Day 2: Policy evaluation and impact assessment

Quasi-experiments

Simon Munzert
Hertie School

Table of contents

1. Quasi-experiments
2. Instrumental variables
3. Interrupted time-series
4. Regression discontinuity
5. Difference-in-differences
6. Trade-offs in impact evaluation

Quasi-experiments

What represents a quasi-experiment?

Quasi-experiments

- Treatment assignment is **not under researcher's control** ↔ "controlled" experiment, RCT
- Treatment assignment follows a **random or as-if random** process; exogenous to outcome
- Construction of **treatment and control group post hoc** (and not always obvious)
- Often also referred to as **natural experiments** because treatment assignment is induced by nature

Different statistical approaches to exploit quasi-experiments

- Instrumental variables (IV)
- Difference-in-differences (DID)
- Regression discontinuity design (RDD)
- Interrupted time-series (ITS)
- Synthetic control
- ...

Example: The Vietnam draft lottery

Scholarly interest

- Economists' interest in the causal effect of military service on earnings (Angrist 1990), health (Angrist et al. 1996), political attitudes (Erikson and Stoker 2011)
- Problem: potential confounding due to self-selection into military service

$\rightarrow$ If those who volunteer for military service are different from those who do not, then the estimated effect of military service on earnings, health, or political attitudes is biased.

The Vietnam draft lottery 1969 as randomization device

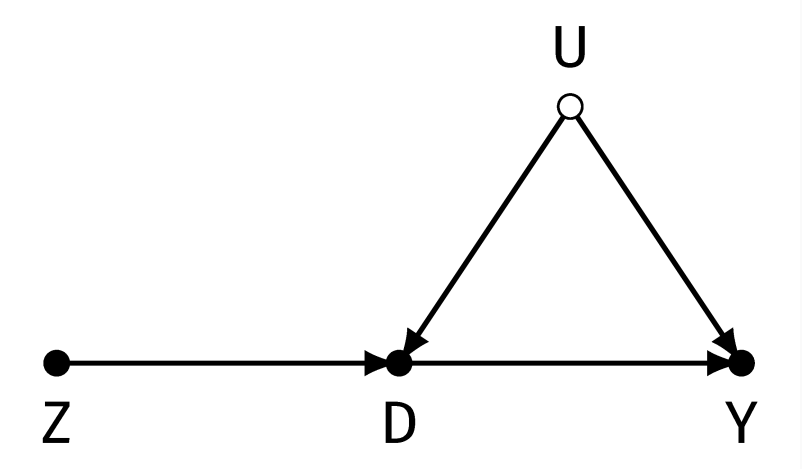
- From 1970 to 1972, random sequence numbers were assigned to each birth date in cohorts of 19-year-olds).
- Men with lottery numbers below a cutoff were sequentially drafted while men with numbers above the cutoff could not be drafted.
- Noncompliance issues! The draft did not perfectly determine military service: many draft-eligible men were exempted for health and other reasons; exempted men volunteered for service.

Instrumental variables

Basic logic of instrumental variables

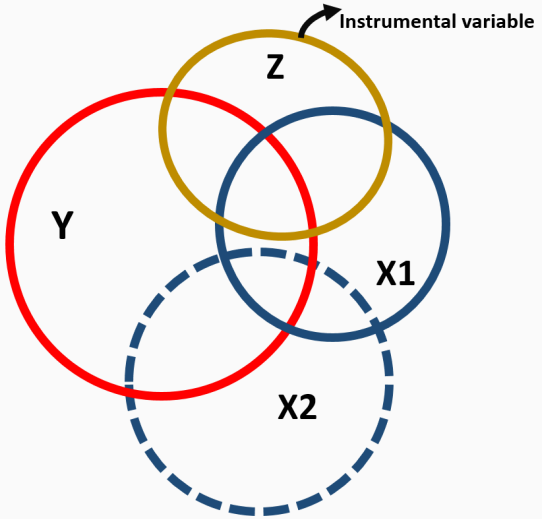
The setting

- Interest is in the causal effect of D on Y
- The relationship is subject to potential confounding
- There is a variable Z (the instrumental variable) that is (as-if) randomly assigned and related to D



Basic idea of instrumental variables (IV):

- Split the variation of D into two parts:
  - one endogenous (related to confounders U)
  - one truly exogenous
- The part of the variation in D that is related to Z but not to U is then used to estimate the causal effect of D on Y.



Sources of IV designs

Table 4.1 Selected sources of instrumental-variables designs

Source of instrumental variable	Units in study group	Treatment variable	Outcome variables
Lotteries			
Military drafts	Soldiers	Military service	Earnings, attitudes
Prize lotteries	Lottery players	Overall income	Political attitudes
Judge lotteries	Prisoners	Prison terms	Recidivism
Training invitations	Job-seekers	Job trainings	Wages
School vouchers	Students	Private-school attendance	Educational achievement
Weather shocks			
Rainfall growth	Countries	Economic growth	Civil war
Natural disasters	Countries	Oil prices	Democracy
Age			
Quarter-of-birth	Students	Education	Earnings
Twin studies			
Twin births	Mothers	Number of children	Earnings
Institutional variation			
Electoral cycles	States	Police presence	Crime
Land tenure types	States	Inequality	Public goods
Historical shocks			
Deaths of leaders	Countries	Colonial annexation	Development
Colonial settler mortality	Countries	Current institutions	Economic growth

`Source`: Dunning, Thad. 2012. Natural experiments in the social sciences: A design-based approach. CUP.

