Contents

1 Propensity Scores and Inverse Probability Weighting

This document provides a rigorous treatment of **propensity score methods** and **inverse probability weighting (IPW)** for estimating causal effects from observational data. We derive the IPW estimator from first principles, explain the intuition behind propensity score balancing, and clarify common conceptual subtleties.

1.1 What You'll Learn

- How propensity scores enable causal inference by balancing covariates
- The mathematical derivation of the IPW identification formula
- Practical implementation: estimators, stabilized weights, and diagnostics

- The relationship between IPW and propensity score matching
- When and why IPW can fail (and what to do about it)

1.2 Prerequisites

- Familiarity with the potential outcomes framework
- Understanding of conditional expectation and the law of iterated expectations
- Basic probability theory (conditional independence, Bayes' rule)

1.3 Table of Contents

1.4 1. Setup and Notation

We work within the standard binary-treatment causal inference framework.

1.4.1 Notation

For each unit $i = 1, \ldots, n$:

- **Treatment:** $T_i \in \{0, 1\}$
  - $T_i = 1$: unit received treatment
  - $T_i = 0$: unit received control
- **Observed outcome:** Y_i
- **Covariates (pre-treatment features):** X_i (can be a vector)
- **Potential outcomes:**
  - $Y_i(1)$: outcome if unit i were treated
  - $Y_i(0)$: outcome if unit i were untreated

Only one potential outcome is observed:

$$Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$$

Target estimand (ATE):

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

1.4.2 Key Assumptions

1. **Consistency:** If $T = t$, then $Y = Y(t)$
2. **Unconfoundedness (Ignorability):**

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

   Given covariates X, treatment assignment is independent of potential outcomes

3. **Positivity (Overlap):**
$$0 < \mathbb{P}(T = 1 \mid X = x) < 1$$
for all relevant x (every covariate region has both treated and control units)

1.5 2. Propensity Scores: Definition and Properties

1.5.1 Definition

The **propensity score** is the conditional probability of receiving treatment given covariates:
$$e(X) := \mathbb{P}(T = 1 \mid X)$$

This single number (even when X is high-dimensional) summarizes how likely treatment was given covariates.

1.5.2 The Balancing Property (Rosenbaum–Rubin)

If treatment is ignorable given X, then it's also ignorable given $e(X)$. More precisely, conditioning on $e(X)$ **balances covariates**:
$$T \perp\!\!\!\perp X \mid e(X)$$

Interpretation: Among units with the same propensity score, the treated and control groups have similar covariate distributions. The propensity score is a **balancing score**.

Why this matters: We can make observational data behave like a randomized trial by balancing on $e(X)$ instead of the full covariate vector X.

1.6 3. Inverse Probability Weighting: The Core Idea

IPW strategy: Reweight the sample so that both the treated and control groups look like the full population, thereby correcting the covariate imbalance induced by non-random treatment assignment.

1.6.1 The Weights

- **Treated units** ($T = 1$): weight $\propto 1/e(X)$
- **Control units** ($T = 0$): weight $\propto 1/(1 - e(X))$

This is why it's called **inverse probability** weighting—we weight by the inverse of the probability of receiving the treatment actually received.

1.7 4. Deriving the IPW Identification Formula

Our goal is to express $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ in terms of observed data. We'll derive the formula for $\mathbb{E}[Y(1)]$ step-by-step.

1.7.1 Step 1: Law of Total Expectation

Condition on X:
$$\mathbb{E}[Y(1)] = \mathbb{E}\big[\mathbb{E}[Y(1) \mid X]\big]$$

1.7.2 Step 2: Apply Unconfoundedness

Unconfoundedness implies $Y(1) \perp\!\!\!\perp T \mid X$, so:
$$\mathbb{E}[Y(1) \mid X] = \mathbb{E}[Y(1) \mid T = 1, X]$$

1.7.3 Step 3: Apply Consistency

By consistency, when $T = 1$, we have $Y = Y(1)$:

$$\mathbb{E}[Y(1) \mid T = 1, X] = \mathbb{E}[Y \mid T = 1, X]$$

Combining steps 1-3:

$$\mathbb{E}[Y(1)] = \mathbb{E}\big[\mathbb{E}[Y \mid T = 1, X]\big]$$

Now we need to express this as an expectation over observed data using weights.

