

Selection Bias and Randomized Controlled Trial

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

March 26, 2026

Causal Effect and Counterfactual

- In many applications, we might be interested in the ATE or the ATT
- These causal effects are defined by **comparing potential outcomes**
- But we only know:
 - ▶ Whether an individual was treated or not
 - ▶ The **observed outcome** of the individual

Causal Effect and Counterfactual

- Causal effects are defined by **potential outcomes**
- For example, the ATT:

$$\text{ATT} = E[Y_i^1 - Y_i^0 \mid D_i = 1] = \underbrace{E[Y_i^1 \mid D_i = 1]}_{\text{observable}} - \underbrace{E[Y_i^0 \mid D_i = 1]}_{\text{counterfactual}}$$

- The first term is observed:
 - ▶ It is the average outcome of the treated group
- The second term, $E[Y_i^0 \mid D_i = 1]$, is the **counterfactual**:
 - ▶ What would the treated group's outcome have been *had they not been treated*?
- This counterfactual is **never observed** — the fundamental problem of causal inference

Constructing the Counterfactual from Data

- Since $E[Y_i^0 | D_i = 1]$ is unobservable, we need to **approximate it** using observable data
- The most natural candidate: use the **untreated group's outcome** as the counterfactual

$$E[Y_i^0 | D_i = 1] \approx E[Y_i^0 | D_i = 0] = E[Y_i | D_i = 0]$$

- This leads to the **Observed Difference in Outcomes (ODO)**:

$$\text{ODO} = E[Y_i | D_i = 1] - E[Y_i | D_i = 0] = \underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Treated}} - \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Untreated}}$$

Selection Bias

Selection Bias

- **Question:** Does ODO recover the true ATT? *Not in general* — it may be contaminated by **selection bias**
- Decompose the ODO by adding and subtracting $E[Y_i^0|D_i = 1]$:

$$\begin{aligned} & \underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{ODO}} \\ &= E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1] + E[Y_i^0|D_i = 1] - E[Y_i^0|D_i = 0] \\ &= \underbrace{E[Y_i^1 - Y_i^0|D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_i^0|D_i = 1] - E[Y_i^0|D_i = 0]}_{\text{Selection Bias}} \end{aligned}$$

Selection Bias

- **Selection Bias** = $E[Y_i^0 | D_i = 1] - E[Y_i^0 | D_i = 0]$:
 - ▶ The difference in *untreated* potential outcomes between the two groups
- If selection bias $\neq 0$, the untreated group is a **bad counterfactual** for the treated group
- **Our goal:** Find a way to make selection bias = 0

Selection Bias: A Numerical Example

- To compute the ATT, we need $E[Y_i^0 | D_i = 1]$ — the counterfactual for the **treated group**

i	D_i	Y_i^1	Y_i^0	Y_i	$Y_i^1 - Y_i^0$
David	1	3	?	3	?
Tina	1	2	?	2	?
Mary	0	?	1	1	?
Bill	0	?	1	1	?
$E[Y_i^1 D_i = 1]$		2.5			
$E[Y_i^0 D_i = 1]$?		

- $E[Y_i^1 | D_i = 1] = 2.5$ is observed — but $E[Y_i^0 | D_i = 1]$ is the unobservable counterfactual
- Without it, we **cannot compute the ATT directly**

Selection Bias: A Numerical Example

- Use the untreated group as the counterfactual and compute ODO

i	D_i	Y_i^1	Y_i^0	Y_i	$Y_i^1 - Y_i^0$
David	1	3	?	3	?
Tina	1	2	?	2	?
Mary	0	?	1	1	?
Bill	0	?	1	1	?
$E[Y_i D_i = 1]$				2.5	
$E[Y_i D_i = 0]$				1.0	

- We can observe both group means and compute:

$$\text{ODO} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = 2.5 - 1.0 = 1.5$$

- But is $\text{ODO} = 1.5$ the true causal effect (ATT)? Only if there is no selection bias

Selection Bias: A Numerical Example

- Suppose we *could* observe all counterfactuals:

i	D_i	Y_i^1	Y_i^0	Y_i	$Y_i^1 - Y_i^0$
David	1	3	2	3	1
Tina	1	2	1	2	1
Mary	0	1	1	1	0
Bill	0	1	1	1	0
$E[Y_i^1 D_i = 1]$		2.5			
$E[Y_i^0 D_i = 1]$		1.5			
$E[Y_i^0 D_i = 0]$		1.0			

