Causal Inference

Lecture 01: Potential Outcomes and Identification

Joao Alipio-Correa

Agenda

1. Introduction: The Running Example
2. Mathematical Preliminaries: Expectations
3. Potential Outcomes
4. The Assumption Package
5. Core Estimands: ATE and ATT
6. Why the Naive Difference Fails
7. Identification: From Estimands to Data
8. What Fails: Where Derivations Break

Introduction: The Running Example

Causal vs. Associational Claims

When we say "canvassing increases turnout," we claim something **beyond description**.

- **Associational:** Turnout is higher among canvassed voters
- **Causal:** Turnout *would change* if we intervened to canvass (or not canvass) the same voters

The problem: campaigns do not canvass at random.

$\Rightarrow$ Without making the causal question explicit, we risk answering a different (associational) question.

The Running Example: Notation

We study a **campaign canvassing program** and **voter turnout**.

Index voters by $i \in \{1, \ldots, n\}$.

Treatment:

$$A_i \in \{0, 1\}, \qquad A_i = 1 \text{ (canvassed)}, \quad A_i = 0 \text{ (not canvassed)}$$

Outcome:

$$Y_i \in \{0, 1\}, \qquad Y_i = 1 \text{ (voted)}, \quad Y_i = 0 \text{ (did not vote)}$$

Right now, (A_i, Y_i) are just observed variables. The causal content begins when we define **potential outcomes**.

Mathematical Preliminaries: Expectations

Expectation as a Population Average

$\mathbb{E}[Y]$ is the **average value** of Y in the population.

For binary $Y \in \{0, 1\}$:

$$\mathbb{E}[Y] = P(Y = 1) = \text{turnout rate in the population}$$

More generally:

$$\mathbb{E}[Y] = \sum_{y} y \cdot P(Y = y)$$

Key interpretation: $\mathbb{E}[Y]$ is the long-run average of Y.

Conditional Expectation: Averages Within Subgroups

$\mathbb{E}[Y \mid A = 1]$ = average outcome among units with $A = 1$.

For binary Y:

$$\mathbb{E}[Y \mid A = 1] = P(Y = 1 \mid A = 1) = \text{turnout rate among canvassed voters}$$

Think of it as "a mean inside a slice of the data."

Example: Computing Conditional Expectations

Canvassing Study

1000 voters: 300 canvassed ($A = 1$), 700 not canvassed ($A = 0$).

Among canvassed: 180 voted. Among non-canvassed: 280 voted.

$$\mathbb{E}[Y \mid A = 1] = \frac{180}{300} = 0.60, \qquad \mathbb{E}[Y \mid A = 0] = \frac{280}{700} = 0.40$$

Overall turnout:

$$\mathbb{E}[Y] = \frac{180 + 280}{1000} = 0.46$$

The Law of Iterated Expectations

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid A]\big] = P(A=1)\,\mathbb{E}[Y \mid A=1] + P(A=0)\,\mathbb{E}[Y \mid A=0]$$

In words: Average within groups, then average across groups (weighted by group size) = overall average.

Verification:

$$\mathbb{E}[Y] = 0.30 \times 0.60 + 0.70 \times 0.40 = 0.18 + 0.28 = 0.46 \quad \checkmark$$

Iterated Expectations with Covariates

Often we condition on covariates X (age, party, prior turnout, etc.).

$$\mathbb{E}[Y] = \mathbb{E}\Big[\mathbb{E}[Y \mid X]\Big]$$

Two-step averaging:

1. Average Y within each stratum of X
2. Average those stratum means over the distribution of X

This will be central when we write:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big[\mathbb{E}[Y \mid X, A = a]\Big]$$

Algebra Trick: Pulling Out Indicators

Because $A \in \{0, 1\}$, multiplying by A **selects the treated group**:

$$\mathbb{E}[A \cdot Y] = P(A = 1) \cdot \mathbb{E}[Y \mid A = 1]$$

Intuition: $A \cdot Y = Y$ for treated units, $= 0$ for untreated units.

