Causal Inference using Difference-in-Differences

Lecture 1: Introduction

Pedro H. C. Sant'Anna
Emory University

January 2025

Introduction and DiD popularity

Importance of Empirical Research

- The availability of richer datasets and the advances in computational power have changed Social Sciences during the last 40 years.
- Currie, Kleven and Zwiers (2020) show that the fraction of empirical research keeps rising.
- A very common goal of empirical research is to uncover/highlight the **casual effect** of a given policy intervention.

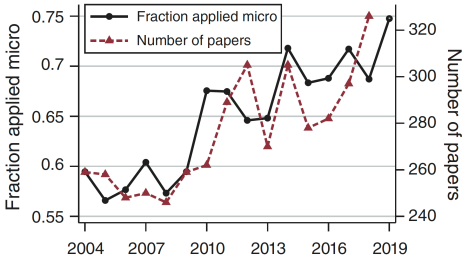


FIGURE 1. APPLIED MICROECONOMICS ARTICLES IN TOP-FIVE JOURNALS

Note: This figure shows the fraction of papers in top-five journals that report an applied microeconomics JEL code (left axis) and the total number of papers in the top-five journals (right axis).

The boom of experimental and quasi-experimental methods

Currie et al. (2020) documented this change well

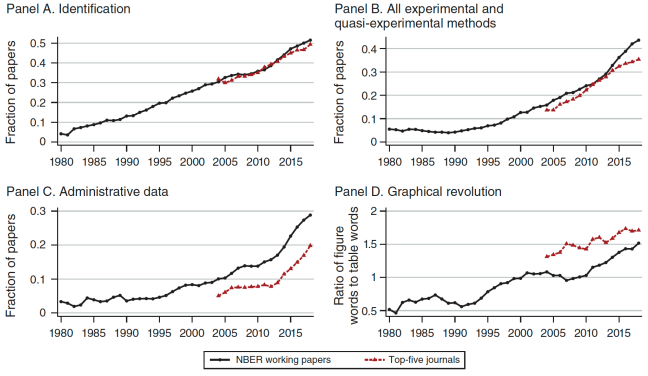


FIGURE 2. THE CREDIBILITY REVOLUTION

Notes: This figure shows different dimensions of the "credibility revolution" in economics: identification (panel A), all experimental and quasi-experimental methods (panel B), administrative data (panel C), and the graphical revolution (panel D). Panel D shows the ratio of the number of "figure" terms to the number of "table" terms mentioned. See Table A.I for a list of terms. The series show five-year moving averages.

What about experiments (or A/B tests)?

Currie et al. (2020)

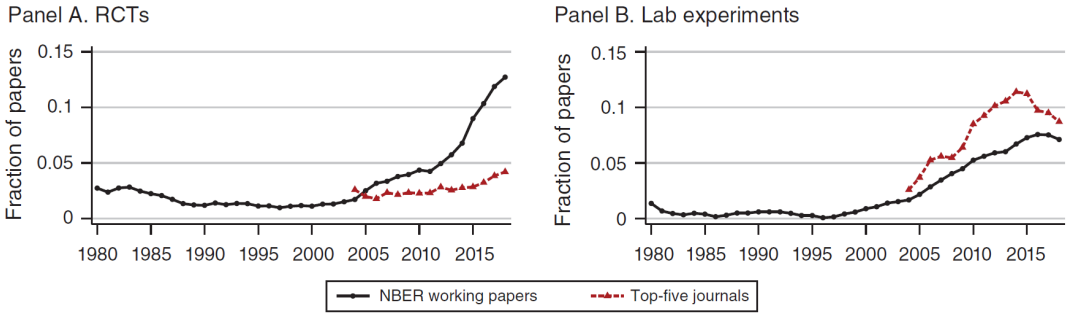


FIGURE 3. EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of experiment. See Table A.I for a list of terms. The series show five-year moving averages.