Strengths and weaknesses of IV

Strengths

- With a valid instrument Z at hand, you (probably) do not have to account for any other confounders of D and Y
- Instrumental variables provide an established way to exploit random or as-if random variation in treatment assignment, putting observational studies closer to the experimental ideal-type design

Weaknesses

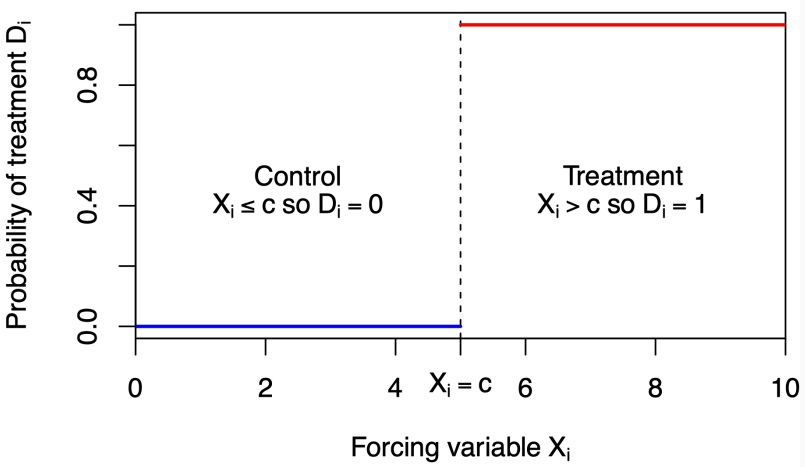
- Rests on strong assumptions that are ultimately untestable; therefore a good story is the crucial element of any plausible IV specification
- IV estimators have been demonstrated to be biased until N is very large even if the assumptions hold, partly due to the fact that supposedly exogenous portions of D are themselves estimates that come along with measurement error
- Good instruments are extremely difficult to find. IV research tends to be driven by opportunistic choice of IV, not by the research question - hardly a good starting point for impact evaluation

Regression discontinuity

Basic logic of regression discontinuity

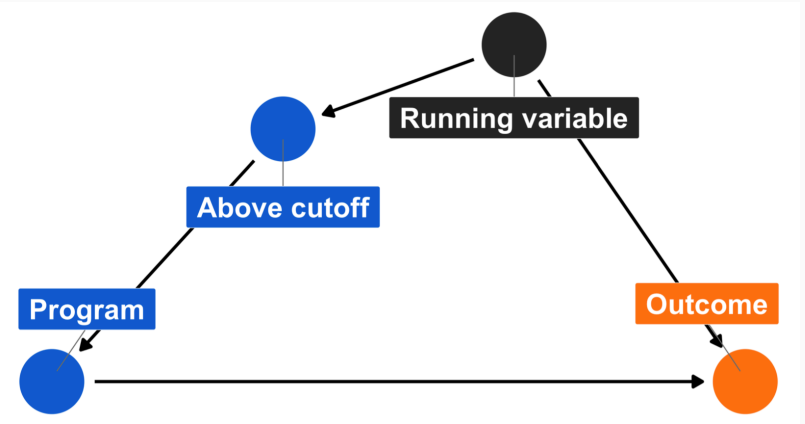
The setting

- A threshold is used to assign treatment/policy
- Examples: admission to a program, taxation, election results, etc.
- Treatment is assigned according to a rule based on another variable (called the **forcing or running variable**)

