Potential Outcomes Framework

Statistics · Module 11 — Causality

Node 11.1

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.1.

The potential outcomes framework (Rubin causal model) is the foundational language for defining causal quantities in statistics. It formalizes what it means for a treatment to cause an effect by indexing outcomes to hypothetical interventions, making explicit the gap between what is observed and what is needed for causal claims. This node defines the estimands (ATE, ATT, CATE), the key assumptions (SUTVA, ignorability, overlap), and the fundamental problem of causal inference. It enables the graphical approach (Node 11.2), identification theory (Node 11.3), and experimental design (Node 11.4).

Formal Definition

For each unit $i \in \{1, \ldots, n\}$ and treatment value $t \in \mathcal{T}$, the **potential outcome** $Y_i(t)$ is the value of the outcome that unit i would exhibit if assigned to treatment t.

For binary treatment $T_i \in \{0, 1\}$:

- $Y_i(1)$: potential outcome under treatment.
- $Y_i(0)$: potential outcome under control.

The **individual treatment effect** (ITE) is

$$\tau_i = Y_i(1) - Y_i(0).$$

The **observed outcome** is

$$Y_i^{\text{obs}} = T_i\, Y_i(1) + (1 - T_i)\, Y_i(0) = Y_i(T_i).$$

Intuition

Causation requires comparing two states of the world: what happened under treatment versus what would have happened under control. The potential outcomes framework takes this literally—it defines both outcomes for every unit, even though only one is ever observed. The unobserved outcome is the **counterfactual**. The fundamental problem of causal inference is that we can never observe both potential outcomes for the same unit, so individual causal effects are never directly identified from data. All causal inference proceeds by making assumptions that link the observable data to the unobservable counterfactual distribution.

Structural Role

Dependencies: Builds on probability spaces (Node 1.3) and estimation theory (Node 3.1). **Enables:** DAGs (Node 11.2) provide a graphical language for the same causal relationships; identifiability (Node 11.3) gives conditions under which causal effects can be recovered; randomized experiments (Node 11.4) achieve identification by design.

The potential outcomes framework and the DAG/structural equation framework (Node 11.2) are complementary formalizations of causality. Potential outcomes define the estimands; DAGs encode the causal structure. The backdoor criterion (Node 11.3) translates between the two.

Mathematical Development

The Fundamental Problem of Causal Inference

For each unit i, only one of $Y_i(0)$ and $Y_i(1)$ is observed. The counterfactual $Y_i(1 - T_i)$ is missing. Therefore:

$$\tau_i = Y_i(1) - Y_i(0) \text{ is never identified from data for any individual } i.$$

Causal inference redirects attention from individual effects to **average effects**, which can be identified under assumptions.

Causal Estimands

Average Treatment Effect (ATE):

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

Average Treatment Effect on the Treated (ATT):

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

Conditional Average Treatment Effect (CATE):

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

The ATE is the population-level average; the ATT restricts to the treated subpopulation; the CATE allows treatment effect heterogeneity across covariate values.

SUTVA: Stable Unit Treatment Value Assumption

SUTVA requires:

1. **No interference:** $Y_i(t_1, \ldots, t_n) = Y_i(t_i)$. Unit i's potential outcome depends only on its own treatment, not on the treatments of others.
2. **No hidden variations of treatment:** There is only one version of each treatment level. If $T_i = t$, then $Y_i^{\text{obs}} = Y_i(t)$ regardless of how treatment t was administered.

SUTVA is required for potential outcomes to be well-defined as individual-level quantities. It fails in settings with network effects (vaccination, social programs), where one unit's treatment affects another's outcome.

Ignorability (Unconfoundedness)

Strong ignorability given covariates X:

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

This states that, conditional on observed covariates, treatment assignment is independent of potential outcomes. It is the assumption that there are no unmeasured confounders.

