 22–23; Ding (2023), pp. 158–160.

Horvitz–Thompson and Hájek IPW estimators

Write $\hat{e}_i = \hat{e}(X_i)$. Both estimators use inverse probabilities, but they normalize the weighted outcomes differently.

Horvitz–Thompson

$$\begin{aligned}\hat{\mu}_{1,\text{HT}} &= \mathbb{P}_n \frac{ZY}{\hat{e}}, \\ \hat{\mu}_{0,\text{HT}} &= \mathbb{P}_n \frac{(1-Z)Y}{1-\hat{e}}, \\ \hat{\tau}_{\text{HT}} &= \hat{\mu}_{1,\text{HT}} - \hat{\mu}_{0,\text{HT}}.\end{aligned}$$

Each weighted total uses the fixed denominator n .

Hájek, or self-normalized IPW

$$\begin{aligned}\hat{\mu}_{1,\text{H}} &= \frac{\mathbb{P}_n ZY/\hat{e}}{\mathbb{P}_n Z/\hat{e}}, \\ \hat{\mu}_{0,\text{H}} &= \frac{\mathbb{P}_n (1-Z)Y/(1-\hat{e})}{\mathbb{P}_n (1-Z)/(1-\hat{e})}, \\ \hat{\tau}_{\text{H}} &= \hat{\mu}_{1,\text{H}} - \hat{\mu}_{0,\text{H}}.\end{aligned}$$

Each treatment group's weights sum to one.

Ding (2023), p. 160; Wager (2026), p. 160, equation (12.9).

Why Hájek normalization helps

Define

$$\hat{a}_1 = \mathbb{P}_n \frac{Z}{\hat{e}(X)}, \quad \hat{a}_0 = \mathbb{P}_n \frac{1 - Z}{1 - \hat{e}(X)}.$$

If every outcome changes to $Y_i + c$, then

$$\hat{\tau}_{\text{HT}}(Y + c) = \hat{\tau}_{\text{HT}}(Y) + c(\hat{a}_1 - \hat{a}_0),$$

which generally changes the finite-sample estimate. For Hájek,

$$\hat{\tau}_{\text{H}}(Y + c) = \{\hat{\mu}_{1,\text{H}}(Y) + c\} - \{\hat{\mu}_{0,\text{H}}(Y) + c\} = \hat{\tau}_{\text{H}}(Y).$$

With the true propensity score, the law of large numbers also gives

$$\frac{\mathbb{P}_n ZY/e(X)}{\mathbb{P}_n Z/e(X)} \xrightarrow{p} \frac{\mathbb{E}\{Y(1)\}}{1} = \mathbb{E}\{Y(1)\},$$

and analogously for control.

Key idea

Hájek is consistent and location invariant. It often stabilizes finite samples, although it need not always have smaller variance.

Ding (2023), Proposition 11.1, pp. 160–162; Wager (2026), pp. 160–162.

Horvitz–Thompson inference with a fitted score

First fit a logistic propensity model,

$$e_\gamma(x) = \text{expit}(\gamma^\top x), \quad \mathbb{P}_n X\{Z - e_\gamma(X)\} = 0.$$

Then compute

$$\hat{\tau}_{\text{HT}} = \mathbb{P}_n H(O; e_\gamma), \quad H(O; e_\gamma) = \frac{ZY}{e_\gamma(X)} - \frac{(1-Z)Y}{1 - e_\gamma(X)}.$$

Both estimates jointly solve the stacked equations

$$\mathbb{P}_n M_{\text{HT}}(O; \tau, \gamma) = 0, \quad M_{\text{HT}}(O; \tau, \gamma) = \begin{pmatrix} H(O; e_\gamma) - \tau \\ X\{Z - e_\gamma(X)\} \end{pmatrix}.$$

Apply the joint sandwich and use its $(1, 1)$ element for $\hat{\tau}_{\text{HT}}$.

Implementation

A whole-row bootstrap must refit the logistic model and reconstruct every weight in each resample.