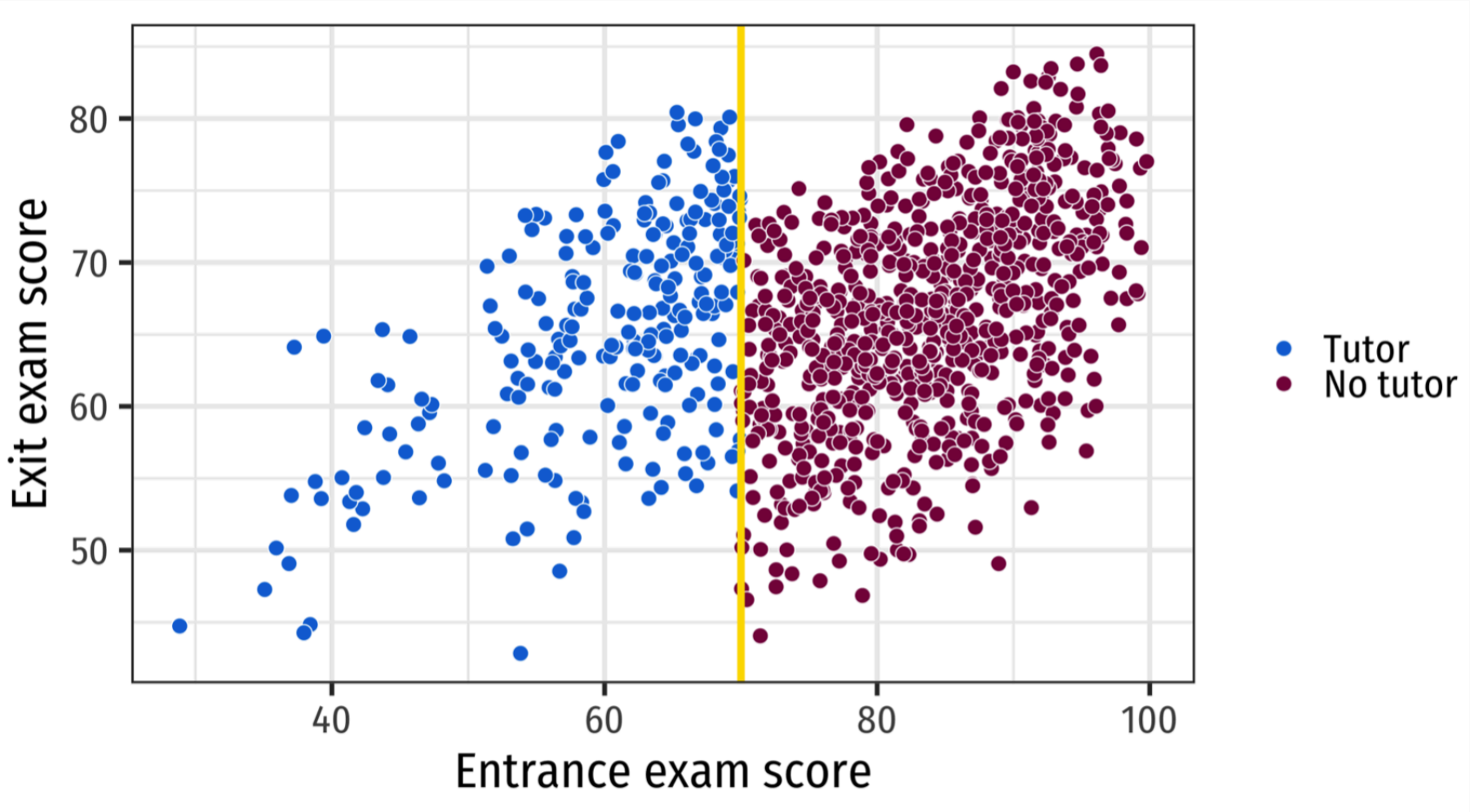


Basic idea of regression discontunity (RDD):

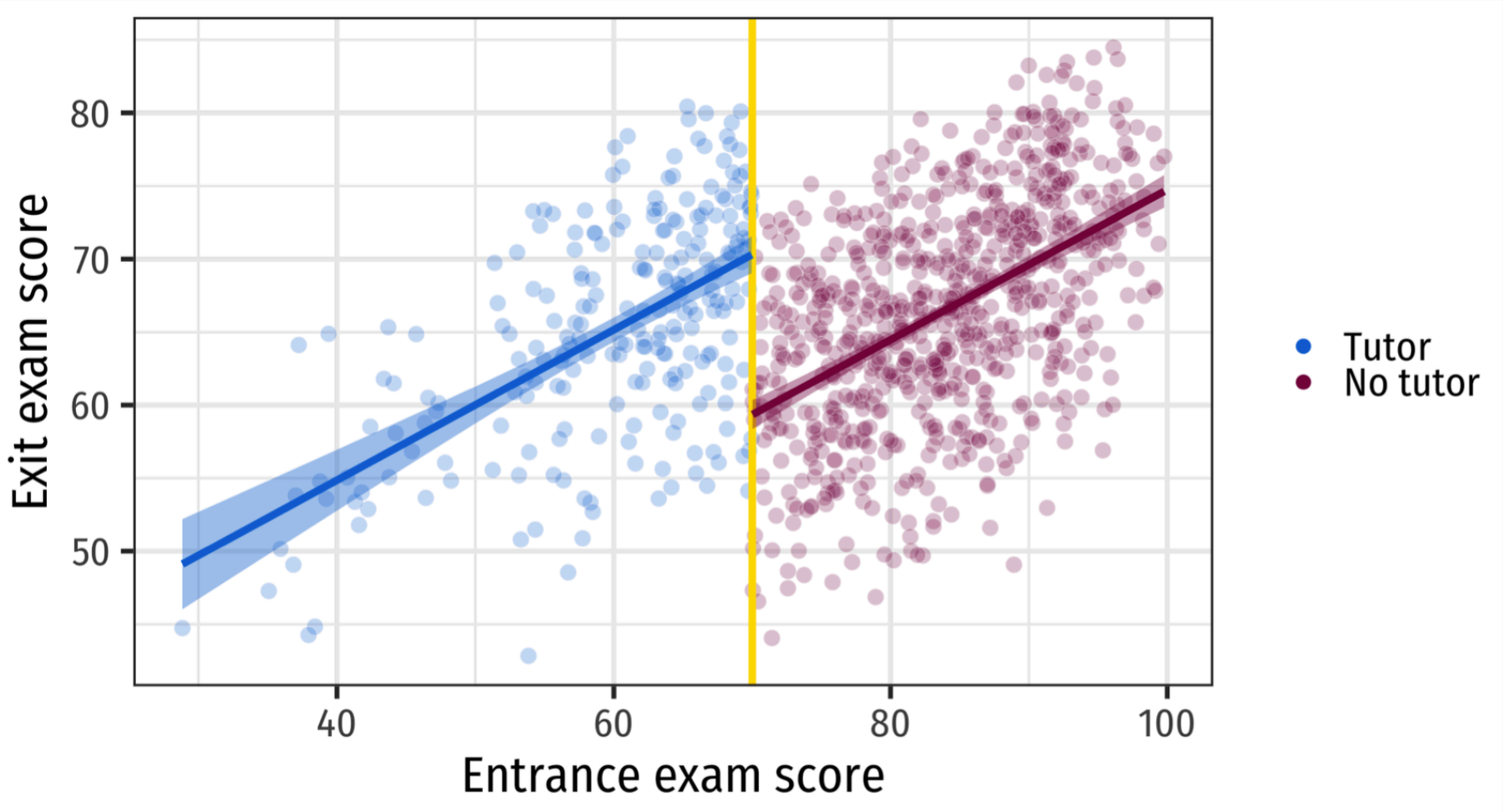
- Whether units end up just below or just above the threshold is assumed to be a matter of chance (**local randomization**)
- A treatment effect estimate is obtained by comparing (predicted) Y-values of those just eligible for treatment with those just ineligible
- Often useful for **analysis in and of a rule-based" world** (administrative programs, elections, etc.)