- $ATT = E[Y_i^1 | D_i = 1] - E[Y_i^0 | D_i = 1] = 2.5 - 1.5 = 1$
- Selection Bias = $E[Y_i^0 | D_i = 1] - E[Y_i^0 | D_i = 0] = 1.5 - 1.0 = 0.5$

$$\Rightarrow \underbrace{ODO}_{1.5} = \underbrace{ATT}_{1} + \underbrace{\text{Selection Bias}}_{0.5}$$

- Key:** David & Tina earn more *even without* treatment

Sources of Selection Bias

- Selection bias arises from **self-selection into treatment**
 - ▶ Individuals choose treatment based on observable *and* unobservable characteristics
 - ▶ E.g., more able students are more likely to attend graduate school, and also earn more regardless
- This non-random selection leads to systematic differences between treated and untreated groups
- **The fundamental challenge of causal inference:** Eliminate selection bias to recover the true causal effect

Causal Effect and Identification Strategy

- Identification strategy tells us what we can learn about a **causal effect** from **observed data**
 - ▶ The main goal is to eliminate **selection bias**
- Identification depends on **assumptions**, not on estimation strategies
 - ▶ Estimation strategies: OLS, MLE, GMM
 - ▶ If an effect is not identified, no estimation method will recover it
- Main strategies to eliminate selection bias:
 - ▶ **RCT**: Do not allow self-selection into treatment
 - ▶ **Matching/Regression**: Control for all observable confounders
 - ▶ **DID, IV, RDD**: Exploit exogenous variation in treatment

Randomized Controlled Trial

Randomized Controlled Trial

- The most credible identification strategy is the **Randomized Controlled Trial (RCT)**
- RCT: Each observation is **randomly assigned** to treatment or control group
- RCT has two key features that eliminate selection bias:
 - 1 Randomly assign treatment
 - 2 Sufficiently “large” sample size

Why Does Random Assignment Work?

- Random assignment (e.g., coin flip) ensures:
 - ▶ The probability of receiving treatment is **unrelated to any confounding factor**
 - ▶ Every observation has the **same probability** of being assigned to the treatment group

- This means potential outcomes are **independent** of treatment assignment:

$$(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i$$

- **Intuition:** D is determined by a coin flip, not by the individual's potential outcomes

Why Does Random Assignment Work?

- Random assignment creates treatment and control groups that are **similar on average** across all characteristics
- This similarity extends to potential outcomes:
 - ▶ $E[Y_i^0 | D_i = 1] = E[Y_i^0 | D_i = 0]$
 - ▶ $E[Y_i^1 | D_i = 0] = E[Y_i^1 | D_i = 1]$
- Therefore, **selection bias is eliminated**:

$$\underbrace{E[Y_i^0 | D_i = 1] - E[Y_i^0 | D_i = 0]}_{\text{Selection Bias}} = 0$$

RCT Identifies ATT, ATU, and ATE

$$\begin{aligned} & \underbrace{E[Y_i | D_i = 1] - E[Y_i | D_i = 0]}_{\text{ODO}} \\ &= \underbrace{E[Y_i^1 - Y_i^0 | D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_i^0 | D_i = 1] - E[Y_i^0 | D_i = 0]}_{\text{Selection Bias}} \\ &= \underbrace{E[Y_i^1 - Y_i^0 | D_i = 1]}_{\text{ATT}} + \underbrace{0}_{\text{Selection Bias}} \\ &= \underbrace{E[Y_i^1 - Y_i^0 | D_i = 0]}_{\text{ATU}} \\ &= \underbrace{E[Y_i^1 - Y_i^0]}_{\text{ATE}} \end{aligned}$$

- In RCT, a simple **observed difference in outcome (ODO)** provides an unbiased estimate of the causal effect

The Role of Large Sample Size

- Random assignment ensures similarity **on average**, but:
 - ▶ If each group has only one individual, random assignment alone cannot guarantee balance
- We also need **large sample size** to ensure that group differences in individual characteristics wash out
- With large samples, the Law of Large Numbers guarantees that treatment and control groups are balanced across all observable *and* unobservable characteristics

- There are two main types of RCT:

1 Lab experiment

- ★ Conducted under highly controlled laboratory conditions
- ★ Individuals play games or tasks as a proxy for real-world treatment

2 Field experiment

- ★ Conducted in a natural setting (school, workplace, market)
- ★ Allows researchers to observe treatment effects in the real world

Empirical Example

Empirical Example

Noy & Zhang (2023)

Shakked Noy & Whitney Zhang (2023) “**Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence**” *Science*