Similarly, $(1 - A)$ selects the control group:

$$\mathbb{E}[(1 - A) \cdot Y] = P(A = 0) \cdot \mathbb{E}[Y \mid A = 0]$$

Potential Outcomes

Defining Potential Outcomes

For each voter i, define **two potential outcomes**:

$$Y_i(1) \quad \text{and} \quad Y_i(0)$$

- $Y_i(1)$ = turnout for voter i if we **set** $A_i = 1$ (canvass them)
- $Y_i(0)$ = turnout for voter i if we **set** $A_i = 0$ (do not canvass)

Think of each voter carrying two "response cards":

Card 1:	$Y_i(1)$ (turnout if canvassed)
Card 2:	$Y_i(0)$ (turnout if not canvassed)

The Fundamental Problem of Causal Inference

The **individual causal effect**:

$$\tau_i \equiv Y_i(1) - Y_i(0)$$

The problem: In the realized world, voter i is either canvassed or not.

$\Rightarrow$ We only ever observe **one** potential outcome per voter.
$\Rightarrow$ The other is a **missing counterfactual**.

Causal inference is a missing-data problem: one potential outcome per unit is systematically unobserved.

Example: Response Cards for Four Voters

Voter i	$Y_i(1)$	$Y_i(0)$	τ_i	Interpretation
1	1	0	1	persuadable voter
2	1	1	0	always-voter
3	0	0	0	never-voter
4	0	1	-1	backfire voter

Voters respond **differently** to the same intervention.

Two potential outcomes per voter, but data will never show both.

The Assumption Package

Why We Need Assumptions

Potential outcomes $(Y(1), Y(0))$ are not directly observed.

To connect them to observed data (Y, A, X), we need **assumptions**.

Each assumption serves a specific role in identification proofs:

- **Consistency** — links counterfactuals to observed outcomes
- **SUTVA** — makes potential outcomes well-defined
- **Exchangeability** — eliminates selection bias
- **Positivity** — ensures we can condition on relevant strata

Assumption: Consistency

If unit i receives treatment a, then the observed outcome equals the potential outcome under a:

$$A_i = a \implies Y_i = Y_i(a), \qquad a \in \{0, 1\}$$

What it allows: Replace $Y(a)$ with observed Y inside the stratum where $A = a$.

At the expectation level:

$$\mathbb{E}[Y \mid A = 1] = \mathbb{E}[Y(1) \mid A = 1], \qquad \mathbb{E}[Y \mid A = 0] = \mathbb{E}[Y(0) \mid A = 0]$$

The Selector Identity

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

Because A_i is binary:

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

Verification:

$$\text{If } A_i = 1: \quad Y_i = 1 \cdot Y_i(1) + 0 \cdot Y_i(0) = Y_i(1)$$

$$\text{If } A_i = 0: \quad Y_i = 0 \cdot Y_i(1) + 1 \cdot Y_i(0) = Y_i(0)$$

The observed dataset reveals **exactly one response card** per voter.

Example: What the Data Reveal

Suppose the campaign canvasses Maria and Carlos, but not Juan and Ana.

Voter	$Y_i(1)$	$Y_i(0)$	Treatment	Observed Y_i	Missing
Maria	1	0	Canvassed	$Y(1) = 1$ (voted)	$Y(0)$
Juan	1	1	Not canvassed	$Y(0) = 1$ (voted)	$Y(1)$
Carlos	0	0	Canvassed	$Y(1) = 0$ (didn't vote)	$Y(0)$
Ana	0	1	Not canvassed	$Y(0) = 1$ (voted)	$Y(1)$

Key insight: For Maria, we see she voted after being canvassed. But we *cannot* know if she would have voted without canvassing—that counterfactual is forever missing.

The missing counterfactual is **structural**, not a sampling issue.

SUTVA: Stable Unit Treatment Value Assumption

SUTVA guarantees that $Y_i(1)$ and $Y_i(0)$ are **well-defined**.

(i) No hidden versions of treatment

- If "canvassing" means different things (2-min vs. 15-min visit), $Y_i(1)$ is ambiguous
- More honest notation: $Y_i(1, v)$ where v indexes versions

(ii) No interference between units

- Voter i's outcome depends only on i's own treatment
- If canvassing i affects roommate j, we need $Y_i(\mathsf{a})$ not $Y_i(a_i)$

Exchangeability: No Unmeasured Confounding

Unconditional exchangeability (ideal benchmark):

$$A \perp\!\!\!\perp Y(a), \qquad a \in \{0, 1\}$$

Treatment assignment is independent of potential outcomes.
Example: In a *randomized experiment*, we flip a coin to decide who gets canvassed. The coin doesn't "know" who would respond to canvassing. $\Rightarrow$ Treated and untreated groups are comparable.

