Inverse probability weighting

1. What would the treated have experienced without treatment?

Suppose we observe whether each person received a treatment, their outcome, and some pre-treatment characteristics. Write these as D, Y, and X: $D = 1$ means treated, $D = 0$ means untreated, and X can contain several characteristics.

Each person has two potential outcomes: $Y(1)$ under treatment and $Y(0)$ without treatment. We observe only the one corresponding to their actual treatment:

$$Y = DY(1) + (1 - D)Y(0).$$

The target in this note, as in the accompanying presentation, is the **average treatment effect on the treated (ATT)**:

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

Here $\mathbb{E}[\cdot \mid D = 1]$ means an average over the treated population. Split the difference into its two means:

$$\tau = \mathbb{E}[Y \mid D = 1] - \underbrace{\mathbb{E}[Y(0) \mid D = 1]}_{\theta}.$$

The first mean is observable: for treated people, $Y = Y(1)$. The second mean, which we call θ, is missing. **All the work is in finding θ.**

Why not use the ordinary control mean?

Controls reveal untreated outcomes. But their characteristics may differ from those of the treated. If higher-risk people are more likely to receive treatment, for example, the controls' average untreated outcome may be a poor guide to what the treated would have experienced.

Inverse probability weighting (IPW) keeps the observed control outcomes but changes how much each control counts. It constructs a control group with the treated group's composition in X. Under an assumption explained below, its weighted mean then recovers θ.

There are two distinct questions to answer:

1. **Identification:** why does the population weighted mean equal the missing mean?
2. **Consistency:** why does the estimate computed from data approach that population mean as the sample grows?

We first make the weighting concrete, then answer these questions in that order.

2. A small example: change how much each control counts

Use the same example as the presentation. A characteristic X takes the values $0, 1, 2$, each in one third of the population. Treatment probabilities are $1/4, 1/2, 3/4$, respectively. We can express those population proportions as counts out of twelve:

Characteristic x	Treated	Controls	Weight per control	Weighted controls
0	1	3	1/3	1
1	2	2	1	2
2	3	1	3	3
Total	6	6		6

At $x = 2$, one control must represent three treated people, so it receives weight three. At $x = 0$, three controls must together represent one treated person, so each receives weight one third. The final column now has exactly the treated composition. These are **population proportions expressed as counts**, not the promised outcome of a random sample of twelve people.

Suppose the mean untreated outcomes within the three groups are $0, 1, 4$, for both treated and control people. Treatment adds $0, 2, 4$, respectively. Thus the observed treated means within the groups are $0, 3, 8$.

The ordinary control mean uses control counts $(3, 2, 1)$:

$$\text{ordinary control mean} = \frac{3 \cdot 0 + 2 \cdot 1 + 1 \cdot 4}{6} = 1.$$

The missing untreated mean for the treated instead uses counts $(1, 2, 3)$:

$$\theta = \frac{1 \cdot 0 + 2 \cdot 1 + 3 \cdot 4}{6} = \frac{7}{3}.$$

Weighting the observed controls produces exactly this second calculation:

$$\text{weighted control mean} = \frac{3(1/3) \cdot 0 + 2(1) \cdot 1 + 1(3) \cdot 4}{3(1/3) + 2(1) + 1(3)} = \frac{7}{3}.$$

The denominator is the **sum of the weights**, counting the weight of every control. This is how any weighted average is formed.

The observed treated mean is $(1 \cdot 0 + 2 \cdot 3 + 3 \cdot 8)/6 = 5$. Therefore,

$$\tau = 5 - \frac{7}{3} = \frac{8}{3} \approx 2.667.$$

The unadjusted difference would be $5 - 1 = 4$. Weighting removes the part of that gap caused by the different mix of X. Next we show why these weights work generally.