- The authors examine the effect of ChatGPT access on workers' productivity in professional writing tasks
- We use this example to illustrate how an RCT eliminates selection bias and identifies the causal effect

AI and the Labor Market

Theoretical Framework: Task-Based Models

- The dominant theoretical framework in labor economics for analyzing automation is the **task-based model** (Acemoglu & Restrepo, 2018, 2022)
 - ▶ Production requires a continuum of **tasks**; each task can be performed by either labor or capital (including AI)
 - ▶ **Automation** = AI takes over tasks previously performed by workers
- Automation has **two opposing effects** on wages:
 - ▶ **Productivity effect (+)**: AI lowers costs \Rightarrow output expands \Rightarrow labor demand rises
 - ▶ **Displacement effect (-)**: Workers are pushed out of automated tasks \Rightarrow labor demand falls
- Net effect on wages and employment is **ambiguous** — depends on which effect dominates
 - ▶ New task creation (e.g., AI engineer, prompt designer) can offset displacement
 - ▶ The presumption that all technologies increase labor demand simply be

AI and the Labor Market

What Does Theory Predict for Generative AI?

- Acemoglu (2024) applies the task-based framework to AI
- AI-based productivity gains can come from three channels:
 - ▶ **Automation:** AI takes over tasks previously performed by workers
 - ▶ **Complementarity:** AI augments worker productivity in remaining tasks
 - ▶ **New task creation:** AI enables entirely new products and job roles
- Acemoglu's conclusion is **cautiously pessimistic**:
 - ▶ Even if AI raises productivity significantly within a task, only a small share of all tasks in the economy are currently exposed to AI — so aggregate effects are limited
 - ▶ Projected aggregate TFP gains over 10 years: less than 0.55%, well below optimistic forecasts (e.g., Goldman Sachs: +15% over 10 years)

AI and the Labor Market

From Theory to Evidence

- Theory gives us the framework — but the net effect of AI on productivity and wages is ultimately an **empirical question**
- We need **causal evidence** to quantify:
 - ▶ How large are the productivity gains at the task level?
 - ▶ Who benefits — high-ability or low-ability workers?
 - ▶ Does AI complement or substitute for labor?

⇒ **Noy & Zhang (2023)** provide exactly this: a randomized controlled trial measuring the causal effect of ChatGPT access on worker productivity in professional writing tasks

Noy & Zhang (2023)

Experimental Design

- **Setting:** Preregistered online experiment with 453 college-educated professionals
 - ▶ Occupations: marketers, grant writers, consultants, data analysts, HR professionals, managers
- **Tasks:** Each participant completed **two** occupation-specific writing assignments
 - ▶ Press releases, short reports, analysis plans, delicate emails
 - ▶ 20–30 minute tasks designed to resemble real work
 - ▶ **Task 1** (pre-treatment): completed by both groups without ChatGPT
- **Incentives:** Base pay \$10 + up to \$14 bonus for output quality
 - ▶ Evaluators were blinded, experienced professionals grading on a 1–7 scale

- **Random Assignment:** 50% assigned to treatment, 50% to control
 - ▶ Treatment ($D_i = 1$), $n_1 = 218$: instructed to register for ChatGPT and permitted to use it on Task 2
 - ▶ Control ($D_i = 0$), $n_0 = 235$: instructed to register for Overleaf (same hassle cost) — no AI access; < 5% used it
- **Outcomes (Y_i):** measured on Task 2 only
 - ▶ Time spent on the task (minutes)
 - ▶ Average evaluator grade (1–7 scale)

Step 1: Balance Test

How to Test Balance

- The first thing we check in any RCT is whether the two groups are **similar *before* treatment**
- If the two groups had different *pre-existing* characteristics, the ODO would be contaminated by selection bias
- **Balance check:** We test whether pre-treatment variables differ significantly across groups
 - ▶ If no pre-treatment variable is significantly different \Rightarrow randomization succeeded

Step 1: Balance Test

Null Hypothesis and t -statistic

- **Null hypothesis H_0** : no difference in pre-treatment variable X between the two groups

$$H_0 : E[X_i | D_i = 1] - E[X_i | D_i = 0] = 0$$

- We test H_0 using a **two-sample t -test**:

$$t = \frac{\bar{X}_1 - \bar{X}_0}{SE}, \quad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$$

where $\bar{X}_d = \frac{1}{n_d} \sum_{i:D_i=d} X_i$, and s_d^2 is the sample variance in group d