Popularity of Difference-in-Differences methods

Currie et al. (2020)

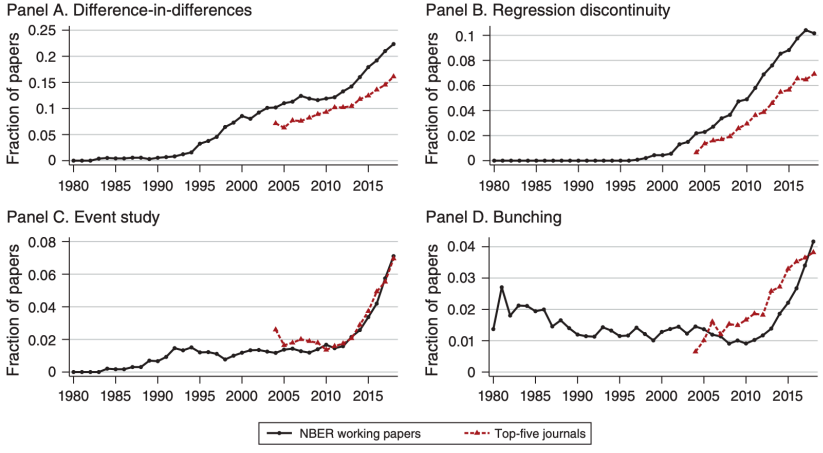
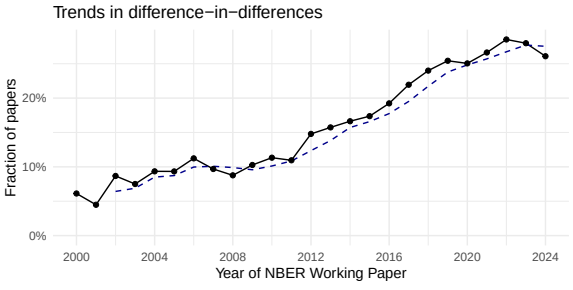
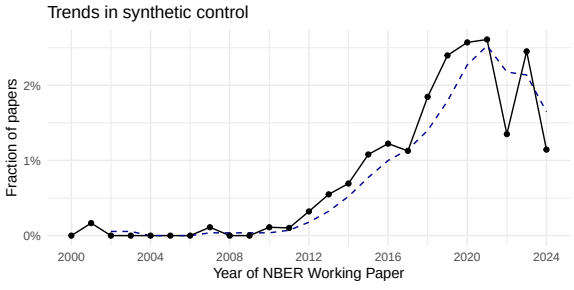


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Recent popularity of DiD methods in empirical work

Goldsmith-Pinkham (2024) built on Currie et al. (2020) and updated the analysis using NBER working papers data that ends in May 2024.

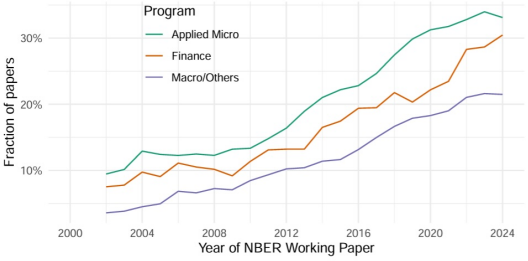




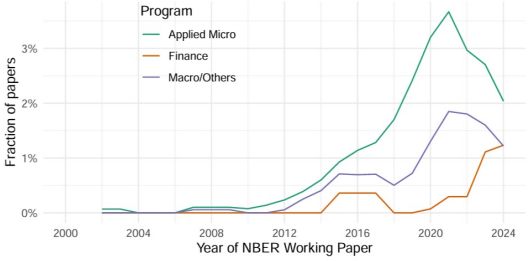
The Y-axes are in different scales.