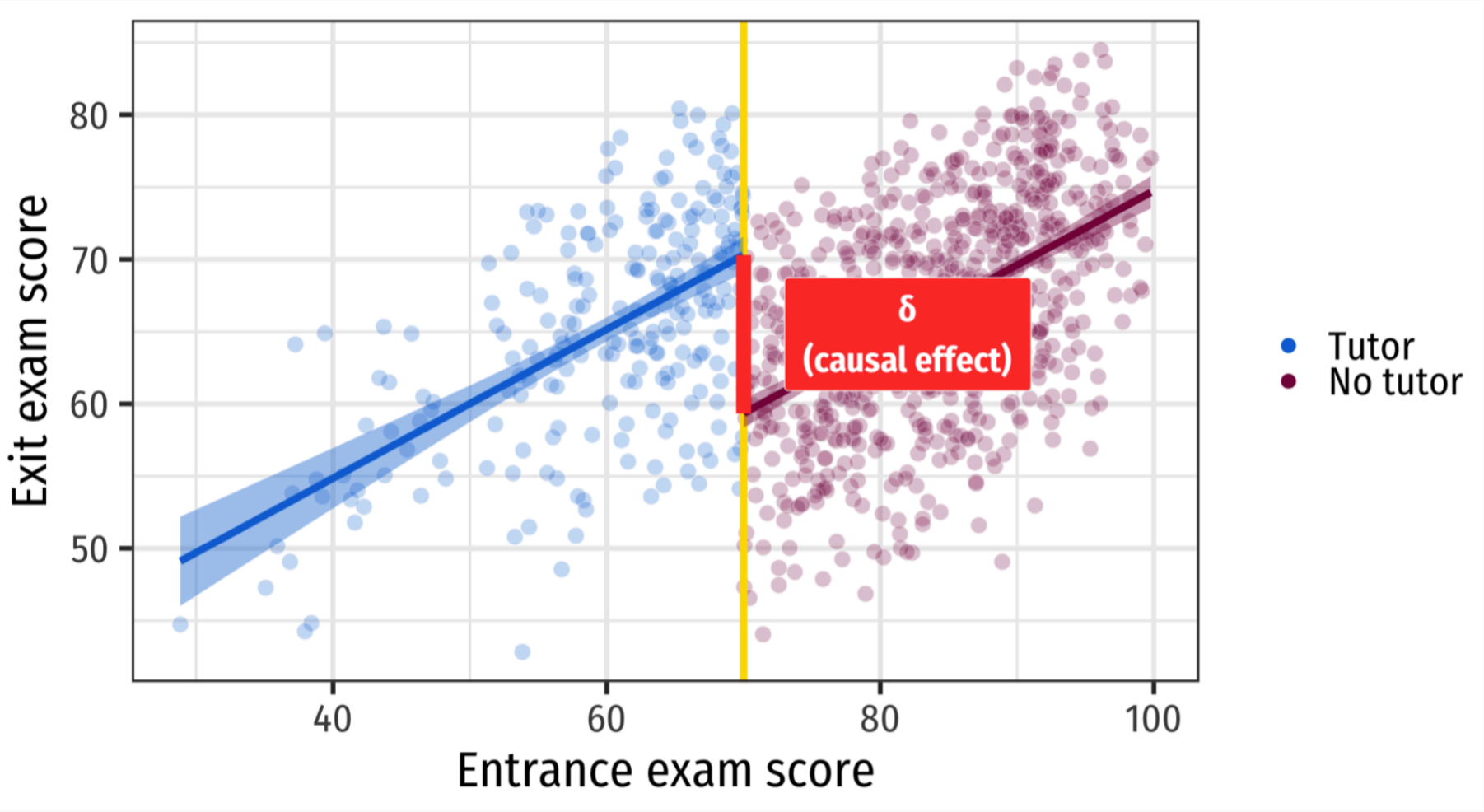
RDD: Example


Exit exam score
80
70
60
50
40
60
80
100
Entrance exam score
Tutor
No tutor

RDD: Example


Exit exam score
80
70
60
50
40
60
80
100
Entrance exam score
Tutor
No tutor

RDD: Example


Exit exam score
80
70
60
50
40
60
80
100
Entrance exam score
δ
(causal effect)
Tutor
No tutor

Sources of RD designs

Table 3.1 Selected sources of regression-discontinuity designs

Source of RD design	Units in study group (at RD threshold)	Treatment variables	Outcome variables
Entrance exams	Students, others	Public recognition of scholastic achievement	Educational achievement
Population thresholds	Municipalities, citizens	Voting technologies	Effective turnout
		Federal funds	Voting behavior
		Cash transfers	Voting behavior
		Electoral rules	Voting behavior
		Politicians' salaries	Candidate entry
Size-based thresholds			
Voter numbers	Voters	Voting by mail	Voting behavior
School size	Students	Class size	Educational achievement
Firm size	Firms	Antibias laws	Productivity
Eligibility criteria			
Poverty rank	Municipalities	Antipoverty programs	Voting behavior
Criminality index	Prisoners	High-security incarceration	Recidivism
Age-based thresholds			
Voting age	Voters	Past voting	Turnout
Birth quarter	Students	Years of education	Earnings
Close elections	Candidates/parties	Incumbency	Candidates' performance
	Firms	Campaign donations	Public works contracts

`Source`: Dunning, Thad. 2012. Natural experiments in the social sciences: A design-based approach. CUP.

Strengths and weaknesses of RDD

Strengths

- **High internal validity:** With enough data available around the threshold and in the absence of sorting, the observational design is pretty close to the experimental ideal
- Intuitive and easy to understand
- Thresholds are everywhere in a policy-fueled world - lots of potential for application

Weaknesses

- **Limited external validity:** Effect is only identified for a small subpopulation → should we expect the same effect among more remote subjects?
- RDD requires lots of data (or extrapolation); sensitivity to choice of functional form
- Subjects potentially know the threshold and may sort themselves into treatment and control groups