- The t -statistic measures whether $\bar{X}_1 - \bar{X}_0$ is away from zero (H_0)
 - ▶ Large $|t| \Rightarrow$ difference is large relative to sampling variability \Rightarrow evidence against H_0

Step 1: Balance Test

Binary Variables

Variable	Treatment	Control	Diff.	t-stat	p-value
Employed	0.960 (0.196)	0.910 (0.286)	0.050 (0.020)	2.50	0.013**
Occ: HR Professional	0.110 (0.314)	0.060 (0.238)	0.046 (0.026)	1.77	0.077*
Occ: Consultant	0.110 (0.314)	0.130 (0.337)	-0.013 (0.030)	-0.45	0.655
Occ: Data Analyst	0.110 (0.314)	0.110 (0.314)	0.000 (0.029)	0.00	1.000
Occ: Grant Writer	0.170 (0.376)	0.160 (0.367)	0.012 (0.035)	0.34	0.724
Occ: Manager	0.410 (0.493)	0.430 (0.496)	-0.017 (0.046)	-0.37	0.719
Occ: Marketer	0.090 (0.287)	0.110 (0.314)	-0.023 (0.028)	-0.82	0.412
Observations	218	235			

Step 1: Balance Test

Continuous Variables

Variable	Treatment	Control	Diff.	<i>p</i> -value
Annual Salary (\$)	71,938	67,764	4,174	0.256
Years in Occupation	10.07	10.49	-0.42	0.630
Task 1 Time (min)	26.58	26.10	0.48	0.681
Task 1 Grade (1-7)	3.77	3.63	0.14	0.244
Job Satisfaction (1-10)	6.34	6.30	0.04	0.834
Self-Efficacy (1-10)	6.90	6.89	0.01	0.954
Observations	218	235		

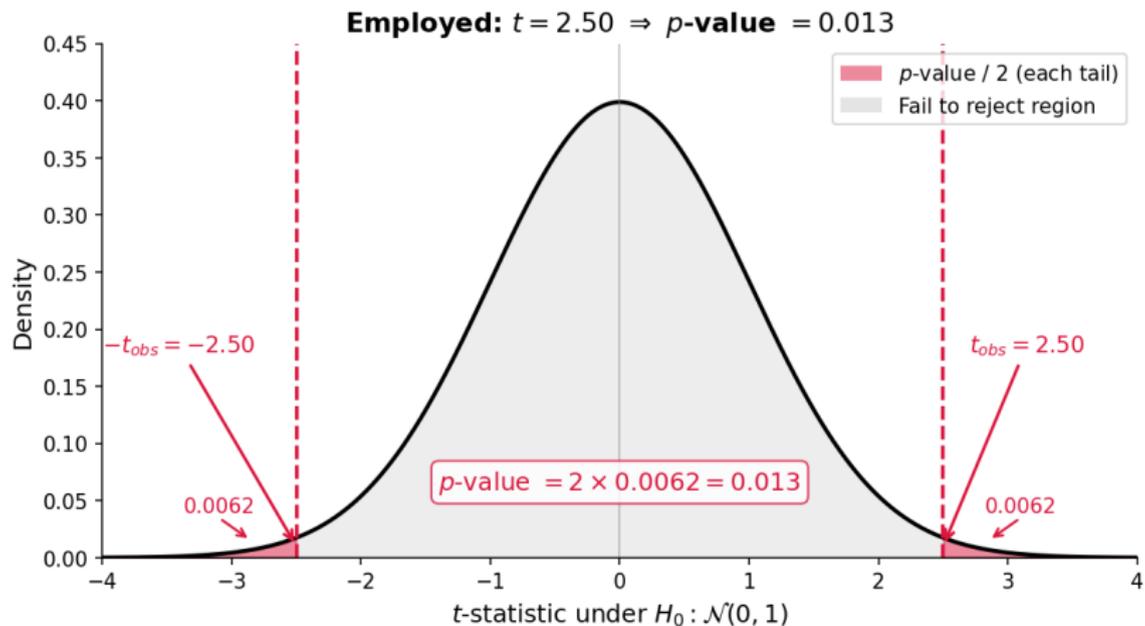
Step 1: Balance Test

From t -statistic to p -value

- To evaluate whether t -statistic is large enough to against H_0 , we need to calculate the **p -value**
 - ▶ **p -value**: the probability of obtaining such t -statistic (or a more extreme value) if the null hypothesis were true
 - ▶ **Central Limit Theorem (CLT)**: When sample size is sufficiently large, t -statistic will have standard normal distribution
 - ▶ Based on standard normal distribution and t -statistic, we can get p -value