Popularity of Difference-in-Differences methods: by fields

Goldsmith-Pinkham (2024): the popularity of DiD by fields



(a) Difference-in-differences

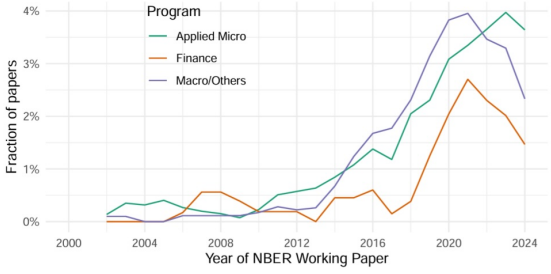


(b) Synthetic controls

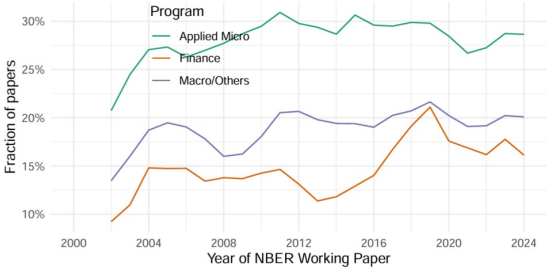
Figure 5: Panel (a) reports the share of papers that mention difference-in-differences or event studies. Figure (b) reports the share of papers that mention synthetic controls (this includes both synthetic difference-in-differences and synthetic control methods). See Table 2 for the breakdown of fields, and the Appendix for definitions on keywords.

Popularity of Difference-in-Differences methods

Goldsmith-Pinkham (2024): Compare previous plot with IV



(a) Bartik and shift-share instruments



(b) Instrumental variables

Figure 6: Panel (a) reports the share of papers that mention Bartik or shift-share instruments. Figure (b) reports the share of papers that mention instrumental variables. See Table 2 for the breakdown of fields, and the Appendix for definitions on keywords.

Why DiD is so popular?

Causality with Observational Data: What can we do?

- In many applications, we do not have access to experimental data.
- Without an experiment, we will rely on **observational data**.
- With **observational data**, we have no choice but rely on **assumptions** to conduct causal inference.
- Different methods rely on different assumptions.
- Our job as researchers is to assess the pros and cons of each method in their ability to answer the questions we (and the business/policy makers/stakeholders) care about.

Causality with Observational Data: What can we do?

- DiD is very popular.
- **WHY?!**
- My guess: data requirements, availability of tools to assess the plausibility of assumptions and easy-to-use software.
- What are the main alternatives to DiD?

  1. Rely on unconfoundedness and leverage **regression, matching, re-weighting** or **double machine learning**.

     Drawback: Rule out selection on unobservables.
     We need to have data on everything that affects treatment timing and outcome of interest (unconfoundedness assumption).

Causality with Observational Data: What else?

- What are the other main alternatives to DiD?

  2. Rely on **Pre-Post analysis**

     Drawback: Does not account for potential trends in outcomes.
     This is more reasonable if we study very short-run effects, but that is not usually the case.

The appeal of Difference-in-Differences

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.
- **DiD combines previous approaches to avoid their pitfalls**.
- **Advantage: Allow for selection on unobservables and time-trends.**
  Not magic: We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would evolve similarly across groups/cohorts - **Parallel Trends assumption**.

  Parallel Trends needs to be discussed and its plausibility assessed!

- **Data Requirements:** We need data from periods before and after treatment to use DiD (and some periods where no unit is treated).

Some DiD Examples

Some DiD Examples

- Card and Krueger (1994): Effect of minimum wage on employment.
  - Compared the changes in wages, employment, and prices at stores in New Jersey (increased minimum wage) relative to stores in Pennsylvania (minimum wage remained fixed).

- Dube, Lester and Reich (2010); Dube, William Lester and Reich (2016), Callaway and Sant'Anna (2021) and many others:
  Effect of minimum wage on different measures of employment
  - Callaway and Sant'Anna (2021) exploit variation in the timing of state minimum wage changes to understand its effect on teen employment.

Some DiD examples

- Meyer, Viscusi and Durbin (1995): Effect of weekly benefit amount on time out of work due to injury.
  - They compared high-earnings (affected by the policy change) and low-earnings (not affected by the policy change) individuals injured before and after increases in the maximum weekly benefit amount. Estimated effects in Kentucky and Michigan.
- Malesky, Nguyen and Tran (2014): Effect of government re-centralization in Vietnam on public services.
  - They compared provinces (and districts) that abolished elected councils in Vietnam to other provinces that did not abolish them, before and after the re-centralization. Analyzed 30 outcomes.

Some DiD examples

- Carey, Miller and Wherry (2020): Effect of Medicaid expansion on access to care and utilization for those who are already insured.
  - They compare different insurance coverage and health care utilization measures among states that opted to expand Medicaid eligibility in 2014 or 2015 with those that did not expand by 2015, before and after the expansion.