Tsiatis et al. (2020), pp. 43–44, equations (2.64)–(2.65); Ding (2023), pp. 161–162.

Hájek as weighted least squares

Use weights $w_i = Z_i/\hat{e}_i + (1 - Z_i)/(1 - \hat{e}_i)$. The coefficient $\hat{\beta}$ from

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \sum_i w_i (Y_i - \alpha - \beta Z_i)^2$$

equals $\hat{\tau}_H$. Because Z is binary, the normal equations set

$$\hat{\alpha} = \hat{\mu}_{0,H}, \quad \hat{\alpha} + \hat{\beta} = \hat{\mu}_{1,H}.$$

Therefore $\hat{\beta} = \hat{\mu}_{1,H} - \hat{\mu}_{0,H} = \hat{\tau}_H$.

Practical use

Standard regression software can compute the Hájek point estimate. Its default WLS standard error is not valid when the propensity score is estimated.

Ding (2023), Proposition 14.1, pp. 196–197.

Hájek inference

With known $e(X)$, the influence function is

$$\phi_H(O) = \frac{Z}{e(X)}\{Y - \mu_1\} - \frac{1 - Z}{1 - e(X)}\{Y - \mu_0\}.$$

If $e(X; \gamma)$ is fitted, stack the two weighted-residual equations

$$\mathbb{P}_n \frac{Z}{e_\gamma(X)}(Y - \mu_1) = 0, \quad \mathbb{P}_n \frac{1 - Z}{1 - e_\gamma(X)}(Y - \mu_0) = 0$$

with the propensity score equation, then apply the sandwich to $\tau = \mu_1 - \mu_0$.

Inference warning

Use the joint sandwich or refit the propensity model and both weighted means in every whole-row bootstrap sample.

Ding (2023), Proposition 14.1, pp. 196–197; Tsiatis et al. (2020), pp. 32–35.

Why fitting the propensity can improve precision

With a correctly specified maximum-likelihood propensity model, let

$$\phi^*(O) = H(O; e) - \tau, \quad s_\gamma(O) = X\{Z - e(X)\},$$

and define $I = \mathbb{E}(s_\gamma s_\gamma^\top)$ and $c = \mathbb{E}(s_\gamma \phi^*)$. The influence function with a fitted propensity score is

$$\phi_{\text{fit}}(O) = \phi^*(O) - c^\top I^{-1} s_\gamma(O).$$

The second term removes the linear projection of the oracle influence function onto the score:

$$\begin{aligned} \text{Var}(\phi_{\text{fit}}) &= \text{Var}(\phi^*) - 2c^\top I^{-1} c + c^\top I^{-1} I I^{-1} c \\ &= \text{Var}(\phi^*) - c^\top I^{-1} c \leq \text{Var}(\phi^*). \end{aligned}$$

Scope of the claim

A correct likelihood fit can improve raw HT by imposing sample balance. The result compares plug-in forms and does not apply to arbitrary fitted scores.

Tsiatis et al. (2020), pp. 44–46; Wager (2026), pp. 25 and 28.

Example: known versus fitted propensity

Suppose treatment is assigned independently with $Z_i \sim \text{Bernoulli}(1/2)$. Write

$$\mu_1 = \mathbb{E}\{Y_i(1)\} = \mathbb{E}(Y_i \mid Z_i = 1), \quad \sigma_1^2 = \text{Var}\{Y_i(1)\} = \text{Var}(Y_i \mid Z_i = 1).$$

The known propensity $e = 1/2$ gives the Horvitz–Thompson estimator

$$\hat{\mu}_1^{\text{known}} = \mathbb{P}_n \frac{Z_i Y_i}{1/2} = \frac{2}{n} \sum_{i=1}^n Z_i Y_i.$$

Alternatively, fit the correctly specified intercept-only propensity model:

$$e(\gamma) = \gamma, \quad \hat{\gamma} = \mathbb{P}_n Z_i = \bar{Z}.$$

Plugging in $\hat{\gamma}$ yields

$$\hat{\mu}_1^{\text{fitted}} = \mathbb{P}_n \frac{Z_i Y_i}{\hat{\gamma}} = \frac{\sum_i Z_i Y_i}{\sum_i Z_i} = \bar{Y}_1.$$

Question

Both are consistent. Why can fitting a known probability reduce variance?