Conditional exchangeability (observational target):

$$A \perp\!\!\!\perp Y(a) \mid X, \qquad a \in \{0, 1\}$$

Within strata of X, treatment is as-if random.
Example: Campaigns target based on *prior voting* (X). Among voters who all voted last time ($X = 1$), whether someone gets canvassed is unrelated to their potential turnout. $\Rightarrow$ Compare like with like.

Positivity (Overlap)

Assumption: Positivity

For all relevant covariate values x:

$$0 < P(A = a \mid X = x) < 1, \qquad a \in \{0, 1\}$$

In words: Both treated and untreated units must exist in each stratum we rely on.

Example of failure: Campaign never canvasses voters under 25.
$\Rightarrow P(A = 1 \mid X = \text{“under 25”}) = 0$
$\Rightarrow$ Cannot estimate $\mathbb{E}[Y \mid X = \text{“under 25”}, A = 1]$

Summary: The Assumption Package

Assumption	What it allows
Consistency	Replace $Y(a)$ with Y in stratum $A = a$
SUTVA	$Y_i(a)$ is well-defined (no versions, no interference)
Exchangeability	Remove conditioning on A from potential outcome means
Positivity	Conditional expectations are defined for relevant strata

Core Estimands: ATE and ATT

The Individual Causal Effect

For voter i:

$$\tau_i \equiv Y_i(1) - Y_i(0)$$

Problem: τ_i is never observable.

- If $A_i = 1$: we observe $Y_i(1)$, but $Y_i(0)$ is missing
- If $A_i = 0$: we observe $Y_i(0)$, but $Y_i(1)$ is missing

Example: Maria was canvassed and voted ($Y = 1$). Her individual effect is $\tau_{\text{Maria}} = Y_{\text{Maria}}(1) - Y_{\text{Maria}}(0) = 1 - ?$
We observe $Y_{\text{Maria}}(1) = 1$. But $Y_{\text{Maria}}(0)$ is unknowable—we cannot rewind time and not canvass her.

$\Rightarrow$ Focus on **average** causal effects.

The Average Treatment Effect (ATE)

Definition: ATE

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

In words: If we canvassed *everyone* vs. *no one*, how would average turnout differ?

- Averages over **all** voters (persuadables, always-voters, never-voters, etc.)
- The ATE compares two **hypothetical worlds**: one where everyone is treated, one where no one is. Neither world needs to actually exist in our data.
- A property of the **population**, not a regression coefficient

But wait: What if we only care about the voters who were *actually* canvassed?

The Average Treatment Effect on the Treated (ATT)

Definition: ATT

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

In words: Among voters who *were* canvassed, what is the average causal effect?

- Averages only over the **treated** subgroup
- Policy-relevant when treatment is targeted, not universal

The fundamental problem reappears: To compute ATT, we need $\mathbb{E}[Y(0) \mid A = 1]$—the average outcome the treated would have had *if they had not been treated*.
But treated units were treated, so we never observe their $Y(0)$!

Example: Seeing ATE vs. ATT

Note: This table shows *both* potential outcomes—something we can **never observe** in real data. This is a teaching device only.

Voter	A_i	$Y_i(1)$	$Y_i(0)$	Observed Y_i	τ_i
1	1	1	1	1	0
2	1	1	0	1	1
3	1	0	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	0	1

$$\text{ATE} = \frac{0+1+0+1+0+1}{6} = 0.50, \qquad \text{ATT} = \frac{0+1+0}{3} = 0.33$$

Different estimands $\Rightarrow$ different answers (and that's fine!).

The Naive Difference Is Neither ATE nor ATT

From the same table:

$$\mathbb{E}[Y \mid A=1] - \mathbb{E}[Y \mid A=0] = \frac{1+1+0}{3} - \frac{0+0+0}{3} = \frac{2}{3} \approx 0.67$$

This is **larger** than both ATE (0.50) and ATT (0.33).