3. Why these particular weights?

The **propensity score** is the probability of treatment at a given value of X:

$$p_0(x) = \mathbb{P}(D = 1 \mid X = x).$$

The subscript zero labels the true population probability. Later, $\hat{p}(x)$ will denote its estimate. Among people with $X = x$, the treated fraction is $p_0(x)$ and the control fraction is $1 - p_0(x)$. Their ratio gives the weight per control:

$$o_0(x) = \frac{p_0(x)}{1 - p_0(x)}.$$

This is the **odds of treatment**. Multiplying the control fraction by the odds gives the treated fraction:

$$\underbrace{(1 - p_0(x))}_{\text{control fraction}} \underbrace{\frac{p_0(x)}{1 - p_0(x)}}_{\text{weight per control}} = \underbrace{p_0(x)}_{\text{treated fraction}}.$$

That cancellation is the central mechanism. It works within every covariate group, including when the overall treated and control groups have different sizes.

When can this recover an untreated outcome for the treated?

Matching the composition is only part of the argument. We also need controls to provide the right untreated mean *within* each covariate group.

Comparable untreated outcomes given X. Assume $Y(0) \perp\!\!\!\perp D \mid X$: once X is held fixed, treatment status conveys no further information about the untreated potential outcome. This is conditional unconfoundedness. In particular, for groups represented among the treated,

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

The first equality uses the causal assumption; the second uses $Y = Y(0)$ for controls. Equality of these conditional means is all the proof needs. Weighting cannot establish that assumption from the observed data.

Available controls. Write $\pi = \mathbb{P}(D = 1)$ and assume $\pi > 0$ and $p_0(X) \leq 1 - \epsilon$ for some $\epsilon > 0$. This overlap condition ensures there are controls wherever there are treated people, and bounds their weights. Values with $p_0(x) = 0$ are allowed: they contain no treated people and receive zero weight.

For the expectations and convergence arguments below, assume independent, identically distributed observations and finite absolute means of both potential outcomes. These are sampling and moment conditions, separate from the causal assumption.

4. The population proof: weighted controls recover the missing mean

Use $\mu_0(X) = \mathbb{E}[Y \mid X, D = 0]$ as shorthand for the control mean within a covariate group. It appears only to explain the proof; IPW does not require fitting an outcome regression.

First find the total weight

The indicator $1 - D$ selects controls: it is one for controls and zero for treated people. Given X, its mean is the control probability $1 - p_0(X)$. Also, $o_0(X)$ is fixed once X is fixed. Thus

$$\mathbb{E}[(1 - D)o_0(X) \mid X] = o_0(X)(1 - p_0(X)) = p_0(X).$$

Average over X. The law of iterated expectations says that averaging conditional means gives the overall mean, $\mathbb{E}[\mathbb{E}[A \mid X]] = \mathbb{E}[A]$. Consequently,

$$\begin{aligned}\mathbb{E}[(1 - D)o_0(X)] &= \mathbb{E}[\mathbb{E}[(1 - D)o_0(X) \mid X]] \\ &= \mathbb{E}[p_0(X)] \\ &= \mathbb{E}[\mathbb{E}[D \mid X]] \\ &= \mathbb{E}[D] = \pi.\end{aligned}$$

Here $p_0(X) = \mathbb{E}[D \mid X]$ because D is a binary indicator. Iterated expectations therefore shows that the average treatment probability is the treated share.

Then find the weighted outcome contribution

Within an X group, controls occur with probability $1 - p_0(X)$ and their mean outcome is $\mu_0(X)$. Treated people contribute zero to $(1 - D)Y$. Therefore,

$$\mathbb{E}[(1 - D)Y \mid X] = (1 - p_0(X))\mu_0(X).$$

Multiply by the odds, and average over X:

$$\begin{aligned}\mathbb{E}[(1 - D)o_0(X)Y] &= \mathbb{E}[\mathbb{E}[(1 - D)o_0(X)Y \mid X]] \\ &= \mathbb{E}[o_0(X)(1 - p_0(X))\mu_0(X)] \\ &= \mathbb{E}[p_0(X)\mu_0(X)].\end{aligned}$$