- Assunção, Gandour, Rocha and Rocha (2020): Effect of rural credit on deforestation.
  - Compared municipalities within the Amazon biome (concession of subsidized rural credit for them are conditional on stricter requirements since 2008), with municipalities outside the border of the Amazon biome (not affected by the policy change), before and after the policy.

Some DiD examples

- Beck, Levine and Levkov (2010): Effect of bank branching deregulation on income distribution in the US.
  - Exploit staggered bank deregulation across states to understand its effect on the Gini index (among other outcomes); see also Baker, Larcker and Wang (2022).
- Venkataramani, Shah, O'Brien, Kawachi and Tsai (2017): Effect of US Deferred Action for Childhood Arrivals (DACA) immigration program on health outcomes.
  - Compared changes in health outcomes among individuals who met key DACA eligibility criteria (based on age at immigration and at the time of policy implementation) before and after program implementation versus changes in outcomes for individuals who did not meet these criteria.

Canonical DiD Estimator

The canonical Difference-in-Differences estimator

- **The canonical DiD estimator** is given by

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=treated,t=post} - \overline{Y}_{g=treated,t=pre}\right) - \left(\overline{Y}_{g=untreated,t=post} - \overline{Y}_{g=untreated,t=pre}\right),$$

where $\overline{Y}_{g=d,t=j}$ is the sample mean of the outcome Y for units in group d in time period j,

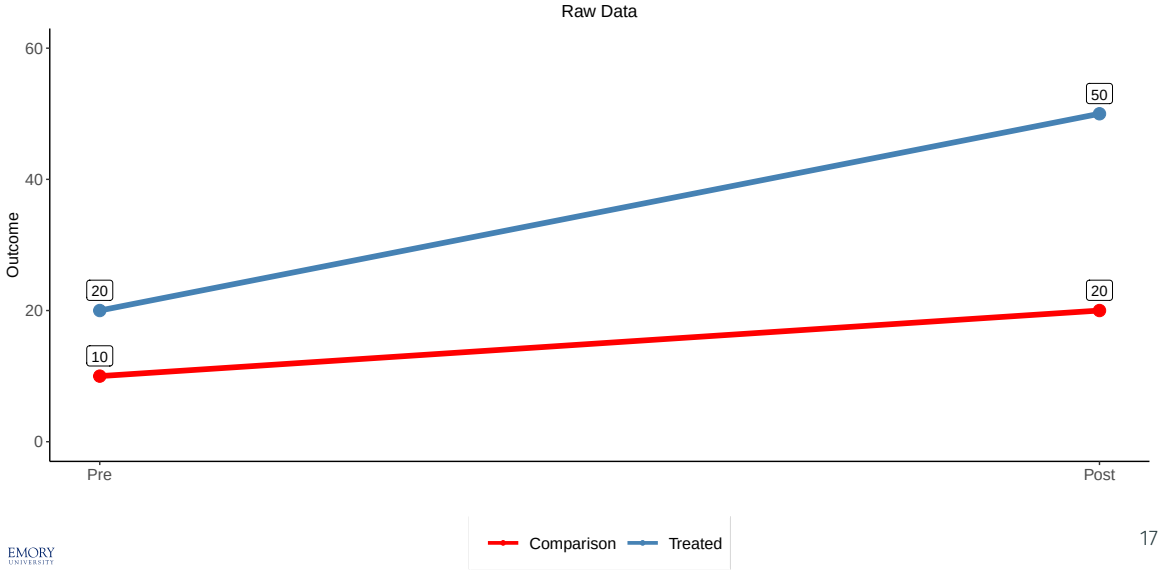
$$\overline{Y}_{g=d,t=j} = \frac{1}{N_{g=d,t=j}} \sum_{i=1}^{N_{all}} Y_i 1\{G_i = d\} 1\{T_i = j\},$$