Why? Treated voters have higher baseline turnout:

$$\mathbb{E}[Y(0) \mid A=1] = \frac{1+0+0}{3} = 0.33, \qquad \mathbb{E}[Y(0) \mid A=0] = \frac{0+0+0}{3} = 0$$

The naive comparison mixes **causal effects** with **selection bias**.

Why the Naive Difference Fails

Rewriting the Naive Difference

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

Start with the naive comparison:

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$$

By consistency:

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

Problem: This compares $Y(1)$ for treated with $Y(0)$ for untreated.

These are **different groups**!

The Canonical Decomposition (Step 1)

Recall ATT: $\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid A = 1] = \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]$

We have:

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

Trick: Add and subtract the same term: $\mathbb{E}[Y(0) \mid A = 1]$

$$= \mathbb{E}[Y(1) \mid A = 1] \; -\mathbb{E}[Y(0) \mid A = 1] + \mathbb{E}[Y(0) \mid A = 1] \; - \mathbb{E}[Y(0) \mid A = 0]$$

The gold terms sum to zero—we haven't changed anything, just rearranged.

The Canonical Decomposition (Step 2)

Recall ATT: $\tau_{\text{ATT}} = \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]$

Now regroup the terms:

$$= \underbrace{\Big(\mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]\Big)}_{\text{this is exactly } \tau_{\text{ATT}}!} + \underbrace{\Big(\mathbb{E}[Y(0) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]\Big)}_{\text{selection bias}}$$

The Canonical Decomposition (Step 2)

Recall ATT: $\tau_{\text{ATT}} = \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]$

Now regroup the terms:

$$= \underbrace{\Big(\mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1]\Big)}_{\text{this is exactly } \tau_{\text{ATT}}!} + \underbrace{\Big(\mathbb{E}[Y(0) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]\Big)}_{\text{selection bias}}$$

$$\underbrace{\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]}_{\text{naive difference}} = \underbrace{\tau_{\text{ATT}}}_{\text{causal}} + \underbrace{\mathbb{E}[Y(0) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]}_{\text{selection bias}}$$

Interpreting the Decomposition

$$\text{Naive difference} = \text{ATT} + \text{Selection bias}$$

ATT (causal): Effect of canvassing on those who were canvassed.

Selection bias (non-causal): Baseline turnout difference between groups *even if no one were canvassed*.

Example: Campaigns target high-propensity voters. These voters would vote at higher rates *even without canvassing*.
$\Rightarrow \mathbb{E}[Y(0) \mid A = 1] > \mathbb{E}[Y(0) \mid A = 0]$
$\Rightarrow$ Selection bias is **positive**, inflating the naive estimate.

Consequence: Naive comparisons overstate the causal effect when campaigns target responsive voters.

Numerical Verification

Voter	A_i	$Y_i(1)$	$Y_i(0)$	Observed Y_i	τ_i
1	1	1	1	1	0
2	1	1	0	1	1
3	1	0	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	0	1

- ATT $= (0 + 1 + 0)/3 = 0.33$
- $\mathbb{E}[Y(0) \mid A = 1] = (1 + 0 + 0)/3 = 0.33, \quad \mathbb{E}[Y(0) \mid A = 0] = (0 + 0 + 0)/3 = 0$
- Selection bias $= 0.33 - 0 = 0.33$

$$\text{Naive difference} = \text{ATT} + \text{Selection bias} = 0.33 + 0.33 = 0.67 \quad \checkmark$$

What the Observed Data Reveal

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From selector identity: $\mathbb{E}[Y] = P(A = 1) \cdot \mathbb{E}[Y(1) \mid A = 1] + P(A = 0) \cdot \mathbb{E}[Y(0) \mid A = 0]$

What we observe: $\mathbb{E}[Y(1) \mid A = 1]$ and $\mathbb{E}[Y(0) \mid A = 0]$

Why? Among treated ($A = 1$), we see $Y = Y(1)$. Among untreated ($A = 0$), we see $Y = Y(0)$.

What we need but don't observe:

- $\mathbb{E}[Y(0) \mid A = 1]$ — What would treated voters' turnout be *if not treated*?
  We can't know: they *were* treated, so we only see their $Y(1)$.
- $\mathbb{E}[Y(1) \mid A = 0]$ — What would untreated voters' turnout be *if treated*?
  We can't know: they *weren't* treated, so we only see their $Y(0)$.