Why does this concern the treated? For covariate groups with $p_0(X) > 0$,

$$\begin{aligned}\mathbb{E}[Y(0) \mid X, D = 1] &= \mathbb{E}[Y(0) \mid X, D = 0] \\ &= \mathbb{E}[Y \mid X, D = 0] = \mu_0(X).\end{aligned}$$

The first equality uses unconfoundedness; the second uses $Y = Y(0)$ for controls. Thus the treated fraction times its mean untreated outcome is

$$\begin{aligned}\mathbb{E}[DY(0) \mid X] &= p_0(X)\mathbb{E}[Y(0) \mid X, D = 1] \\ &= p_0(X)\mu_0(X).\end{aligned}$$

Where $p_0(X) = 0$, both contributions are zero and no treated mean is needed. Averaging over X now gives

$$\begin{aligned}\mathbb{E}[p_0(X)\mu_0(X)] &= \mathbb{E}[\mathbb{E}[DY(0) \mid X]] \\ &= \mathbb{E}[DY(0)] \\ &= \pi\mathbb{E}[Y(0) \mid D = 1] = \pi\theta.\end{aligned}$$

The indicator D retains the treated people's untreated potential outcomes and assigns zero to everyone else; their overall contribution is their population share times their group mean.

Divide to obtain a mean

We have shown that the weighted outcome contribution is $\pi\theta$ and the total weight is π. Since $\pi > 0$,

$$\boxed{\frac{\mathbb{E}[(1-D)o_0(X)Y]}{\mathbb{E}[(1-D)o_0(X)]} = \frac{\pi\theta}{\pi} = \theta.}$$

This is identification: a function of observed data equals the missing mean.

5. From a population identity to something we can calculate

In data, replace the population means by sample averages and the unknown treatment probabilities by estimates. For observations $i = 1, \dots, n$:

1. Estimate $\hat{p}(X_i)$, the treatment probability given the observed characteristics.
2. Assign each control weight $\hat{o}_i = \hat{p}(X_i)/(1 - \hat{p}(X_i))$.
3. Compute the weighted control mean and subtract it from the treated mean.

Writing $n_1 = \sum_{i=1}^{n} D_i$ for the number treated, the estimator is

$$\boxed{\hat{\tau} = \underbrace{\frac{\sum_{i=1}^{n} D_i Y_i}{n_1}}_{\text{treated mean}} - \underbrace{\frac{\sum_{i=1}^{n} (1 - D_i)\hat{o}_i Y_i}{\sum_{i=1}^{n} (1 - D_i)\hat{o}_i}}_{\text{weighted control mean } \hat{\theta}}}.$$

This requires at least one treated observation, positive total control weight, and control propensity estimates below one. Under the consistency conditions below, the required denominators are positive with probability tending to one.

The control term is an ordinary weighted average. If the total control weight is S, each control receives normalized weight $\hat{o}_i/S$, and these normalized weights sum to one. This is often called normalized IPW or the Hájek form.

For example, two controls with outcomes $2, 8$ and weights $1, 3$ have weighted mean $(1 \cdot 2 + 3 \cdot 8)/(1+3) = 6.5$. Multiplying both weights by ten leaves the mean unchanged. It is their relative size that matters.

Why does the presentation also divide by the treated share?

At the population truth, the previous page showed $\mathbb{E}[(1 - D)o_0(X)] = \pi$. We can therefore also write

$$\tau = \frac{1}{\pi}\mathbb{E}\left[DY - (1 - D)o_0(X)Y\right].$$

Replacing π by n_1/n and expectations by sample averages gives another IPW estimator, one that divides both outcome sums by n_1.

These two sample formulas are not generally identical: the sum of estimated control weights need not equal n_1. Both are consistent under the conditions below. We use the normalized version throughout because each term is visibly a mean. Normalization by itself does not guarantee that the weighted controls' covariate composition exactly matches the treated in a finite sample.

