Contents

Let's build a clean mental model of **propensity scores** and **inverse probability weighting (IPW)**, starting from first principles and then getting into the "why this works" math. I'll keep the notation explicit and I'll narrate each derivation step instead of doing the usual paper-wizard teleportation.

0.1 1) Setup and notation (the chessboard)

We're in the standard binary-treatment causal inference world.

- **Units:** $i = 1, \dots, n$
- **Treatment:** $T_i \in \{0, 1\}$ $T_i = 1$ means treated, $T_i = 0$ means control.
- **Observed outcome:** Y_i
- **Covariates (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):

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

0.1.1 Key Assumptions

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

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

Meaning: once you condition on X, treatment is "as good as random".

3. **Positivity (overlap):**

$$0 < \mathbb{P}(T = 1 \mid X = x) < 1$$

for relevant x. (No covariate region where treatment is impossible or guaranteed.)

0.2 2) Propensity score: what it is and why it's magical

Definition (propensity score):

$$e(X) := \mathbb{P}(T = 1 \mid X)$$

It's a single number (even when X is huge) summarizing how likely treatment was given covariates.

0.2.1 The key theorem (Rosenbaum–Rubin intuition)

If treatment is ignorable given X, then it's also ignorable given $e(X)$ in the sense that **conditioning on $e(X)$ balances covariates**:

$$T \perp X \mid e(X)$$

Translation: among people with the same propensity score, the treated and control groups should look similar in X. So propensity score is a **balancing score**.

This is why propensity methods exist at all: you can try to "make observational data behave like a randomized trial" by balancing on $e(X)$.

0.3 3) IPW in one sentence (then we'll earn it)

IPW idea: reweight the sample so that the treated group looks like the full population (and the control group also looks like the full population), correcting the covariate imbalance induced by non-random treatment.

Weights:

- For treated ($T = 1$): weight $\propto 1/e(X)$
- For control ($T = 0$): weight $\propto 1/(1 - e(X))$

This is why it's called **inverse probability** weighting.

0.4 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.

0.4.1 Step 1: Law of Total Expectation

Condition on X:

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

0.4.2 Step 2: Apply Unconfoundedness

Unconfoundedness implies $Y(1) \perp\!\!\!\perp T \mid X$, so:

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

0.4.3 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.

0.4.4 Step 4: Convert "conditional-on-treated" into a weighted unconditional expectation

Consider the random variable:

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

Take its conditional expectation given X:

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

Now expand $\mathbb{E}[T \cdot Y \mid X]$ using iterated expectation conditioning on T:

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

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

- if $T = 1$: $T \cdot Y = 1 \cdot Y = Y$
- if $T = 0$: $T \cdot Y = 0$

So:

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

Plug back:

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

But T is Bernoulli with $\mathbb{E}[T \mid X] = e(X)$. Also $\mathbb{E}[Y \mid T = 1, X]$ is just a function of X (call it $m_1(X)$). Then:

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

Therefore:

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

Now take expectation over X:

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

Boom: we got the identification formula:

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

Similarly:

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

So the ATE is:

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

0.5 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:

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

0.6 6) Stabilized weights (less "numerical drama")

Plain IPW can explode when $\hat{e}(X)$ is near 0 or 1. Stabilized weights reduce variance:

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

These preserve (approximately) sample size and tend to behave better.

0.7 7) What IPW is really doing (a geometric intuition)

Imagine your treated sample is "over-represented" in some covariate region where treatment is likely. IPW says:

- If you were very likely to be treated ($e(X)$ large), you don't represent many "missing" people $\rightarrow$ small weight.
- If you were unlikely to be treated ($e(X)$ small) but you *did* get treated, you are rare and informative $\rightarrow$ huge weight.

So IPW is creating a **pseudo-population** where treatment is independent of X.

0.8 8) Diagnostics you should basically never skip

1. **Overlap / positivity check:** look at distributions of $\hat{e}(X)$ by treatment group.
2. **Weight diagnostics:** max weight, percentiles, effective sample size.
3. **Balance after weighting:** standardized mean differences (SMD) of covariates should shrink toward 0.

If weights are insane, it's not "a minor inconvenience"—it's the data screaming "you don't have overlap."

0.9 9) Going deeper: ATE vs ATT via weighting

Sometimes you don't want ATE. You want:

- **ATT**: effect on treated $\mathbb{E}[Y(1) - Y(0) \mid T = 1]$

A classic weighting scheme for ATT:

- treated weight $= 1$
- control weight $= \hat{e}(X)/(1 - \hat{e}(X))$

This reweights controls to look like treated.

0.10 10) Where propensity score methods fit (big picture)

Propensity score is used as a tool for:

- **Weighting** (IPW, ATT weights, overlap weights)
- **Matching** (nearest neighbor on $e(X)$)
- **Stratification** (bin into propensity score strata)
- **Covariate adjustment** (regress outcome on T and $e(X)$, sometimes)

Weighting is often the cleanest for estimation + diagnostics, but it's also the most sensitive to overlap issues.