1.7.4 Step 4: The Weighting Trick

Consider the weighted random variable:

$$\frac{T \cdot Y}{e(X)}$$

Take its conditional expectation given X:

$$\mathbb{E}\left[\frac{T \cdot Y}{e(X)} \mid X\right] = \frac{1}{e(X)} \cdot \mathbb{E}[T \cdot Y \mid X]$$

Expand $\mathbb{E}[T \cdot Y \mid X]$ using iterated expectation:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{E}\big[\mathbb{E}[T \cdot Y \mid T, X] \mid X\big]$$

Inside the inner expectation, T is fixed (either 0 or 1): * If $T = 1$: $T \cdot Y = Y$ * If $T = 0$: $T \cdot Y = 0$

Therefore:

$$\mathbb{E}[T \cdot Y \mid T, X] = T \cdot \mathbb{E}[Y \mid T, X]$$

Substituting back:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{E}\big[T \cdot \mathbb{E}[Y \mid T, X] \mid X\big]$$

Since T is Bernoulli with $\mathbb{E}[T \mid X] = e(X)$:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{P}(T = 1 \mid X) \cdot \mathbb{E}[Y \mid T = 1, X] = e(X) \cdot \mathbb{E}[Y \mid T = 1, X]$$

Therefore:

$$\mathbb{E}\left[\frac{T \cdot Y}{e(X)} \mid X\right] = \frac{1}{e(X)} \cdot e(X) \cdot \mathbb{E}[Y \mid T = 1, X] = \mathbb{E}[Y \mid T = 1, X]$$

Taking expectation over X:

$$\mathbb{E}\left[\frac{T \cdot Y}{e(X)}\right] = \mathbb{E}\big[\mathbb{E}[Y \mid T = 1, X]\big] = \mathbb{E}[Y(1)]$$

1.7.5 The IPW Identification Formula

$$\boxed{\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{T \cdot Y}{e(X)}\right]}$$

Similarly:

$$\boxed{\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{(1-T) \cdot Y}{1 - e(X)}\right]}$$

Therefore, the ATE is:

$$\boxed{\text{ATE} = \mathbb{E}\left[\frac{T \cdot Y}{e(X)} - \frac{(1-T) \cdot Y}{1 - e(X)}\right]}$$

1.8 5. The Sample IPW Estimator

In practice, we estimate $e(X)$ using a propensity score model $\hat{e}(X)$ (e.g., logistic regression, gradient boosting, random forest, neural network).

The sample IPW estimators are:

$$\widehat{\mathbb{E}[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)}, \quad \widehat{\mathbb{E}[Y(0)]} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}$$

The IPW estimate of ATE is:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}\right)$$

1.9 6. Stabilized Weights

Plain IPW weights can become extremely large when $\hat{e}(X)$ is near 0 or 1, leading to high variance. **Stabilized weights** reduce this variance:

$$w_i^{\text{stab}} = \begin{cases} \frac{\mathbb{P}(T=1)}{\hat{e}(X_i)} & \text{if } T_i = 1 \\ \frac{\mathbb{P}(T=0)}{1-\hat{e}(X_i)} & \text{if } T_i = 0 \end{cases}$$

where $\mathbb{P}(T=1)$ is the marginal probability of treatment (sample proportion).

Advantages: * Preserve approximate sample size * Reduce variance while maintaining consistency * More stable in practice

---

1.10 7. Geometric Intuition: What IPW Really Does

IPW creates a **pseudo-population** where treatment is independent of covariates.

The logic: * If you were **very likely** to be treated ($e(X)$ large), you don't represent many "missing" people → **small weight** * If you were **unlikely** to be treated ($e(X)$ small) but you *did* get treated, you are rare and informative → **large weight**

This reweighting corrects for selection bias by making the treated group look like the full population (and similarly for controls).

1.11 8. Essential Diagnostics

These diagnostics are **not optional**—they reveal whether IPW is appropriate for your data.