Proof that ignorability identifies the ATE. Under ignorability:

$$\mathbb{E}[Y(t) \mid X] = \mathbb{E}[Y(t) \mid T = t, X] = \mathbb{E}[Y^{\text{obs}} \mid T = t, X],$$

where the first equality uses independence and the second uses consistency ($Y^{\text{obs}} = Y(T)$). Therefore:

$$\tau(x) = \mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x],$$

and integrating over X:

$$\tau_{\text{ATE}} = \mathbb{E}_X[\tau(X)] = \mathbb{E}_X\big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\big].$$

Overlap (Positivity)

The **overlap** condition requires

$$0 < e(x) = \mathbb{P}(T = 1 \mid X = x) < 1 \quad \text{for all } x \text{ in the support of } X.$$

Without overlap, some covariate values have no treated (or no control) units, and $\mathbb{E}[Y \mid T = t, X = x]$ is undefined. Overlap ensures every subpopulation contains both treated and control units.

Selection Bias Decomposition

The naive estimator $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$ generally does not equal the ATE:

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \tau_{\text{ATE}} + \underbrace{\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]}_{\text{Selection bias}}.$$

Selection bias is the difference in baseline (control) outcomes between treated and control groups. It vanishes under ignorability (or randomization) but is generically nonzero in observational studies.

Derivation. Write $\mathbb{E}[Y \mid T = 1] = \mathbb{E}[Y(1) \mid T = 1]$ and $\mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y(0) \mid T = 0]$. Add and subtract $\mathbb{E}[Y(0) \mid T = 1]$:

$$\mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid T = 1]}_{\tau_{\text{ATT}}} + \underbrace{\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]}_{\text{Selection bias}}.$$

If additionally $\tau_{\text{ATT}} = \tau_{\text{ATE}}$ (homogeneous effects), the decomposition simplifies to the formula above.

Propensity Score

The **propensity score** $e(x) = \mathbb{P}(T = 1 \mid X = x)$ is a balancing score.

Theorem (Rosenbaum–Rubin, 1983). Under ignorability given X,

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

Proof. It suffices to show $\mathbb{P}(T = 1 \mid Y(0), Y(1), e(X)) = e(X)$. By the law of iterated expectations:

$$\mathbb{P}(T = 1 \mid Y(0), Y(1), e(X)) = \mathbb{E}[\mathbb{P}(T = 1 \mid Y(0), Y(1), X) \mid Y(0), Y(1), e(X)].$$

By ignorability, $\mathbb{P}(T = 1 \mid Y(0), Y(1), X) = \mathbb{P}(T = 1 \mid X) = e(X)$. Since $e(X)$ is measurable with respect to $\sigma(e(X))$, the outer expectation gives $e(X)$.

This reduces the problem from conditioning on the high-dimensional X to conditioning on the scalar $e(X)$, enabling practical estimation (Node 11.5).

Worked Examples

Example 1: Selection Bias in a Job Training Program

A job training program is offered to unemployed workers. Outcome: earnings Y one year later. Treatment: $T = 1$ if enrolled. Observed: $\bar{Y}_1 = \$22{,}000$, $\bar{Y}_0 = \$18{,}000$.

Naive estimate: $\hat{\tau}^{\text{naive}} = \$4{,}000$.

But workers who enrolled may be more motivated (higher $Y(0)$). Suppose the true selection bias is $\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] = \$2{,}500$. Then the true ATT is $\$4{,}000 - \$2{,}500 = \$1{,}500$.

Under ignorability given education and prior earnings X, the ATE is identified by the adjustment formula. If X does not capture motivation, ignorability fails and the naive estimate is biased.

Example 2: Computing ATE Under Ignorability

Binary treatment, binary covariate $X \in \{0, 1\}$ with $\mathbb{P}(X = 1) = 0.6$.

X	$\mathbb{E}[Y \mid T = 1, X]$	$\mathbb{E}[Y \mid T = 0, X]$	$\tau(X)$
0	5.0	3.0	2.0
1	8.0	5.5	2.5

$\tau_{\text{ATE}} = 0.4 \times 2.0 + 0.6 \times 2.5 = 0.8 + 1.5 = 2.3$.

Naive difference in means (without conditioning): suppose $\mathbb{P}(T = 1 \mid X = 1) = 0.7$, $\mathbb{P}(T = 1 \mid X = 0) = 0.3$. Then treated units are disproportionately $X = 1$.

$\mathbb{E}[Y \mid T = 1] = \frac{0.3 \times 0.4 \times 5.0 + 0.7 \times 0.6 \times 8.0}{0.3 \times 0.4 + 0.7 \times 0.6} = \frac{0.6 + 3.36}{0.12 + 0.42} = \frac{3.96}{0.54} = 7.33$.