with

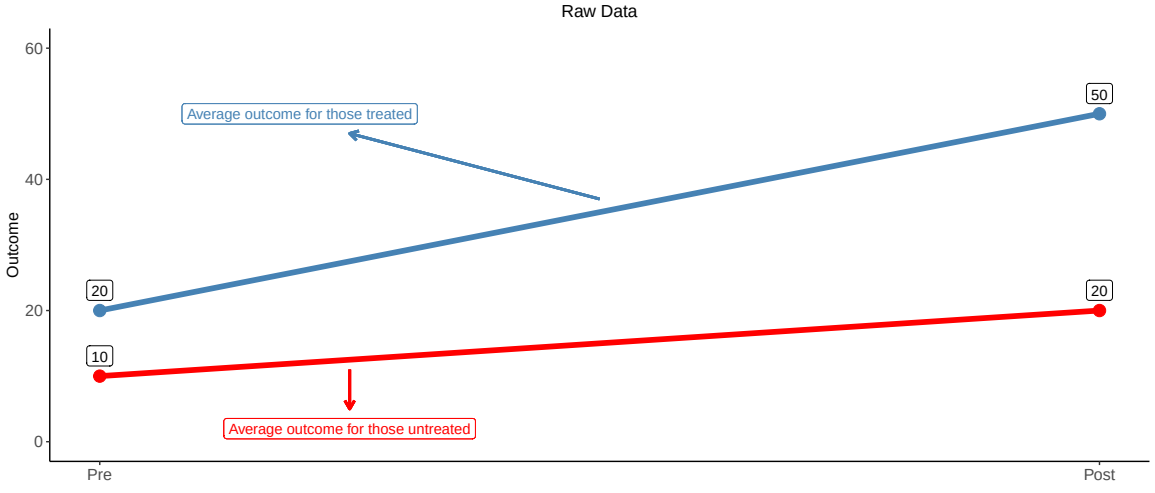
$$N_{g=d,t=j} = \sum_{i=1}^{N_{all}} 1\{G_i = d\} 1\{T_i = j\},$$

G_i and T_i are group and time dummy, respectively, and Y_i is the "pooled" outcome data.