1.11.1 1. Overlap/Positivity Check

Plot the distribution of $\hat{e}(X)$ separately for treated and control groups. Look for: * Sufficient overlap between distributions * No propensity scores near 0 or 1 * No regions where one group is absent

1.11.2 2. Weight Diagnostics

Examine: * Maximum weight * Weight percentiles (95th, 99th) * Effective sample size: $n_{\text{eff}} = \frac{(\sum w_i)^2}{\sum w_i^2}$

Extreme weights indicate poor overlap and unstable estimates.

1.11.3 3. Balance After Weighting

Compute standardized mean differences (SMD) for each covariate:

$$\text{SMD} = \frac{\bar{X}_{\text{treated}} - \bar{X}_{\text{control}}}{\sqrt{(s^2_{\text{treated}} + s^2_{\text{control}})/2}}$$

After weighting, SMD should be close to 0 (typically < 0.1).

Warning: If weights are extreme, the data is telling you there's insufficient overlap. This is a fundamental problem, not a minor inconvenience.

1.12 9. Beyond ATE: Estimating ATT

Sometimes the **Average Treatment Effect on the Treated (ATT)** is more relevant:

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

This is the effect for those who actually received treatment.

1.12.1 ATT Weighting Scheme

- **Treated units:** weight = 1
- **Control units:** weight = $\frac{\hat{e}(X)}{1-\hat{e}(X)}$

This reweights controls to match the covariate distribution of the treated group.

1.13 10. Propensity Score Methods: The Big Picture

The propensity score is a versatile tool used in multiple ways:

1.13.1 Methods

1. **Weighting** (IPW, ATT weights, overlap weights)
   - Uses all data
   - Sensitive to overlap violations
   - Clean theoretical properties
2. **Matching** (nearest neighbor on $e(X)$)
   - Discards unmatched units
   - Robust to extreme propensities
   - Often estimates ATT
3. **Stratification** (bin into propensity score strata)
   - Simple and intuitive
   - Can lose efficiency
4. **Covariate adjustment** (regression with T and $e(X)$)
   - Combines propensity score with outcome modeling

Trade-offs: Weighting is cleanest for estimation and diagnostics but most sensitive to overlap. Matching is more robust but discards data.

1.14 11. Limitations and Practical Considerations

1.14.1 The Fundamental Limitation

Propensity score methods only adjust for observed confounders in X. If an important confounder is unmeasured, IPW cannot fix the bias.

1.14.2 Practical Strategies

1. **Domain knowledge:** Build a defensible set of covariates X based on subject-matter expertise
2. **Sensitivity analysis:** Use methods like Rosenbaum bounds or E-values to assess robustness to unmeasured confounding
3. **Negative controls:** Test for residual confounding using outcomes that shouldn't be affected by treatment
4. **Alternative identification strategies:** Consider instrumental variables, front-door criterion, or natural experiments
5. **Doubly robust estimators:** Use AIPW or TMLE, which combine propensity scores with outcome modeling for improved stability

1.14.3 Implementation Pipeline

A typical IPW workflow: 1. Fit propensity score model $\hat{e}(X)$ 2. Compute weights 3. Check diagnostics (overlap, balance, weight distribution) 4. Compute weighted means 5. Estimate standard errors (robust sandwich estimator or bootstrap)

Next step: Doubly robust estimators (AIPW/TMLE) often outperform plain IPW, especially in high-dimensional settings common in genomics.

1.15 12. Deep Dive: Clarifying Key Concepts

This section addresses common conceptual subtleties that arise when learning IPW.

1.15.1 12.1 Why $\mathbb{E}[Y \mid T = 1, X]$ is "Just a Function of X"

When we write $\mathbb{E}[Y \mid T = 1, X]$, this might seem like a random variable since Y, T, and X are all random. However, **conditioning freezes randomness**.