Tsiatis et al. (2020), pp. 45–46, equations (2.66)–(2.67); adapted from the earlier course slides.

Known propensity: the extra variance term

First center and scale the known-propensity estimator:

$$\sqrt{n}(\hat{\mu}_1^{\text{known}} - \mu_1) = \frac{2}{\sqrt{n}} \sum_{i=1}^n \left(Z_i Y_i - \frac{\mu_1}{2} \right).$$

Because $\mathbb{E}(Z_i Y_i) = \mu_1/2$, the summands have mean zero. By the law of total variance,

$$\begin{aligned} \text{Var}(Z_i Y_i) &= \mathbb{E}\{\text{Var}(Z_i Y_i \mid Z_i)\} + \text{Var}\{\mathbb{E}(Z_i Y_i \mid Z_i)\} \\ &= \mathbb{E}(Z_i \sigma_1^2) + \text{Var}(Z_i \mu_1) \\ &= \frac{\sigma_1^2}{2} + \frac{\mu_1^2}{4}. \end{aligned}$$

Therefore, the central limit theorem gives

$$\sqrt{n}(\hat{\mu}_1^{\text{known}} - \mu_1) \rightsquigarrow N(0, 4 \text{Var}(Z_i Y_i)) = N(0, 2\sigma_1^2 + \mu_1^2).$$

Source of the penalty

The fixed denominator $n/2$ leaves random treated-count variation, contributing μ_1^2 .

Tsiatis et al. (2020), p. 45, equations (2.68)–(2.69); adapted from the earlier course slides.

Fitted propensity: the variance reduction

For the ratio estimator, an exact rearrangement gives

$$\sqrt{n}(\hat{\mu}_1^{\text{fitted}} - \mu_1) = (\mathbb{P}_n Z_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i(Y_i - \mu_1).$$

Now $\mathbb{P}_n Z_i \xrightarrow{p} 1/2$, while

$$\begin{aligned} \text{Var}\{Z_i(Y_i - \mu_1)\} &= \mathbb{E}[Z_i \text{Var}(Y_i - \mu_1 \mid Z_i)] + \text{Var}[Z_i \mathbb{E}(Y_i - \mu_1 \mid Z_i)] \\ &= \frac{\sigma_1^2}{2} + 0. \end{aligned}$$

Hence the central limit theorem and Slutsky's theorem imply

$$\sqrt{n}(\hat{\mu}_1^{\text{fitted}} - \mu_1) \rightsquigarrow N(0, 2\sigma_1^2).$$

Equivalently, the known-score fluctuation decomposes as

$$\sqrt{n}(\hat{\mu}_1^{\text{known}} - \mu_1) = \frac{2}{\sqrt{n}} \sum_i Z_i(Y_i - \mu_1) + \mu_1 \frac{1}{\sqrt{n}} \sum_i (2Z_i - 1).$$

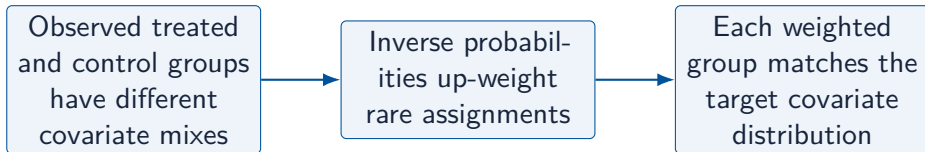
The fitted denominator absorbs the treatment-count term, reducing asymptotic variance by exactly μ_1^2 .

Tsiatis et al. (2020), pp. 45–46, equations (2.70)–(2.71); adapted from the earlier course slides.