Interrupted time-series

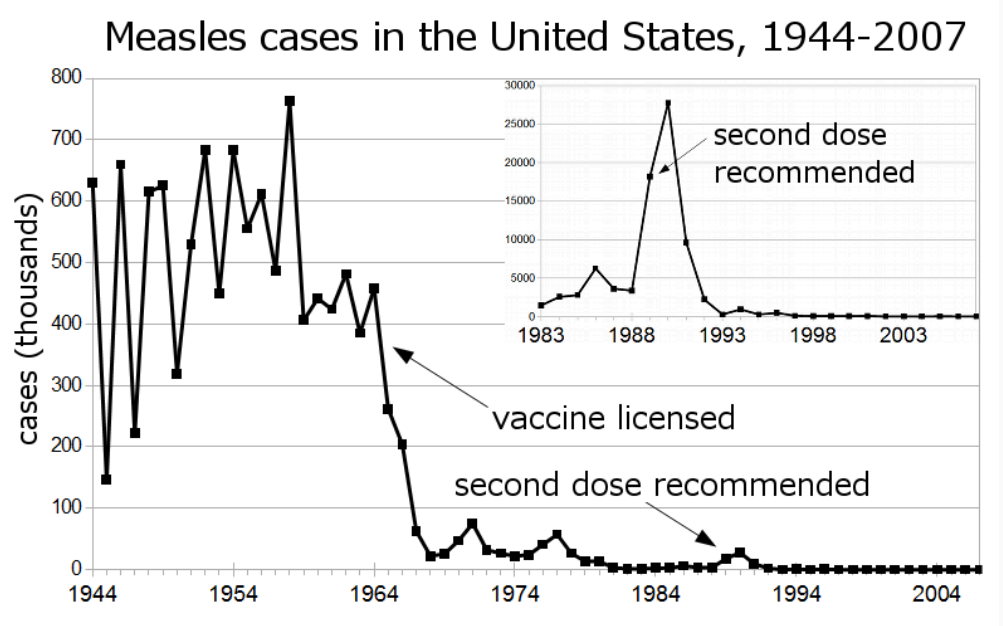
Basic logic of interrupted time-series

The setting

- Time-series data allow observing the same subject or unit in different causal states at different points in time
- The treatment is introduced at a known point in time (the interruption")

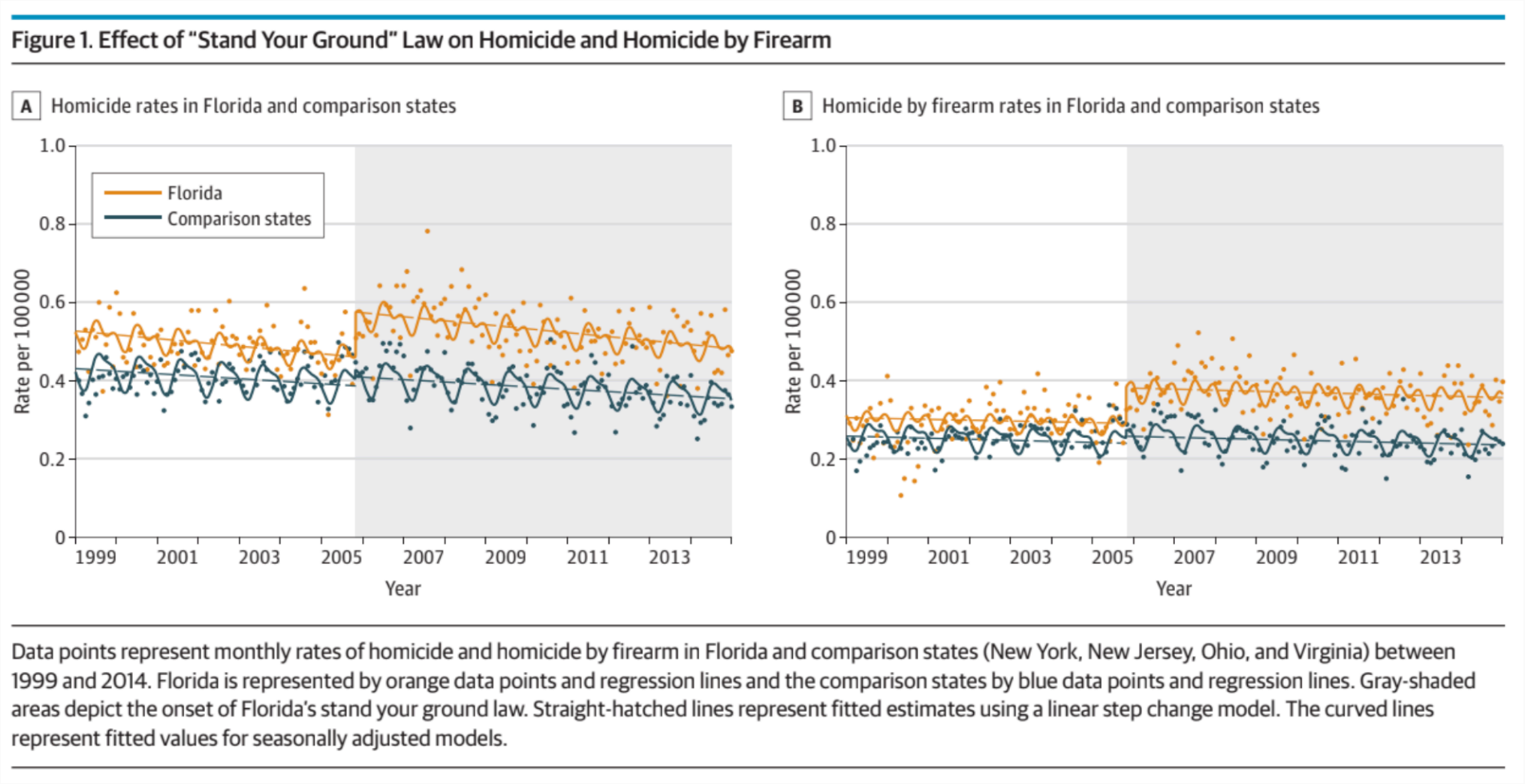
The logic of interrupted time-series (ITS)

- The effect of the treatment is estimated by comparing the pre- and post-interruption trends in the outcome variable
- Basic ITS model: $Y_t = f(t) + \beta D_t + e_t$
- In words: Y is some function in time $f(t)$ and the treatment indicator D
- Analogy to RDD with time as the running variable



Example: The effect of "Stand Your Ground" Law on homicide

Hertie School



`Source` Humphreys et al. (2017), JAMA Internal Medicine.