Key insight: Once you condition on $T = 1$ and $X = x$, the expectation is a **deterministic number**:

$$\mathbb{E}[Y \mid T = 1, X = x] \in \mathbb{R}$$

As x varies, this defines a **function**:

$$m_1(x) := \mathbb{E}[Y \mid T = 1, X = x]$$

This is no different from linear regression:

$$\mathbb{E}[Y \mid X = x] = \beta^\top x$$

which everyone treats as a function of x.

Why conditioning "freezes" randomness: Inside the conditional expectation $\mathbb{E}[Y \mid T = 1, X = x]$: * $T = 1$ is fixed (not random) * $X = x$ is fixed (not random) * Only Y varies across hypothetical repetitions

The result is a single number, not a random variable.

1.15.2 12.2 Unpacking the Key Derivation Step

In Step 4 of the IPW derivation, we used:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{E}\big[T \cdot \mathbb{E}[Y \mid T, X] \mid X\big]$$

Let's justify this carefully using iterated expectation:

Step A: Apply the law of iterated expectations:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{E}\big[\mathbb{E}[T \cdot Y \mid T, X] \mid X\big]$$

Step B: Evaluate the inner expectation. Once T is fixed, it's just a constant:

$$\mathbb{E}[T \cdot Y \mid T, X] = T \cdot \mathbb{E}[Y \mid T, X]$$

Step C: Substitute back:

$$\mathbb{E}[T \cdot Y \mid X] = \mathbb{E}\big[T \cdot \mathbb{E}[Y \mid T, X] \mid X\big]$$

Step D: Since T is Bernoulli given X, only the $T = 1$ branch contributes:

$$\mathbb{E}[T \cdot \mathbb{E}[Y \mid T, X] \mid X] = e(X) \cdot \mathbb{E}[Y \mid T = 1, X] + (1 - e(X)) \cdot 0$$

Therefore:

$$\mathbb{E}[T \cdot Y \mid X] = e(X) \cdot \mathbb{E}[Y \mid T = 1, X]$$

This is a weighted average over the two possible values of T, not magic.

1.15.3 12.3 IPW vs. Propensity Score Matching

Are they the same? No—they are related but distinct approaches.

1.15.3.1 Propensity Score Matching (Discrete Geometry)

Strategy: For each treated unit, find control units with similar $e(X)$ and compare outcomes.

Characteristics: * Discards unmatched units * Local comparisons * Often estimates ATT * Balance achieved by **selection**

Metaphor: Carving out a subset where treated and control units resemble each other.

1.15.3.2 IPW (Continuous Geometry)

Strategy: Keep everyone, but reweight so treatment is independent of covariates.

Characteristics: * Uses all observations * Global reweighting * Naturally estimates ATE * Balance achieved by **rescaling**

Metaphor: Warping the population density to simulate a randomized trial.

1.15.3.3 The Key Difference

- **Matching asks:** "Who should I compare to whom?"
- **IPW asks:** "How much should each observation count?"

Matching is discrete and combinatorial; IPW is continuous and expectation-based.

1.15.3.4 Practical Implications

- **Matching:** Robust to extreme propensities but discards data
- **IPW:** Uses all data but can be unstable under poor overlap
- **Matching:** Harder to analyze asymptotically
- **IPW:** Fits naturally into semiparametric theory

You can view IPW as an infinite-sample, smooth analogue of matching, but they are fundamentally different approaches.

1.15.4 12.4 Key Takeaways

1. **Conditional expectations are functions:** $\mathbb{E}[Y \mid T = 1, X]$ is a deterministic function of X because conditioning freezes randomness.
2. **IPW is not matching:** IPW is a population-reweighting strategy, not a discrete matching procedure.
3. **The derivation is inevitable:** Once you understand that conditioning freezes randomness and that we can reweight to balance covariates, the IPW formula follows naturally from iterated expectations.

1.16 Next Steps

This document covered IPW in depth. Natural extensions include:

- **Doubly robust estimation (AIPW/TMLE):** Combines propensity scores with outcome modeling for improved robustness
- **Overlap weights:** Alternative weighting scheme that emphasizes regions of good overlap
- **Sensitivity analysis:** Methods for assessing robustness to unmeasured confounding
- **High-dimensional propensity scores:** Regularization and machine learning approaches for genomics applications