Weighting restores covariate balance

With the true propensity score, any integrable function $h(X)$ satisfies

$$\mathbb{E} \left\{ \frac{Z}{e(X)} h(X) \right\} = \mathbb{E}\{h(X)\}, \quad \mathbb{E} \left\{ \frac{1-Z}{1-e(X)} h(X) \right\} = \mathbb{E}\{h(X)\}.$$



Key idea

The propensity model matters because its weights must create balance, not merely predict treatment well.

Ding (2023), pp. 163–164; Wager (2026), Chapter 7, pp. 84–91.

Propensity-score diagnostics

Overlap

Compare score distributions by group.
Limited common support signals weak comparisons.

Weight stability

Inspect the largest weights and the effective sample size.

Covariate balance

Compare covariate means and distributions after weighting.

Estimand sensitivity

Trimming may improve overlap, but it changes the target population.

Practical rule

Diagnose the weighted sample before interpreting the weighted outcome contrast.

Ding (2023), pp. 160–164 and 267–269; Wager (2026), pp. 24–27.

Regression standardization using the propensity score

Let $S = e(X)$ and define

$$m_z^S(s) = \mathbb{E}(Y \mid Z = z, S = s), \quad z \in \{0, 1\}.$$

Because the propensity score is a balancing score,

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid S, \quad \tau = \mathbb{E}\{m_1^S(S) - m_0^S(S)\}.$$

This gives a two-stage outcome-regression estimator:

1. Fit $\hat{e}(X)$.
2. Fit separate, sufficiently flexible outcome curves $\hat{m}_1^S(s)$ and $\hat{m}_0^S(s)$ using the scalar score.
3. Predict both outcomes for every unit and average their difference.

Inference

The score is estimated in the first stage. Use a stacked sandwich or a whole-row bootstrap that refits both stages.

Tsiatis et al. (2020), pp. 39–41, equations (2.54)–(2.56); Ding (2023), pp. 193–196.

Identification through the propensity score

Let $S = e(X)$ and define the observed score-specific outcome means

$$m_z^S(s) = \mathbb{E}(Y \mid Z = z, S = s).$$

Balance and conditional exchangeability imply, for $z \in \{0, 1\}$,

$$\begin{aligned}\mathbb{E}\{Y(z)\} &= \mathbb{E}[\mathbb{E}\{Y(z) \mid S\}] \\ &= \mathbb{E}[\mathbb{E}\{Y(z) \mid Z = z, S\}] \\ &= \mathbb{E}[\mathbb{E}(Y \mid Z = z, S)] \\ &= \mathbb{E}\{m_z^S(S)\}.\end{aligned}$$

Therefore,

$$\tau = \mathbb{E}\{m_1^S(S) - m_0^S(S)\}.$$

Tsiatis et al. (2020), pp. 39–41, equations (2.54)–(2.56); Ding (2023), pp. 193–196.

Propensity-adjusted OLS and its estimand

Consider the population OLS regression of Y on $\{1, Z, e(X)\}$, and let b_e be the coefficient on Z . The Frisch–Waugh–Lovell residual from regressing Z on $\{1, e(X)\}$ is $Z - e(X)$. Hence

$$\begin{aligned} b_e &= \frac{\mathbb{E}[\{Z - e(X)\}Y]}{\mathbb{E}[\{Z - e(X)\}^2]} \\ &= \frac{\mathbb{E}[e(X)\{1 - e(X)\}\{\mu_1(X) - \mu_0(X)\}]}{\mathbb{E}[e(X)\{1 - e(X)\}]} \end{aligned}$$

The numerator follows by writing $Y = ZY(1) + (1 - Z)Y(0)$ and conditioning on X .

Interpretation

The coefficient emphasizes covariate values with the most treatment overlap. It need not equal the population average effect.

For the population average

Model the two score-specific response curves flexibly, allow treatment-score interactions, and standardize both fitted outcomes.

Ding (2023), Theorem 14.1 and Corollary 14.1, pp. 193–196.