$\mathbb{E}[Y \mid T = 0] = \frac{0.7 \times 0.4 \times 3.0 + 0.3 \times 0.6 \times 5.5}{0.7 \times 0.4 + 0.3 \times 0.6} = \frac{0.84 + 0.99}{0.28 + 0.18} = \frac{1.83}{0.46} = 3.98$.

Naive estimate: $7.33 - 3.98 = 3.35$, which overstates the true ATE of 2.3 because treated units have higher X values (confounding).

Cross-Links

AI. Counterfactual reasoning in reinforcement learning is the sequential decision analogue: the agent's potential outcomes under different action sequences are the value functions, and the fundamental problem of causal inference becomes the exploration-exploitation tradeoff.

Economics. The Roy model is a structural potential outcomes model with self-selection: agents choose treatment based on their comparative advantage, generating endogenous selection. The potential outcomes framework underpins the credibility revolution in empirical economics (Angrist, Imbens, Rubin).

Linear Algebra. The potential outcomes for n units form the matrix $\mathbf{Y} = [Y_i(t)]_{n \times |\mathcal{T}|}$, of which only one entry per row is observed. The fundamental problem is a missing data problem with a structured missingness pattern determined by the treatment assignment vector.

Quantum Mechanics. In quantum mechanics, counterfactual reasoning arises in the analysis of Bell inequalities: the potential outcomes framework applied to measurement settings formalizes the assumption of local realism, and Bell's theorem shows that no assignment of potential outcomes can reproduce quantum correlations.

Structural Compression

Invariant

Causal estimands are defined on the joint distribution of potential outcomes $(Y(0), Y(1))$; identification requires assumptions linking this unobserved joint distribution to the observed data distribution $p(Y^{\text{obs}}, T, X)$.

Mechanism

Potential outcomes index all counterfactual worlds; observed data indexes only one world per unit. Ignorability $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X$ equates the observational conditional with the interventional distribution. Overlap ensures every covariate stratum has both treated and control units. The propensity score reduces the conditioning dimension from X to the scalar $e(X)$.

Cross-System Echo

The potential outcomes framework is the frequentist formalization of causality; in the structural equation modeling tradition (DAGs, Node 11.2), the same causal quantities are defined via interventions $do(T = t)$. In economics, the Roy model and the Heckman selection model are structural potential outcomes models with endogenous treatment choice. In AI, off-policy evaluation is the sequential analogue of the fundamental problem of causal inference.

Exercises

1. Show that $\tau_{\text{ATE}} = \tau_{\text{ATT}}$ if and only if $\mathbb{E}[Y(1) - Y(0) \mid T = 1] = \mathbb{E}[Y(1) - Y(0) \mid T = 0]$, i.e., the treatment effect is the same for treated and untreated subpopulations.
2. Prove the selection bias decomposition: $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \tau_{\text{ATT}} + (\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0])$.
3. Under ignorability and overlap, derive the inverse probability weighting (IPW) representation: $\tau_{\text{ATE}} = \mathbb{E}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]$.
4. Show that the propensity score is a balancing score: $\mathbb{E}[T \mid X, e(X)] = e(X)$, and use this to prove $\mathbb{E}[h(X) \mid T = 1, e(X)] = \mathbb{E}[h(X) \mid T = 0, e(X)]$ for any function h under ignorability.
5. Construct a numerical example with $n = 4$ units where the observed difference in means equals the ATE but individual treatment effects are heterogeneous, illustrating the ecological fallacy in reverse.

6. Consider a setting where SUTVA is violated (e.g., vaccination with herd immunity). Define the potential outcomes $Y_i(\mathbf{t})$ as a function of the full treatment vector $\mathbf{t} = (t_1, \ldots, t_n)$. How many potential outcomes does each unit have? Why does this make identification dramatically harder?

7. Argue, structurally, why the fundamental problem of causal inference—the impossibility of observing both potential outcomes for the same unit—is not a limitation of experimental design but a definitional feature of individual-level causation, and identify the assumptions that convert this individual-level impossibility into a population-level identification result.

Statistics · Module 11 — Causality · Node 11.1
Concepts: random-variables, expectation-value, conditional-probability
Prerequisites: 1.3, 3.1
Enables: 11.2, 11.3, 11.4

hbar.university — own.your.cognition