Strengths and weaknesses of ITS

Strengths

- Similar to RDD
- Intuitive and easy to understand
- With policies, a starting (or end) date often can be easily defined

Weaknesses

- Time is a problematic running variable: many things can happen at the same time as the policy change
- Policies can be anticipated, which contaminates effects around the interruption

Strategies to increase the confidence in ITS inference

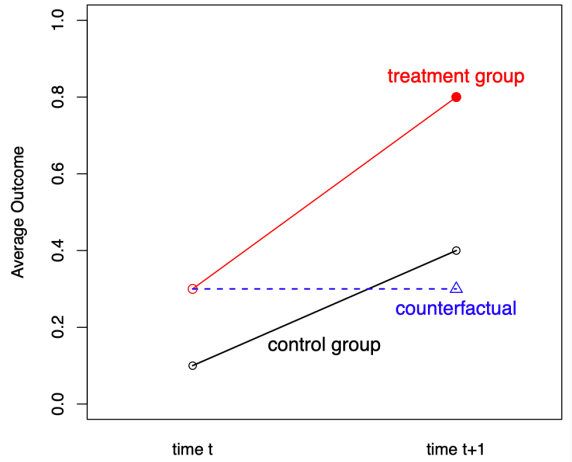
- Estimate effect on **alternative outcomes** that should (not) be affected by the treatment (latter case: "placebo tests")
- Adjust for **trends in other variables** that may affect or be related to the underlying time series of interest
- Assess the **impact of the termination** of the cause in addition to its initiation

Difference-in-differences

Basic logic of difference-in-differences

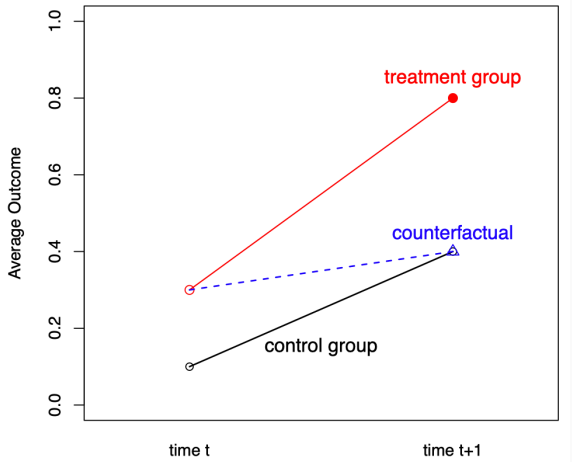
The setting

- How should we conduct causal inference when repeated measurements of treateds and controls are available?
- Simplest case: Before-after measure of treatment and control group
- Two types of variations:
  - Time variation: change over time within each group
  - Group/cross-sectional variation: difference between treatment and control group within each time period
- Before-and-after and cross-sectional designs