AIPW combines predictions and weighted residuals

Define the estimated score for unit i :

$$\hat{\psi}_i = \underbrace{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}_{\text{outcome-regression contrast}} + \underbrace{\frac{Z_i\{Y_i - \hat{\mu}_1(X_i)\}}{\hat{e}(X_i)}}_{\text{treated residual correction}} - \underbrace{\frac{(1 - Z_i)\{Y_i - \hat{\mu}_0(X_i)\}}{1 - \hat{e}(X_i)}}_{\text{control residual correction}}.$$

The AIPW estimator averages these scores:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i.$$

Interpretation

Outcome regression supplies the initial counterfactual predictions. Inverse weighting uses observed residuals to correct systematic prediction errors.

Wager (2026), pp. 29–30; Ding (2023), pp. 169–172.

Two representations of AIPW

Algebra gives two exactly equivalent forms. The first starts from outcome regression:

$$\hat{\tau}_{\text{AIPW}} = \underbrace{\mathbb{P}_n\{\hat{\mu}_1(X) - \hat{\mu}_0(X)\}}_{\hat{\tau}_{\text{REG}}} + \mathbb{P}_n \left[\frac{Z\{Y - \hat{\mu}_1(X)\}}{\hat{e}(X)} - \frac{(1 - Z)\{Y - \hat{\mu}_0(X)\}}{1 - \hat{e}(X)} \right]. \quad (\text{R})$$

The second starts from inverse weighting:

$$\hat{\tau}_{\text{AIPW}} = \underbrace{\mathbb{P}_n \left\{ \frac{ZY}{\hat{e}(X)} - \frac{(1 - Z)Y}{1 - \hat{e}(X)} \right\}}_{\hat{\tau}_{\text{IPW}}} + \mathbb{P}_n \left[\hat{\mu}_1(X) \left\{ 1 - \frac{Z}{\hat{e}(X)} \right\} - \hat{\mu}_0(X) \left\{ 1 - \frac{1 - Z}{1 - \hat{e}(X)} \right\} \right]. \quad (\text{W})$$

Key idea

Representation (R) proves robustness to propensity misspecification. Representation (W) proves robustness to outcome-model misspecification.

Wager (2026), pp. 30–31, equations (3.3)–(3.4).

Route 1: correct outcome regressions

Suppose $\hat{\mu}_z(x)$ converges to the true $\mu_z(x)$, while $\hat{e}(x)$ may be misspecified. In representation (R),

$$\mathbb{E} \left[\frac{Z\{Y - \mu_1(X)\}}{\tilde{e}(X)} \mid X \right] = \frac{\mathbb{P}(Z = 1 \mid X)}{\tilde{e}(X)} \underbrace{\mathbb{E}\{Y - \mu_1(X) \mid Z = 1, X\}}_0 = 0.$$

The same argument applies to the control residual. Therefore,

$$\mathbb{E}\{\psi(O; \mu_0, \mu_1, \tilde{e})\} = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} = \tau.$$

First route to consistency

Correct outcome regressions make the weighted residual corrections mean zero, even when the propensity model is wrong.

Wager (2026), pp. 30–31; Ding (2023), pp. 170–171.

Route 2: correct propensity score

Suppose $\widehat{e}(x)$ converges to the true $e(x)$, while the working outcome functions $\tilde{\mu}_z(x)$ may be misspecified. In representation (W), the two augmentation terms have conditional mean zero:

$$\begin{aligned}\mathbb{E} \left\{ 1 - \frac{Z}{e(X)} \middle| X \right\} &= 0, \\ \mathbb{E} \left\{ 1 - \frac{1-Z}{1-e(X)} \middle| X \right\} &= 0.\end{aligned}$$

Therefore, for any fixed working functions $\tilde{\mu}_1$ and $\tilde{\mu}_0$,

$$\mathbb{E} \left[\tilde{\mu}_1(X) \left\{ 1 - \frac{Z}{e(X)} \right\} - \tilde{\mu}_0(X) \left\{ 1 - \frac{1-Z}{1-e(X)} \right\} \right] = 0.$$

Second route to consistency

The augmentation vanishes, leaving the consistent IPW estimator even when the outcome functions are wrong.