0.11 11) The "I can't block all confounders" reality check

This is the uncomfortable truth: **propensity score methods only adjust for observed confounders** in X. If an important confounder is missing, weighting doesn't magically fix it.

What you can do anyway (practical survival kit):

- Use domain knowledge to build a defensible X.
- Sensitivity analysis (e.g., Rosenbaum bounds / E-values style thinking).
- Negative controls (if you have them).
- Instrumental variables / front-door / natural experiments (if structure allows).
- Prefer doubly robust estimators (AIPW / TMLE), which are often more stable.

(We can go there next; it's a natural continuation from IPW.)

0.11.1 Follow-up Q&A

0.12 Part I — Why is

$\mathbb{E}[Y \mid T = 1, X]$ "just a function of X"?

Let's be very precise about what kind of mathematical object this is.

0.12.1 1) Conditional expectation is a function, not a random variable

When we write $\mathbb{E}[Y \mid T = 1, X]$, we are **not** talking about something random anymore.

- Y is random
- T is random
- X is random

But **once you condition on $T = 1$ and $X = x$**, the expectation is a **number**.

Formally:

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

So as x varies, this defines a **function of x**:

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

That's all that's meant by "it's just a function of X".

No probability left. No randomness left. Just a deterministic mapping:

$$x \longmapsto \text{average outcome among treated with covariates } x$$

This is no different in spirit from linear regression:

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

which everyone happily treats as a function of x.

0.12.2 2) Why conditioning on $T = 1$ doesn't make it random again

This is a subtle psychological trap.

You might think: "Wait, T is random — so how can conditioning on $T = 1$ give something deterministic?"

Answer: **conditioning freezes randomness**.

Once we condition on an event, it is no longer random *within that conditional world*.

Inside the conditional expectation $\mathbb{E}[Y \mid T = 1, X = x]$:

- $T = 1$ is fixed
- $X = x$ is fixed
- Only Y varies across hypothetical repetitions

So the result is a single number.

This is why, inside expectations, we are allowed to treat $\mathbb{E}[Y \mid T = 1, X]$ as a function $m_1(X)$.

0.12.3 3) The key step you highlighted, unpacked slowly

You highlighted this transformation:

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

Let's justify it carefully.

0.12.3.1 Step A: Use iterated expectation

$$\mathbb{E}[TY \mid X] = \mathbb{E}\left[\mathbb{E}[TY \mid T, X] \mid X\right]$$

This is always valid.

0.12.3.2 Step B: Evaluate the inner expectation Inside $\mathbb{E}[TY \mid T, X]$, once T is fixed, it is just a constant (0 or 1). So:

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

No probability trickery here — it's just pulling out a constant.

0.12.3.3 Step C: Substitute back

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

Now comes the crucial observation:

- $\mathbb{E}[Y \mid T = 1, X]$ is a function of X
- $\mathbb{E}[Y \mid T = 0, X]$ is another function of X

So when you average over $T \mid X$, only the $T = 1$ branch survives.

0.12.4 4) Why only the treated branch survives

Condition on X. Then:

$$T = \begin{cases} 1 & \text{with probability } e(X) \\ 0 & \text{with probability } 1 - e(X) \end{cases}$$

So:

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

Hence:

$$\mathbb{E}[TY \mid X] = e(X) \cdot m_1(X)$$

That step is not magic — it's literally a weighted average over the two possible values of T.

0.13 Part II — Is IPW just a systematic way to do propensity score matching?

Short answer: **no — they are cousins, not the same creature**.

Long answer: they solve the *same problem* (covariate imbalance) using *different geometry*.

0.13.1 1) What matching does (discrete geometry)

Propensity score matching says:

> "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*

Think of matching as **carving out a subset** of the data where treated and control units resemble each other.

0.13.2 2) What IPW does (continuous geometry)

IPW says:

> "Keep everyone, but reweight them so that treatment is independent of covariates."

Characteristics:

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

Think of IPW as **warping the population density** to simulate a randomized trial.

0.13.3 3) Why they are not the same

Matching answers:

> "Who should I compare to whom?"

IPW answers:

> "How much should each observation count?"

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

You *can* view IPW as an **infinite-sample, smooth analogue** of matching — but that's an intuition, not an identity.

0.13.4 4) Practical consequences (why this distinction matters)

- Matching is robust to extreme propensities but throws away data
- IPW keeps data but can explode under poor overlap
- Matching is harder to analyze asymptotically
- IPW drops neatly into estimating equations and semiparametric theory

This is why:

- IPW is favored in theory and pipelines
- Matching is favored for intuition and presentation
- Doubly robust estimators blend IPW with outcome modeling

0.14 Final mental anchor

Two takeaways worth engraving:

1. $\mathbb{E}[Y \mid T = 1, X]$ is **not random** — it is a deterministic function of X, because conditioning freezes randomness.
2. IPW is **not matching with math polish**. It is a **population-reweighting strategy** that creates a pseudo-randomized world.

Once these two ideas click, the IPW derivation stops feeling like algebraic sleight of hand and starts feeling inevitable.

Next natural step: show how **AIPW/TMLE** repair IPW's fragility by adding outcome models — that's where causal inference really starts to feel like engineering rather than hope.