Why odds weights, if the name says "inverse probability"?

For the ATT, $1/(1 - p_0(X))$ compensates for the chance of appearing among controls; the extra factor $p_0(X)$ targets the treated population. For the average effect in the *whole* population (ATE), the usual weights instead are $1/p_0(X)$ for treated people and $1/(1 - p_0(X))$ for controls. Those target a different population.

6. Consistency when the true propensities are known

Consistency means $\hat{\tau} \xrightarrow{p} \tau$: for any fixed tolerance, the probability that the estimation error exceeds that tolerance tends to zero as n grows. It does not mean the estimate equals τ in a particular sample.

First imagine that we know p_0 and use the true weights $o_0(X_i)$. This isolates ordinary sampling error from the error in estimating propensities.

The treated sample mean approaches the treated population mean

Divide its numerator and denominator by n:

$$\frac{\sum_i D_i Y_i}{\sum_i D_i} = \frac{n^{-1}\sum_i D_i Y_i}{n^{-1}\sum_i D_i}.$$

The **law of large numbers** says a sample average of i.i.d. observations with a finite absolute mean converges in probability to its expectation. Apply it twice:

$$\frac{1}{n}\sum_i D_i Y_i \xrightarrow{p} \mathbb{E}[DY] = \pi\mathbb{E}[Y \mid D = 1], \qquad \frac{1}{n}\sum_i D_i \xrightarrow{p} \pi.$$

Because the denominator approaches a positive number, the ratio approaches $\mathbb{E}[Y \mid D = 1]$. Taking ratios preserves convergence when the limiting denominator is nonzero.

The weighted control mean approaches the missing mean

The same reasoning applies to its two sample averages:

$$\frac{1}{n}\sum_i (1 - D_i) o_0(X_i) Y_i \xrightarrow{p} \mathbb{E}[(1 - D) o_0(X) Y] = \pi\theta,$$

$$\frac{1}{n}\sum_i (1 - D_i) o_0(X_i) \xrightarrow{p} \mathbb{E}[(1 - D) o_0(X)] = \pi.$$

The law of large numbers applies because overlap bounds the weights and the outcomes have finite absolute means. The equalities on the right are precisely what the population proof established. Dividing gives

$$\hat{\theta}_{\text{true weights}} \xrightarrow{p} \frac{\pi\theta}{\pi} = \theta.$$

Subtract the limits

Differences also preserve convergence, so

$$\hat{\tau}_{\text{true weights}} \xrightarrow{p} \mathbb{E}[Y \mid D = 1] - \theta = \tau.$$

This is the bridge: **identification gives the correct population limit; the law of large numbers takes the sample estimator to that limit.** We must still show that estimating the weights does not change it.

7. Why estimating the propensities need not change the limit

Estimated weights use the data, so we cannot simply treat their weighted outcomes as i.i.d. observations. Instead, compare the fitted-weight sums with the true-weight sums we have already proved converge.

Here is a transparent set of **sufficient conditions**. Suppose $0 \le \hat{p}(x) \le 1-\epsilon$ and $p_0(x) \le 1-\epsilon$ throughout the covariate support, and suppose the largest propensity estimation error tends to zero:

$$r_n := \sup_x |\hat{p}(x) - p_0(x)| \xrightarrow{p} 0.$$

This is uniform consistency: eventually, even the largest error is small with high probability. It is stronger than necessary, but lets us give a short, complete proof, including when propensities are fitted on the same sample.

Small propensity errors imply small weight errors

Subtract the two odds fractions using a common denominator:

$$\left| \frac{\hat{p}(x)}{1-\hat{p}(x)} - \frac{p_0(x)}{1-p_0(x)} \right| = \frac{|\hat{p}(x) - p_0(x)|}{(1-\hat{p}(x))(1-p_0(x))} \le \frac{r_n}{\epsilon^2}.$$

The two denominators are each at least ϵ. Thus the largest weight error, $\Delta_n := \sup_x |\hat{o}(x) - o_0(x)|$, also converges in probability to zero.