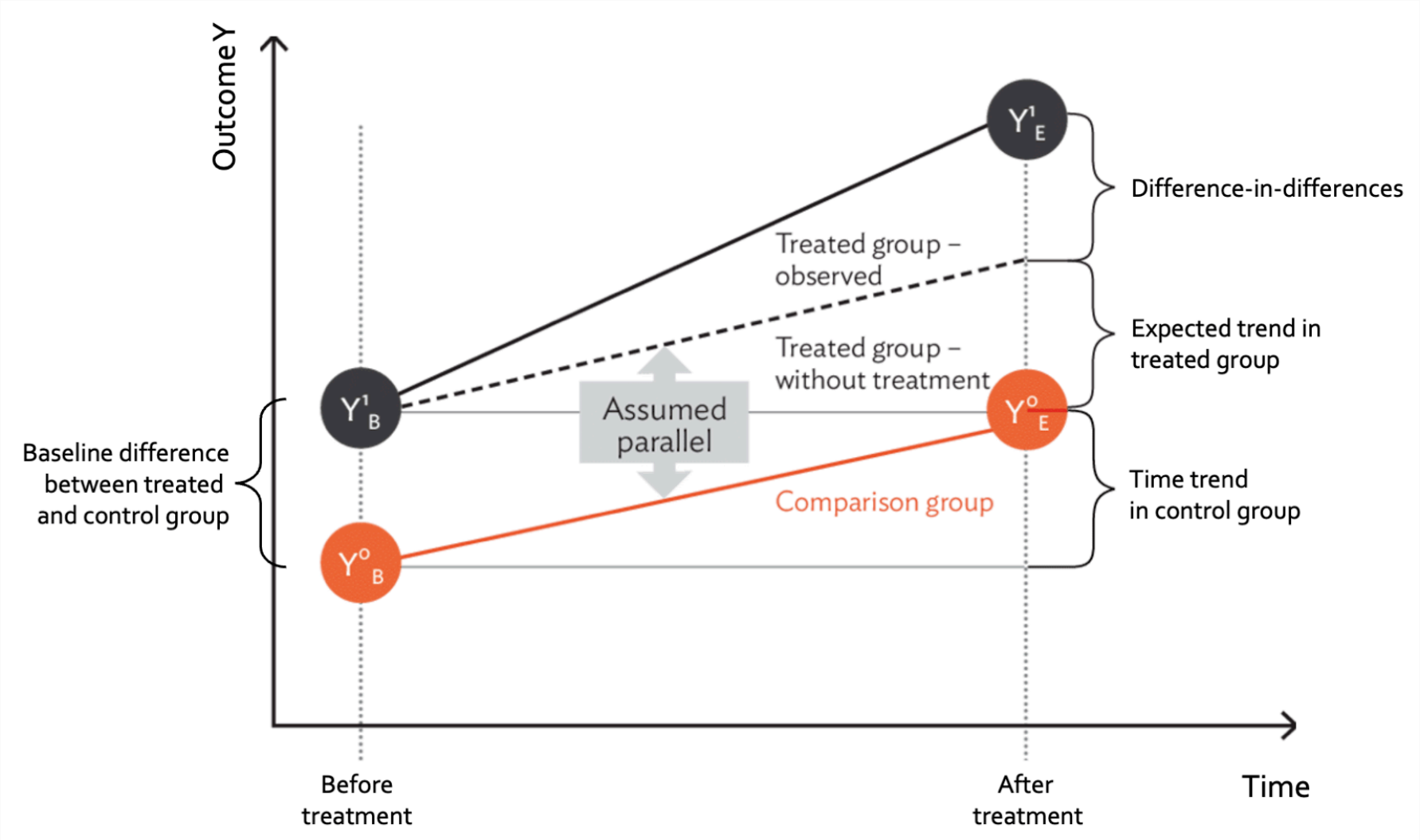


The DID idea

- Can we exploit both variations?
- That's what Difference-in-Differences (Diff-in-diff, DID, DD) tries to achieve
- The DID estimator is the difference between the change in the treatment group and the change in the control group



Visual illustration of DID



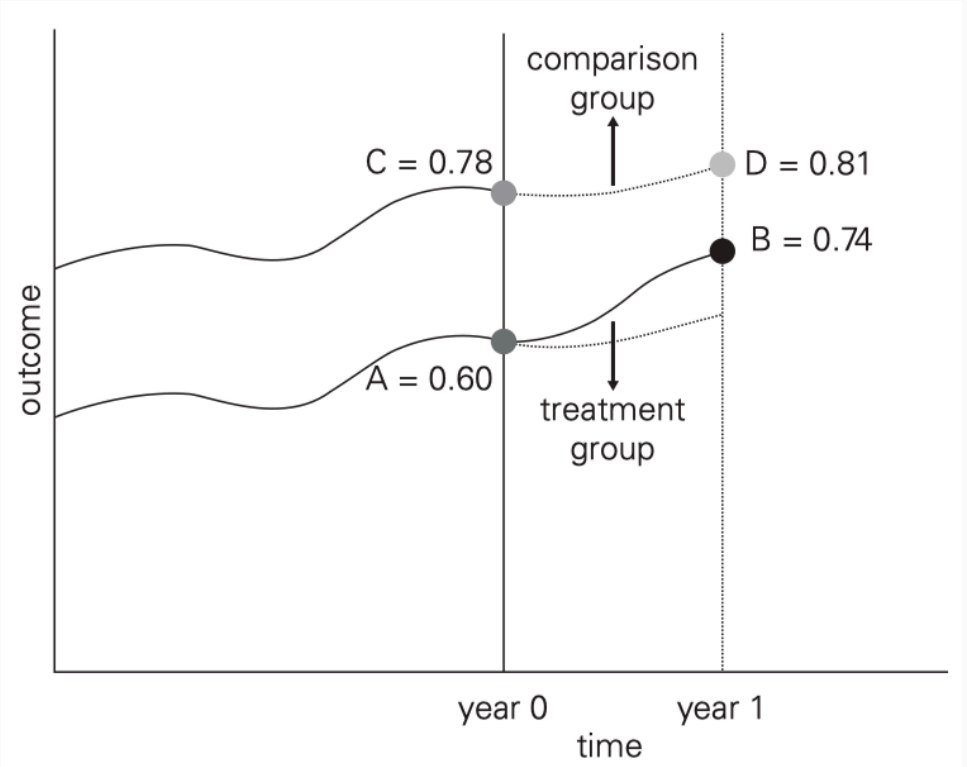
Example: Garbage collection and recycling (fictitious case)

The scenario

- Several boroughs within a large metropolitan area send out information material on value/costs of garbage recycling system. (Not a city-wide campaign!) → the **treatment group**
- Other boroughs do business as usual → the **control group**

Data collection

- Ex-post, we analyze fraction of "correctly" discarded garbage (placed in appropriate recycling boxes)
- Data obtained from City's garbage collection service, in the pre- and post-reform year, for treat. vs control boroughs



How to evaluate the impact of the intervention?

Example: Garbage collection and recycling (fictitious case)

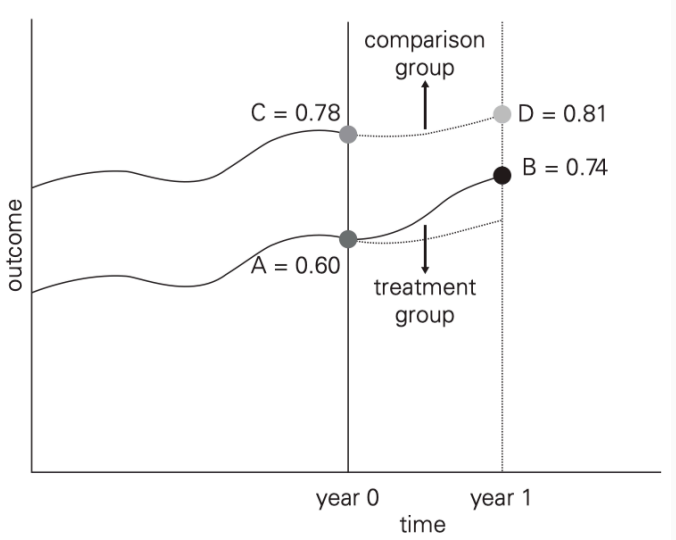
Hertie School

Simple pre-post comparison

- Comparing average outcomes in the treatment group before (year 0) and after the policy change (year 1)
- Result: $B - A = 0.74 - 0.60 = 0.14 \rightarrow$ misleading result!

Ex-post "between" comparison

- Comparing year 1 outcomes between treatment and control group
- Result: $B - D = 0.74 - 0.81 = -0.07 \rightarrow$ misleading result!