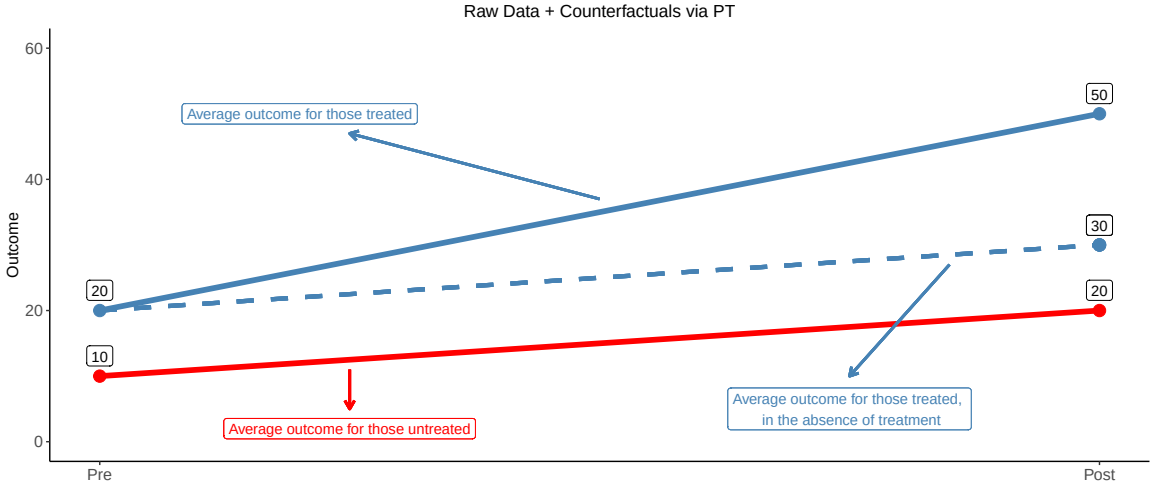
Difference-in-Differences via graphs



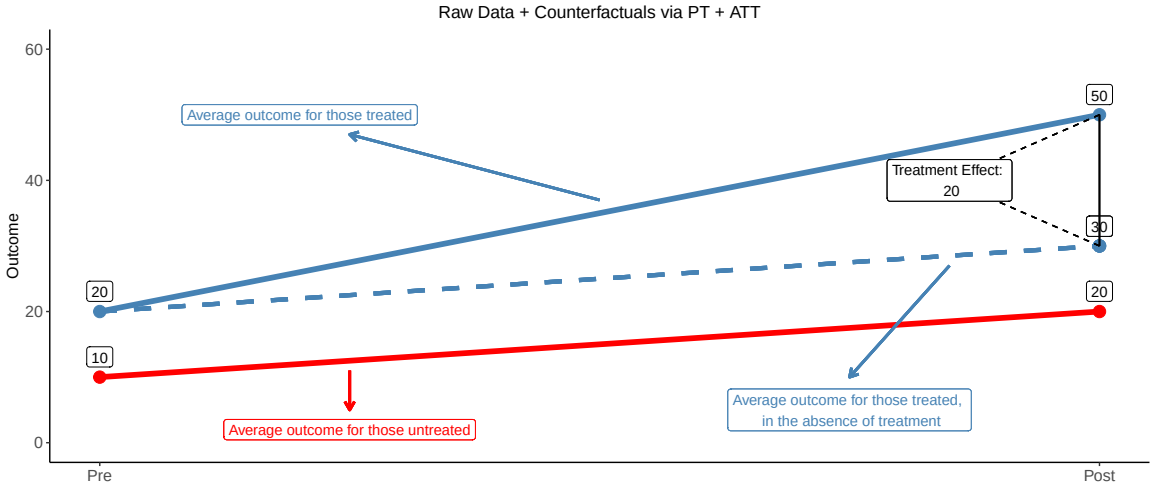
Difference-in-Differences via graphs



Difference-in-Differences via graphs



Difference-in-Differences via graphs



But what kind of treatment effect

parameter $\widehat{\theta}_n^{DiD}$ is actually recovering?

We need to talk about:

1. Potential outcomes

2. Assumptions

Potential Outcomes

Causality with potential outcomes

- We will adopt the **Rubin Causal Model** and define potential outcomes.
- Potential outcomes will reflect the time you are first treated (we can "play" with this later).
- Let $Y_{i,t}(g)$ be the potential outcome for unit i, at time t, if this unit is first treated at time period g.
- T periods: $t = 1, ..., T$.
- Let $G_i \in \mathcal{G} \subset \{1, ..., T\} \cup \{\infty\}$ denote the time unit i is first-treated, with the notion that if a unit is "never-treated", $G_i = \infty$.
- Observed outcome data in time period t for unit i is given by $Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$.

Causality with potential outcomes - The “never treated” group

- We call a group “never treated” if this set of units remains untreated in all periods in our data.
- With two time periods $t = 1, 2$, we call the group of units still not exposed to treatment by time $t = 2$ the “never treated”.
  - This is the case even if some of these units are eventually treated at time $t = 3$ (which we do not have access to this data yet).
- This is an abuse of notation but can help us with intuition.

Causality with potential outcomes in the canonical 2x2 DiD setup

- Let's focus on the **Canonical 2x2 setup**.
- There are n units available, $i = 1, 2, \ldots, n$.
- There are two time periods available, $t = 1$ and $t = 2$.
- A subset of all units are treated at time $g = 2$ (treated units), and a subset of units remain untreated at time $t = 2$, so $\mathcal{G} = \{2, \infty\}$.
- For units that are treated in time period $g = 2$, we observe $Y_{i,t=1}(2)$ and $Y_{i,t=2}(2)$.
- For the "never treated" units $g = \infty$, we observe $Y_{i,t=1}(\infty)$ and $Y_{i,t=2}(\infty)$.

"Traditionally", we call these potential outcomes $Y_{i,t}(1)$ and $Y_{i,t}(0)$, instead of $Y_{i,t}(2)$ and $Y_{i,t}(\infty)$. However, that notation is hard to extend to setups with variations in treatment timing.

Causality with potential outcomes in the canonical 2x2 DiD setup

- **Treatment Effect**
  - The treatment effect or causal effect of the treatment on the outcome of unit i at time t is the difference between its two potential outcomes:

$$Y_{i,t}(2) - Y_{i,t}(\infty)$$

- **Observed outcomes**
  - Observed outcomes at time t are realized as

$$Y_{i,t} = 1\{G_i = 2\}Y_{i,t}(2) + 1\{G_i = \infty\}Y_{i,t}(\infty).$$

- **Fundamental problem of causal inference**
  - At time t we cannot observe both potential outcomes $Y_{i,t}(2)$ and $Y_{i,t}(\infty)$.