Identification: From Estimands to Data

What Is Identification?

Identification = expressing a causal estimand *entirely* in terms of observed data.

The task: Rewrite $\tau = \mathbb{E}[Y(1) - Y(0)]$ using only $P(Y, A, X)$.

(Here: Y = observed turnout, A = treatment received, X = pre-treatment covariates)

- "Identified" ≠ "estimated"
- "Identified" ≠ "we have a good estimator"

Meaning: "If two causal worlds produce the same $P(Y, A, X)$, they imply the same τ."

Translation: Different underlying causal structures could generate identical data. If our assumptions rule out alternatives, only one value of τ is compatible with what we observe.

Identification Under Randomization: Setup

Recall: $A \perp\!\!\!\perp Y(a)$ means $\mathbb{E}[Y(a) \mid A = a] = \mathbb{E}[Y(a)]$ (Unconditional Exchangeability)

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

Goal: Identify $\mathbb{E}[Y(a)]$—the mean potential outcome under treatment a.

This is a **counterfactual** quantity: the average turnout if *everyone* received treatment a.

Why it's hard: We don't observe $Y(a)$ for everyone—only for those with $A = a$.

Strategy: Use assumptions to rewrite $\mathbb{E}[Y(a)]$ in terms of observed data.

Identification Under Randomization: Step 1

Recall: $A \perp\!\!\!\perp Y(a)$ means $\mathbb{E}[Y(a) \mid A = a] = \mathbb{E}[Y(a)]$ (Unconditional Exchangeability)

Goal: Identify $\mathbb{E}[Y(a)]$.

Start with what we want:

$$\mathbb{E}[Y(a)]$$

Identification Under Randomization: Step 1

Recall: $A \perp\!\!\!\perp Y(a)$ means $\mathbb{E}[Y(a) \mid A = a] = \mathbb{E}[Y(a)]$ (Unconditional Exchangeability)

Goal: Identify $\mathbb{E}[Y(a)]$.

Start with what we want:

$$\mathbb{E}[Y(a)]$$

Apply **unconditional exchangeability**: $A \perp\!\!\!\perp Y(a)$

Since treatment is independent of potential outcomes, conditioning on A doesn't change the mean:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a) \mid A = a]$$

Identification Under Randomization: Step 2

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a) \mid A = a]$$

Identification Under Randomization: Step 2

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a) \mid A = a]$$

Apply **consistency**: Among units with $A = a$, we have $Y = Y(a)$.

$$\mathbb{E}[Y(a) \mid A = a] = \mathbb{E}[Y \mid A = a]$$

Identification Under Randomization: Step 2

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y(a) \mid A = a]$$

Apply **consistency**: Among units with $A = a$, we have $Y = Y(a)$.

$$\mathbb{E}[Y(a) \mid A = a] = \mathbb{E}[Y \mid A = a]$$

Result:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y \mid A = a]$$

The counterfactual mean equals the observed conditional mean!

Identification Under Randomization: The ATE

We showed: $\mathbb{E}[Y(a)] = \mathbb{E}[Y \mid A = a]$

Therefore:

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

Under randomization, the **naive difference *is* the ATE**.

Why Randomization Works: Selection Bias = 0

Under unconditional exchangeability $A \perp\!\!\!\perp Y(a)$:

Step 1: Independence means the distribution of $Y(0)$ is the same regardless of A:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}[Y(0) \mid A = 0] = \mathbb{E}[Y(0)]$$

Why Randomization Works: Selection Bias = 0

Under unconditional exchangeability $A \perp\!\!\!\perp Y(a)$:

Step 1: Independence means the distribution of $Y(0)$ is the same regardless of A:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}[Y(0) \mid A = 0] = \mathbb{E}[Y(0)]$$

Step 2: Therefore, the selection bias term is:

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

Why Randomization Works: Selection Bias = 0

Under unconditional exchangeability $A \perp\!\!\!\perp Y(a)$:

Step 1: Independence means the distribution of $Y(0)$ is the same regardless of A:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}[Y(0) \mid A = 0] = \mathbb{E}[Y(0)]$$

Step 2: Therefore, the selection bias term is:

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

Step 3: From the canonical decomposition:

$$\text{Naive difference} = \text{ATT} + 0 = \text{ATT}$$

Why Randomization Works: Selection Bias = 0

Under unconditional exchangeability $A \perp\!\!\!\perp Y(a)$:

Step 1: Independence means the distribution of $Y(0)$ is the same regardless of A:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}[Y(0) \mid A = 0] = \mathbb{E}[Y(0)]$$

Step 2: Therefore, the selection bias term is:

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

Step 3: From the canonical decomposition:

$$\text{Naive difference} = \text{ATT} + 0 = \text{ATT}$$

Step 4: Similarly, $\mathbb{E}[Y(1) \mid A = 1] = \mathbb{E}[Y(1)]$, so ATT = ATE.

$\Rightarrow$ **Naive difference = ATT = ATE** under randomization.

Why We Need Covariates

In **observational settings**, unconditional exchangeability rarely holds.

Example: Campaigns target canvassing based on prior voting history.
$\Rightarrow$ Canvassed voters differ systematically from non-canvassed voters.
$\Rightarrow A \not\perp\!\!\!\perp Y(a)$

Hope: Once we account for the targeting criteria (X), the remaining variation in treatment is as-if random.

$\Rightarrow$ We assume **conditional** exchangeability: $A \perp\!\!\!\perp Y(a) \mid X$

G-Formula: Step 1 (Iterated Expectations)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Goal: Identify $\mathbb{E}[Y(a)]$.

Start with what we want:

$$\mathbb{E}[Y(a)]$$

G-Formula: Step 1 (Iterated Expectations)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Goal: Identify $\mathbb{E}[Y(a)]$.

Start with what we want:

$$\mathbb{E}[Y(a)]$$

Apply the **law of iterated expectations**:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X]\Big\}$$

Meaning: Average $Y(a)$ within each stratum of X, then average across strata.

(This is just a statistical identity—no causal assumption yet.)

G-Formula: Step 2 (Exchangeability)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X]\Big\}$$

G-Formula: Step 2 (Exchangeability)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X]\Big\}$$

Apply **conditional exchangeability**: Within each stratum $X = x$, treatment is independent of $Y(a)$.

$$\mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a]$$

G-Formula: Step 2 (Exchangeability)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

From Step 1:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X]\Big\}$$

Apply **conditional exchangeability**: Within each stratum $X = x$, treatment is independent of $Y(a)$.

$$\mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a]$$

So:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X, A = a]\Big\}$$

Meaning: We can focus on units with $A = a$ within each stratum.

G-Formula: Step 3 (Consistency)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X, A = a]\Big\}$$

G-Formula: Step 3 (Consistency)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X, A = a]\Big\}$$

Apply **consistency**: Among units with $A = a$, we have $Y = Y(a)$.

$$\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y \mid X, A = a]$$

G-Formula: Step 3 (Consistency)

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y(a) \mid X, A = a]\Big\}$$

Apply **consistency**: Among units with $A = a$, we have $Y = Y(a)$.

$$\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y \mid X, A = a]$$

$$\mathbb{E}[Y(a)] = \mathbb{E}\Big\{\mathbb{E}[Y \mid X, A = a]\Big\}$$

The G-Formula for ATE

$$\tau = \mathbb{E}\Big\{\mathbb{E}[Y \mid X, A = 1]\Big\} - \mathbb{E}\Big\{\mathbb{E}[Y \mid X, A = 0]\Big\}$$

This is the ATE! It tells us the average effect across the whole population.

In words:

1. Within each stratum $X = x$, compare treated vs. untreated mean outcomes
2. Average those stratum-specific differences over the **population** distribution of X

Example: If X = prior voting, we compare canvassed vs. non-canvassed turnout *separately* among prior voters and non-voters, then combine using population proportions.

Identifying the ATT: The Challenge

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

First term: $\mathbb{E}[Y(1) \mid A = 1]$
This is easy! By consistency, $\mathbb{E}[Y(1) \mid A = 1] = \mathbb{E}[Y \mid A = 1]$.
(Among treated units, we observe $Y(1)$ directly.)

Second term: $\mathbb{E}[Y(0) \mid A = 1]$
This is the problem. We need the average $Y(0)$ among treated units—but treated units have $A = 1$, so we only see their $Y(1)$, never their $Y(0)$.