Small weight errors imply small errors in the sample averages

The change in the weighted outcome average is bounded by

$$\left| \frac{1}{n}\sum_i (1-D_i)\hat{o}_i Y_i - \frac{1}{n}\sum_i (1-D_i) o_0(X_i) Y_i \right| \le \Delta_n \frac{1}{n}\sum_i (1-D_i)|Y_i| \xrightarrow{p} 0.$$

Why does the product vanish? Its first factor approaches zero; its second approaches the finite number $\mathbb{E}[(1-D)|Y|]$ by the law of large numbers. The bound holds for each realized dataset, without requiring independence between the two factors.

For the average total weight, the corresponding bound is even simpler:

$$\left| \frac{1}{n}\sum_i (1-D_i)\hat{o}_i - \frac{1}{n}\sum_i (1-D_i) o_0(X_i) \right| \le \Delta_n \frac{1}{n}\sum_i (1-D_i) \le \Delta_n \xrightarrow{p} 0.$$

Estimated-weight averages therefore have the same limits as true-weight averages: the numerator approaches $\pi\theta$ and the denominator approaches $\pi > 0$. Their ratio approaches θ, and subtracting it from the treated mean gives $\hat{\tau} \xrightarrow{p} \tau$.

8. Making the conditions concrete

In the example, estimate treatment probabilities by group frequencies

With $X \in \{0, 1, 2\}$, let n_x be the number of observations in group x and n_{1x} its treated count. Estimate

$$\hat{p}(x) = \frac{n_{1x}}{n_x}.$$

Both counts divided by n are sample averages of indicators. They converge to $\mathbb{P}(D = 1, X = x)$ and $\mathbb{P}(X = x) > 0$. Their ratio therefore converges to $p_0(x)$. There are only three groups, so the maximum of their three absolute estimation errors also approaches zero: this gives uniform consistency.

The true probabilities are at most 3/4, so, with probability approaching one, all three estimated probabilities are at most 7/8. The proof on the previous page then applies with $\epsilon = 1/8$ on that event; events whose probabilities vanish do not alter convergence in probability.

There is also a useful exact calculation. If group x contains n_{0x} controls, then, whenever the relevant counts are positive,

$$\hat{o}(x) = \frac{n_{1x}/n_x}{n_{0x}/n_x} = \frac{n_{1x}}{n_{0x}}, \qquad n_{0x}\hat{o}(x) = n_{1x}.$$

Writing $\bar{Y}_{0x}$ for the observed control mean in group x, this implies

$$\hat{\theta} = \frac{\sum_x n_{1x}\bar{Y}_{0x}}{\sum_x n_{1x}}.$$

It is the control group means averaged using the *treated* group shares. This particular frequency estimator balances the groups exactly; a general fitted propensity model need not. Groups with treated people but no controls prevent this calculation. Under the stated population conditions and finitely many positive-probability groups, that problem disappears with probability tending to one.

What the result does and does not promise

The propensity estimates must be adequate. A fitted model does not automatically converge to p_0. If it converges to a wrong function p^*, the weighted control mean, under analogous convergence conditions, approaches

$$\frac{\mathbb{E}[(1-D)o^*(X)Y]}{\mathbb{E}[(1-D)o^*(X)]}, \qquad o^*(X) = \frac{p^*(X)}{1-p^*(X)},$$

provided these expectations exist and the denominator is positive. That ratio need not equal θ: normalization makes it a mean, not necessarily the right mean.

Overlap and the causal assumption still matter. Probabilities near one give large weights and noisy estimates. Capping weights can change the limiting answer if the cap changes the true weights. Unmeasured differences in untreated outcomes within X groups cannot be repaired by these weights.

The argument requires both links: the assumptions make weighted controls represent the treated counterfactual, and accurate weight estimation plus averaging makes the sample estimate approach it. No finite-sample unbiasedness is claimed.