Fundamental problem of causal inference: Missing data problem

Unit	Data				
	$Y_{i,t=1}(2)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$	G_i
1	?	?	✓	✓	∞
2	✓	✓	?	?	2
3	?	?	✓	✓	∞
4	✓	✓	?	?	2
⋮	⋮	⋮	⋮	⋮	⋮
n	✓	✓	?	?	2

✓: Observed data

?: Missing data (unobserved counterfactuals)

Causality with potential outcomes in the canonical 2x2 DiD setup

- **Problem:**
  - Causal inference is difficult because it involves missing data.
  - At time t, how can we find $Y_{i,t}(2) - Y_{i,t}(\infty)$?
- **"Cheap" solution - Rule out heterogeneity.**
  - $Y_{i,t}(2), Y_{i,t}(\infty)$ constant across units/time.
- **But Causal inference is all about heterogeneity.**
  - In these cases, the "cheap solution" doesn't work, and we need to find other paths.
  - We need to find more appealing assumptions.
  - We will talk about these soon!

Causal parameters of interest

Target parameters in the 2x2 DiD Setup

- Once we embrace treatment effect heterogeneity, recovering unit-specific treatment effects becomes hard, if not impossible.
- We will focus on causal effects in an average sense.
- **Let's first focus on the 2x2 DiD setup**.
- We will also focus on the effect in the post-treatment period. (Guess why?)

Parameters of interest in the 2x2 DiD Setup

- **ATT**

  The Average Treatment Effect on the Treated at time period $t = 2$ is

$$ATT = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = 2\right]$$

- **ATU**

  The Average Treatment Effect on the Untreated at time period $t = 2$ is

$$ATU = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = \infty\right]$$

- **ATE**

  The (overall) Average Treatment Effect at time period $t = 2$ is

$$ATE = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty)\right]$$

Parameters of interest in the 2x2 DiD Setup

These parameters answer different questions:

- **ATT**: What is the average effect of the policy/treatment among units that actually received the treatment by time $t = 2$?
- **ATU**: What is the average effect of the policy/treatment among units that did not receive the treatment by time $t = 2$ if they were to receive the treatment?
- **ATE**: What is the overall average effect of the policy/treatment if everybody were to be treated at time $t = 2$?

What if we have multiple groups?

Potential parameters of interest in the multi-group DiD setups

- **ATT(g, t)**
  The average treatment effect of being first-treated in period $g < \infty$ (compared to never-being treated), among units first-treated in period g, at time period t is

$$ATT(g,t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = g\right]$$

- **ATU(g, t)**
  The average treatment effect of being first-treated in period g (compared to never-being treated), among the never-treated units, at period t is

$$ATU(g,t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = \infty\right]$$

- **ATE(g, t)**
  The (overall) average treatment effect of being first-treated in period g (compared to never-being treated) at period t is

$$ATE(g,t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(\infty)\right]$$

But we do not need to fix the baseline!

Parameters of interest in the multi-group DiD setups

- **ATT(g', g, t|g^*)**
  The average treatment effect of switching first-treatment time from g' to g, among units first treated in period g', at time t:

$$ATT(g', g, t|g^*) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(g')|G_i = g^*\right]$$

- **ATU(g', g, t|∞)**
  The average treatment effect of switching first-treatment time from g' to g, among never-treated units, at time t is

$$ATU(g', g, t|\infty) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(g')|G_i = \infty\right] = ATU(g, t) - ATU(g', t)$$

- **ATE(g',g, t)**
  The (overall) a average treatment effect of switching first-treatment time from g' to g, at time period t is

$$ATE(g', g, t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(g')\right] = ATE(g, t) - ATE(g', t)$$

What if treatment can turn on and off?

What if treatment is multi-valued/continuous?

Exercise

Exercise with treatments turning on and off

- Time to check how well we follow the **principles** of building causal parameters in different setups.

- Let's consider a case with 3 time periods, $t = 1, 2, 3$.
  - At $t = 1$, no unit is treated.
  - At $t = 2$, some units are treated, and others remain untreated.
  - At $t = 3$, some previously treated units remain treated, and some turn treatment off. In addition, among not-yet-treated units, some remain untreated, but others become treated.