Wager (2026), pp. 30–31; Ding (2023), pp. 170–171.

Limits of weak double robustness

It does guarantee

Under the causal assumptions and regularity conditions, AIPW is consistent if either the outcome regressions or the propensity-score model is consistently estimated.

It does not guarantee

- protection when both models are wrong
- protection from unmeasured confounding
- stability under poor overlap
- valid confidence intervals automatically

Scope

Weak double robustness is a large-sample consistency statement. It is not a substitute for research design or diagnostics.

Wager (2026), p. 31; Ding (2023), pp. 175–177.

Strong double robustness

Use cross-fitting and assume strong overlap. Suppose the nuisance estimators satisfy

$$\|\hat{\mu}_z - \mu_z\|_2 = o_p(n^{-\alpha_\mu}), \quad \|\hat{e} - e\|_2 = o_p(n^{-\alpha_e}).$$

The leading remainder is a product of the two nuisance errors. By Cauchy–Schwarz,

$$|R_n| \lesssim \max_z \|\hat{\mu}_z - \mu_z\|_2 \|\hat{e} - e\|_2 = o_p\{n^{-(\alpha_\mu + \alpha_e)}\}.$$

If $\alpha_\mu + \alpha_e \geq 1/2$, then $R_n = o_p(n^{-1/2})$ and feasible AIPW is oracle equivalent:

$$\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \tau) \rightsquigarrow N(0, V^*), \quad V^* = \text{Var}\{\mu_1(X) - \mu_0(X)\} + \mathbb{E} \left\{ \frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right\}.$$

Key idea

The product rate drives the result. For example, two $o_p(n^{-1/4})$ nuisance rates suffice for root- n inference.

Wager (2026), pp. 31–37, especially Theorem 3.2 and equations (3.5), (3.9), and (3.15).

AIPW as a joint M-estimator

Let $m_z(X; \beta) = m(z, X; \beta)$, $e_\gamma(X) = \text{expit}(\gamma^\top X)$, and $\theta = (\tau, \gamma^\top, \beta^\top)^\top$. We can write the target equation as

$$g_\tau(O; \theta) = \frac{ZY}{e_\gamma(X)} - \frac{(1-Z)Y}{1-e_\gamma(X)} - \frac{Z-e_\gamma(X)}{e_\gamma(X)}m_1(X; \beta) \\ - \frac{Z-e_\gamma(X)}{1-e_\gamma(X)}m_0(X; \beta) - \tau.$$

Estimate all parameters together from

$$\mathbb{P}_n M_{\text{AIPW}}(O; \theta) = 0, \quad M_{\text{AIPW}}(O; \theta) = \begin{pmatrix} g_\tau(O; \theta) \\ X\{Z - e_\gamma(X)\} \\ m_\beta(Z, X; \beta)\{Y - m(Z, X; \beta)\} \end{pmatrix}.$$

Key idea

The first equation is AIPW. The second fits the logistic propensity model, and the third fits the outcome regression by least squares.

Tsiatis et al. (2020), pp. 47–49, especially equations (2.72)–(2.79).

Sandwich inference for AIPW

For the stacked estimating equations, compute

$$\hat{A} = -\mathbb{P}_n \frac{\partial M_{\text{AIPW}}(O; \hat{\theta})}{\partial \theta^\top}, \quad \hat{B} = \mathbb{P}_n M_{\text{AIPW}}(O; \hat{\theta}) M_{\text{AIPW}}(O; \hat{\theta})^\top.$$

Then

$$\hat{\Sigma} = \hat{A}^{-1} \hat{B} \hat{A}^{-\top}, \quad \text{SE}(\hat{\tau}_{\text{AIPW}}) = \sqrt{\hat{\Sigma}_{11}/n}.$$

One model is correct

The target equation remains unbiased at the nuisance limits. Use the full sandwich because nuisance fitting can contribute at first order.