How can we learn about $Y(0)$ for the treated group?

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 1

Recall: $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big]$ (Law of Iterated Expectations)

We want: $\mathbb{E}[Y(0) \mid A = 1]$

Apply iterated expectations *within the treated group*:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 1] \,\Big|\, A = 1\Big\}$$

Meaning: Average $Y(0)$ by strata of X among treated, then average over the treated group's X distribution.

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 2

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

From Step 1:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 1] \,\Big|\, A = 1\Big\}$$

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 2

Recall: $A \perp\!\!\!\perp Y(a) \mid X$ means $\mathbb{E}[Y(a) \mid X, A = a] = \mathbb{E}[Y(a) \mid X]$ (Exchangeability)

From Step 1:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 1] \,\Big|\, A = 1\Big\}$$

Apply **conditional exchangeability**: Within stratum $X = x$, $Y(0)$ is independent of A.

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

This is the key step! It says: among people with the same X, the treated and untreated have the same average $Y(0)$.

So we can "borrow" information about $Y(0)$ from the untreated group.

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 3

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 0] \,\Big|\, A = 1\Big\}$$

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 3

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 0] \,\Big|\, A = 1\Big\}$$

Apply **consistency**: Among untreated ($A = 0$), we observe $Y = Y(0)$.

$$\mathbb{E}[Y(0) \mid X, A = 0] = \mathbb{E}[Y \mid X, A = 0]$$

Identifying $\mathbb{E}[Y(0) \mid A = 1]$: Step 3

Recall: $A = a \Rightarrow Y = Y(a)$, so $\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y(a) \mid A = a]$ (Consistency)

From Step 2:

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y(0) \mid X, A = 0] \,\Big|\, A = 1\Big\}$$

Apply **consistency**: Among untreated ($A = 0$), we observe $Y = Y(0)$.

$$\mathbb{E}[Y(0) \mid X, A = 0] = \mathbb{E}[Y \mid X, A = 0]$$

$$\mathbb{E}[Y(0) \mid A = 1] = \mathbb{E}\Big\{\mathbb{E}[Y \mid X, A = 0] \,\Big|\, A = 1\Big\}$$

We estimate $Y(0)$ for treated by looking at untreated outcomes *among people with similar* X, then averaging over the treated group's X distribution.

ATT Formula

$$\tau_{\text{ATT}} = \mathbb{E}[Y \mid A = 1] - \mathbb{E}\Big\{\mathbb{E}[Y \mid X, A = 0] \,\Big|\, A = 1\Big\}$$

Key difference from ATE:

- ATE: outer average over **population** distribution of X
- ATT: outer average over **treated group's** distribution of X

Example: Computing ATE and ATT

Setup

$X \in \{0, 1\}$ = voted last election. Observed turnout rates:

X	$\mathbb{E}[Y \mid X, A = 1]$	$\mathbb{E}[Y \mid X, A = 0]$	Diff
0 (didn't vote)	0.35	0.20	0.15
1 (voted)	0.80	0.75	0.05

Population: $P(X = 0) = 0.40$, $P(X = 1) = 0.60$
Treated group: $P(X = 0 \mid A = 1) = 0.20$, $P(X = 1 \mid A = 1) = 0.80$

ATE = $\mathbb{E}\{\mathbb{E}[Y|X, A = 1] - \mathbb{E}[Y|X, A = 0]\}$ over population:

$$\tau = P(X = 0) \cdot 0.15 + P(X = 1) \cdot 0.05 = 0.40 \times 0.15 + 0.60 \times 0.05 = 0.09$$

ATT = same differences, but weighted by treated distribution:

$$\tau_{\text{ATT}} = 0.20 \times 0.15 + 0.80 \times 0.05 = 0.07$$

Why ATE ≠ ATT Here

Campaign targeted high-propensity voters ($X = 1$).

High-propensity voters have **smaller** treatment effects (0.05 vs. 0.15).

- ATT weights by treated distribution ⇒ more weight on small effect
- ATE weights by population distribution ⇒ more weight on large effect

The difference (9% vs. 7%) is not a contradiction—it's a difference in **which population** we average over.

What Fails: Where Derivations Break

Why This Section Matters

Each identification step has a **specific prerequisite**.