- Let $\boldsymbol{d} = (d_1, d_2, d_3)$ be a sequence of treatments, where $(d_1, d_2, d_3) \in \{0, 1\}^3$.

Exercise with treatments turning on and off

- **Question 1:** Define potential outcomes depending on potential treatment sequences.
- **Question 2:** Define the average treatment effect at time t of taking a specific treatment sequence compared to never being treated, among units that take that given specified treatment sequence.
- **Question 3:** Define the average treatment effect at time t of taking a specific treatment sequence compared to never being treated, among units that remained untreated until $t = 3$.
- **Question 4:** Define the overall average treatment effect at time t of taking a specific treatment sequence compared to never being treated.

Exercise with continuous and multi-valued treatments

- Now let's consider the case where treatment is continuous or multi-valued.
- For simplicity, let's focus on the case with 2 time periods, $t = 1, 2$.
  - At $t = 1$, no unit is treated (everybody with dose $d = 0$).
  - At $t = 2$, some units are treated with dose $d > 0$, and others remain untreated ($d = 0$).

Exercise with continuous and multi-valued treatments

- **Question 5:** Define potential outcomes depending on treatment dosages.
- **Question 6:** Define the overall average treatment effect at time $t = 2$ of receiving dosage d versus not receiving any treatment.
- **Question 7:** Define the overall average treatment effect at time $t = 2$ of receiving dosage d versus receiving dosage d'.
- **Question 8:** Define the average treatment effect at time $t = 2$ of receiving dosage d versus not receiving any treatment, among units who received dosage d.

Exercise with continuous and multi-valued treatments

- **Question 9:** Define the average treatment effect at time $t = 2$ of marginally increasing treatment dosage d, among units who received dosage d. Discuss the discrete and continuous cases separately.

- **Question 10:** Define the overall average treatment effect at time $t = 2$ of marginally increasing treatment dosage d. Discuss the discrete and continuous case separately.

- **Question 11:** The above marginal average treatment effects are "local" to a dosage d. Can you think of a more aggregate treatment effect measure that may summarize the above marginal average treatment effects across different dosages d?

References

Assunção, Juliano, Clarissa Gandour, Romero Rocha, and Rudi Rocha, "The Effect of Rural Credit on Deforestation: Evidence from the Brazilian Amazon," *Economic Journal*, 2020, *130* (626), 290–330.

Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Beck, Thorsten, Ross Levine, and Alexey Levkov, "Big bad banks? The winners and losers from bank deregulation in the United States," *Journal of Finance*, 2010, *65* (5), 1637–1667.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Card, David and Alan Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 1994, *84* (4), 772–793.

Carey, Colleen M., Sarah Miller, and Laura R. Wherry, "The Impact of Insurance Expansions on the Already Insured: The Affordable Care Act and Medicare†," *American Economic Journal: Applied Economics*, 2020, *12* (4), 288–318.

Currie, Janet, Henrik Kleven, and Esmée Zwiers, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, May 2020, *110*, 42–48.

Dube, Arindrajit, T William Lester, and Michael Reich, "Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties," *The Review of Economics and Statistics*, 2010, *92* (4), 945–964.

_ , T. William Lester, and Michael Reich, "Minimum wage shocks, employment flows, and labor market frictions," *Journal of Labor Economics*, 2016, *34* (3), 663–704.

Goldsmith-Pinkham, Paul, "Tracking the Credibility Revolution across Fields," *arXiv:2405.20604*, 2024.

Malesky, Edmund J., Cuong Viet Nguyen, and Anh Tran, "The impact of recentralization on public services: A difference-in- differences analysis of the abolition of elected councils in Vietnam," *American Political Science Review*, 2014, *108* (1), 144–168.

Meyer, Bruce D., W. Kip Viscusi, and David L. Durbin, "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *The American Economic Review*, 1995, *85* (3), 322–340.

Venkataramani, Atheendar S., Sachin J. Shah, Rourke O'Brien, Ichiro Kawachi, and Alexander C. Tsai, "Health consequences of the US Deferred Action for Childhood Arrivals (DACA) immigration programme: a quasi-experimental study," *The Lancet Public Health*, 2017, *2* (4), e175–e181.