Bootstrap alternative. Resample whole rows, refit both nuisance models, and recompute AIPW in every sample.

Tsiatis et al. (2020), pp. 32–35 and 49; Ding (2023), pp. 172–175.

Both models are correct

The target moment is orthogonal to both nuisance parameters. The variance of the AIPW scores gives the same first-order result.

Comparison of the three estimators

Estimator	Models	Core operation	Main vulnerability
Outcome regression	$\mu_1(x), \mu_0(x)$	Predict both conditional means for every unit, then average their difference	Misspecified response surfaces and unsupported extrapolation
IPW (HT or Hájek)	$e(x)$	Reweight observed outcomes; Hájek normalizes weights within each treatment group	Extreme weights, limited overlap, and assignment-model error
AIPW	$\mu_1(x), \mu_0(x), e(x)$	Average regression contrasts plus weighted residual corrections	Both models wrong, limited overlap, or invalid causal assumptions

Key idea

Agreement across estimators can be reassuring, but it cannot test ignorability. Large disagreements often reveal modeling or overlap problems worth investigating.

A practical observational-study workflow

1. State the treatment, outcome, target population, and average causal effect.
2. Use subject-matter knowledge to choose pretreatment covariates and assess ignorability.
3. Examine overlap before fitting an effect estimator.
4. Fit the outcome and propensity models with a prespecified strategy.
5. Estimate the effect with OR, IPW, and AIPW as complementary views of the same target.
6. Diagnose outcome fit, weight stability, and weighted covariate balance.
7. Quantify uncertainty with a stacked sandwich or whole-row bootstrap that refits every nuisance model.

Priority

Estimator choice matters only after the causal question, assumptions, and target population are clear.

Synthesis of Wager (2026), Chapters 2–3 and 12, and Ding (2023), Chapters 10–14.

Summary

- Random assignment makes treatment groups comparable, so difference in means identifies the average effect.
- Observational studies replace randomization with consistency, conditional exchangeability, and overlap.
- Outcome regression standardizes two fitted potential-outcome means. In nonlinear models, the treatment coefficient alone is generally not the average effect.
- The propensity score can replace X in a correctly specified outcome regression, followed by standardization. A simple additive score adjustment generally targets an overlap-weighted effect instead.

Key idea

Identification comes from the design and assumptions. Regression or weighting then turns the identified formula into an estimator.

Summary: weighting and inference

- Horvitz–Thompson uses unnormalized inverse weights. Hájek normalizes within treatment groups and is invariant to shifts in the outcome.
- Fitting a correct propensity model can reduce the raw HT estimator's variance by enforcing sample balance, but the claim does not extend to arbitrary fitted models.
- AIPW has a regression-plus-residual form and an IPW-plus-augmentation form. These give the two weak robustness arguments.
- Use a stacked sandwich or a whole-row nonparametric bootstrap that refits every nuisance model.
- For parametric working models, the stacked sandwich accounts for both nuisance fits. Under cross-fitting and the product-rate condition, strong double robustness justifies the AIPW score variance.

Bottom line: Separate identification, point estimation, and inference; each requires its own conditions.

References

-  Peng Ding.
A First Course in Causal Inference.
2023. Chapters 2, 4, 10–14, and Appendix A.6.
-  Stefan Wager.
Causal Inference: A Statistical Learning Approach.
Draft, September 8, 2026. Chapters 1–3 and 12.
-  Anastasios A. Tsiatis, Marie Davidian, Shannon T. Holloway, and Eric B. Laber.
Dynamic Treatment Regimes: Statistical Methods for Precision Medicine.
CRC Press, 2020. Chapter 2.
-  Paul R. Rosenbaum and Donald B. Rubin.
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1):41–55, 1983.
-  James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao.
Estimation of regression coefficients when some regressors are not always observed.
Journal of the American Statistical Association, 89(427):846–866, 1994.