Step 1: Balance Test

Visualizing the p -value



Step 1: Balance Test

From t -statistic to p -value

- We usually select an arbitrarily pre-defined threshold value γ , which is referred to as the **level of significance**
 - ▶ By convention, γ is commonly set to 0.05 or 0.01
- **Decision rule:**
 - ▶ If $p\text{-value} < \gamma$: **Reject** the null hypothesis
 - ▶ If $p\text{-value} \geq \gamma$: **Fail to reject** the null hypothesis

Step 1: Balance Test

Reading the Results

- **Employed :**

- ▶ $t = 2.50$, $p\text{-value} = 0.013 < 0.05 \Rightarrow$ **reject** H_0
- ▶ The two groups are **significantly different** in employment status

- **Occ: Consultant :**

- ▶ $t = -0.45$, $p\text{-value} = 0.655 \geq 0.05 \Rightarrow$ **fail to reject** H_0
- ▶ **No significant difference** between the two groups in consultant share

- In a balance test, we *want* most variables to have large p -values

- ▶ It means the two groups look similar before treatment

Step 2: Main Results

ODO as Causal Effect

Outcome (Task 2)	Treatment	Control	Diff. (ODO)	p -value
Time Spent (min)	17.0 (17.3)	27.0 (17.3)	-10.0 (1.5)	< 0.001***
Average Grade (1-7)	4.53 (1.37)	3.80 (1.37)	+0.73 (0.13)	< 0.001***
Observations	218	235		

Step 2: Main Results

Summary

- ChatGPT significantly reduced task time by 10 minutes (-37%)
 - ▶ $p < 0.001$: **reject** H_0 — the two groups are significantly different
- ChatGPT significantly raised output quality by 0.73 points ($+19\%$)
 - ▶ $p < 0.001$: **reject** H_0 — the two groups are significantly different
- Because the RCT eliminates selection bias, $ODO = ATE$
 - ▶ ChatGPT *causes* workers to finish faster *and* produce higher-quality output

Additional Results

- **Heterogeneous effects:** ChatGPT helped lower-ability workers the most
 - ▶ Workers who performed poorly on Task 1 saw the largest improvements on Task 2
 - ▶ Workers who already performed well maintained their quality while also saving time
 - ▶ ChatGPT **compressed the gap** between high- and low-ability workers

- **What makes this credible?**

- ▶ Random assignment ensures treatment and control groups are comparable *before* treatment
- ▶ Selection bias = 0 \Rightarrow the simple difference in means recovers the true causal effect
- ▶ Without the RCT, we could not rule out that more productive workers simply chose to use ChatGPT

RCT: The Gold Standard of Causal Inference

- RCT is widely regarded as the **gold standard** for causal inference because:
 - ▶ It **directly eliminates** selection bias by design, without relying on modeling assumptions
 - ▶ Identification rests on **one transparent assumption**: random assignment
 - ▶ The simple difference in means is an **unbiased estimator** of ATE, ATT, and ATU
 - ▶ Results are **easy to interpret** and communicate to policymakers
- Other methods (DID, IV, RDD, matching) are often judged by how closely they approximate the ideal of an RCT

Limitations of RCT: Practical Challenges

- Despite its appeal, RCTs are often **difficult to implement** in practice:
 - ▶ **Cost:** Running large-scale field experiments requires substantial time and resources
 - ▶ **Compliance:** Individuals may not comply with their assigned treatment (non-compliance problem)
 - ▶ **Attrition:** Participants may drop out differentially, reintroducing selection bias
 - ▶ **Generalizability:** Results from a specific RCT may not generalize to other settings (external validity)
 - ▶ **Ethics:** Randomly assigning individuals to the control group may be unfair if the treatment is potentially beneficial

Suggested Readings

- Chapter 1 and 2, *Mastering Metrics: The Path from Cause to Effect*
- Chapter 2, *Mostly Harmless Econometrics*
- Chapter 4, *Causal Inference: The Mixtape*

References

- Acemoglu, D. (2024). “The Simple Macroeconomics of AI.” *Economic Journal*, 134(641), 3155–3185.
- Acemoglu, D., & Restrepo, P. (2018). “The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment.” *American Economic Review*, 108(6), 1488–1542.
- Acemoglu, D., & Restrepo, P. (2022). “Tasks, Automation, and the Rise in U.S. Wage Inequality.” *Econometrica*, 90(5), 1973–2016.