When a prerequisite fails:

- A particular equality breaks
- The observed-data formula no longer equals the estimand

We examine five failure modes:

1. Interference
2. Anticipation
3. Post-treatment conditioning
4. Unmeasured confounding
5. Positivity violations

Failure 1: Interference

Problem: Voter i's outcome depends on others' treatment.

Example: Canvassing Maria reminds her roommate Ana to vote.

Why the selector identity fails:
We wrote: $Y_i = A_i \cdot Y_i(1) + (1 - A_i) \cdot Y_i(0)$
But if Ana's outcome depends on Maria's treatment:

$$Y_{\text{Ana}} \neq A_{\text{Ana}} \cdot Y_{\text{Ana}}(1) + (1 - A_{\text{Ana}}) \cdot Y_{\text{Ana}}(0)$$

Because $Y_{\text{Ana}}(0)$ isn't a single number—it depends on whether Maria was canvassed!

Correct notation: $Y_{\text{Ana}}(A_{\text{Ana}}, A_{\text{Maria}})$

Warning

Interference breaks SUTVA. Standard formulas require redesigning what "treatment" means.

Failure 2: Anticipation

Problem: Units change behavior because they *expect* treatment.

Example: Government announces a new tax policy for next year. Firms change investment *this year* in anticipation.

The notation problem:
We defined $Y_i(0)$ as outcome under "no treatment." But what does $A = 0$ mean?

- $A = 0$ and policy never announced?
- $A = 0$ and policy announced but not yet implemented?

Correct notation: $Y_i(\text{treatment path})$, e.g., $Y_i(\text{announced, not implemented})$

Consistency fails: We substitute Y for $Y(0)$ among $A = 0$ units. But their observed Y reflects anticipation—it's not the "clean" no-treatment outcome.

Warning

Anticipation contaminates the control condition. What you call $Y(0)$ isn't what you meant.

Failure 3: Post-Treatment Conditioning

Problem: Conditioning on a variable M that is *affected by* treatment.

Example: M = "enthusiasm" measured after canvassing. Comparing turnout among voters with same enthusiasm.

The notation problem:
M has its own potential outcomes: $M(1)$ and $M(0)$.
When we condition on M, we're mixing:

- Treated units with $M(1) = m$
- Untreated units with $M(0) = m$

These are **different types of people**! A treated person with high enthusiasm ($M(1) = \text{high}$) may be very different from an untreated person with high enthusiasm ($M(0) = \text{high}$).

Warning

Even if $A \perp\!\!\!\perp Y(a) \mid X$, generally $A \not\perp\!\!\!\perp Y(a) \mid X, M$. Bad controls induce bias.

Failure 4: Unmeasured Confounding

Problem: Variables affect both A and $Y(a)$, but aren't in X.

Example: U = "persuadability score" (unobserved). High-U voters are targeted for canvassing *and* respond more to canvassing.

The notation problem:
We assumed: $A \perp\!\!\!\perp Y(a) \mid X$
But the truth is: $A \perp\!\!\!\perp Y(a) \mid X, U$ (we'd need to condition on U too)
Since we only condition on X:

$$\mathbb{E}[Y(a) \mid X, A = 1] \neq \mathbb{E}[Y(a) \mid X, A = 0]$$

Within strata of X, treated units have higher U on average $\Rightarrow$ higher $Y(a)$.

Warning

Unmeasured confounders invalidate exchangeability. Selection bias remains hidden in your estimate.

Failure 5: Positivity Violations

Problem: $P(A = a \mid X = x) = 0$ for some relevant x.

Example: Campaign never canvasses voters under 25.
$\Rightarrow \mathbb{E}[Y \mid X = \text{“under 25”}, A = 1]$ is undefined (no units to average).

What breaks: G-formula requires this quantity.

Positivity violations can be **structural** (targeting rules), not just small-sample.

Warning

Positivity violations make conditional expectations undefined. Identification fails for affected strata.

The Habit We Want

Whenever you see an observed-data formula:

1. Point to the **exact equalities** that produced it
2. Identify the **exact assumptions** that justify each equality
3. Ask whether those assumptions are **plausible** in your setting

This prevents turning a clean identification argument into an **assumption-blind plug-in procedure**.

Questions?