Example: Garbage collection and recycling (fictitious case)

Simple pre-post comparison

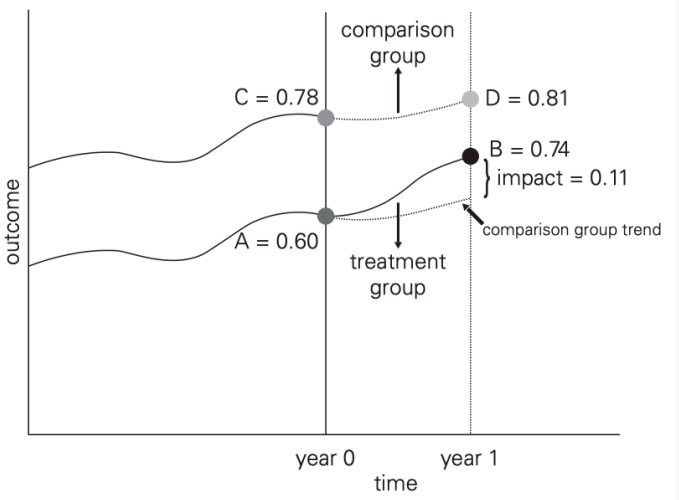
- Comparing average outcomes in the treatment group before (year 0) and after the policy change (year 1)
- Result: $B - A = 0.74 - 0.60 = 0.14 \rightarrow$ misleading result!

Ex-post "between" comparison

- Comparing year 1 outcomes between treatment and control group
- Result: $B - D = 0.74 - 0.81 = -0.07 \rightarrow$ misleading result!

Compute Difference-in-Differences

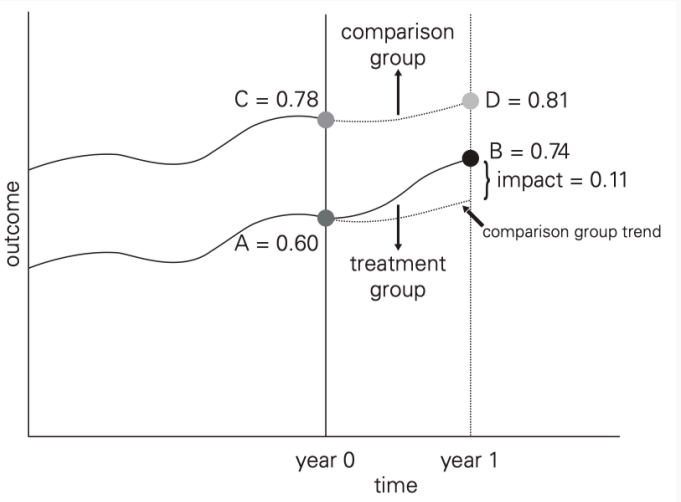
- Comparing change ("difference") in outcomes in the treatment group with changes in outcomes in the control group
- DID result = $(B - A) - (D - C) = 0.11$
- Causal impact measure (under certain important assumptions)



Example: Garbage collection and recycling (fictitious case)

DID identifies "true" causal impact only under...

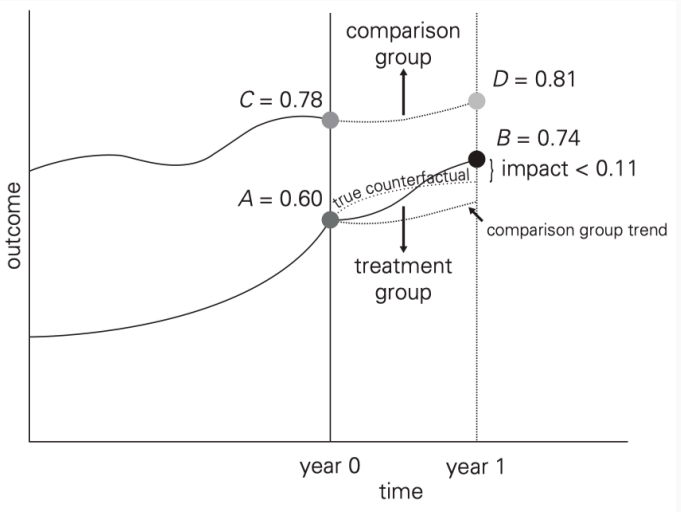
- **Parallel trends**: W/o policy change the outcome in the treatment group would have evolved perfectly "parallel" to outcome in control group (recall: the unobserved, perfect counterfactual!) - Tricky: not testable, one can only check parallel pre-trends
- Group **compositions remain stable** during evaluation period
- **No spillover** from treatment into control group (or vice versa)



Example: Garbage collection and recycling (fictitious case)

DID identifies "true" causal impact only under...

- **Parallel trends**: W/o policy change the outcome in the treatment group would have evolved perfectly "parallel" to outcome in control group (recall: the unobserved, perfect counterfactual!) - Tricky: not testable, one can only check parallel pre-trends
- Group **compositions remain stable** during evaluation period
- **No spillover** from treatment into control group (or vice versa)



Discussion

- Which regions engaged in policy reform? What does that imply for Diff-in-Diff?
- How could the design of the intervention have been strengthened a priori?

Puzzle

What happened?

1. What do you think happened to the man in the picture?

2. What is the likely cause of the observed skin damage?

3. What does this have to do with quasi-experiments?

What happened?

1. What do you think happened to the man in the picture?
2. What is the likely cause of the observed skin damage?
3. What does this have to do with quasi-experiments?

Some more information

> "A 69-year-old man presented with a history of gradual thickening and wrinkling of the skin on the left side of his face. The physical examination showed hyperkeratosis with accentuated ridging, multiple open comedones, and areas of nodular elastosis." (from the Abstract)

- Source: J. Gordon and J. Brieva, NEJM, 2012, Unilateral Dermatoheliosis.
- The patient reported that he had driven a delivery truck for 28 years. Ultraviolet A (UVA) rays transmit through window glass, penetrating the epidermis and upper layers of dermis.

Trade-offs in impact evaluation

The internal vs. external validity trade-off

Internal validity

"What works (in the context of the study's parameters)?"

- How well does the study design rule out rival explanations for the observed effects?
- Is the constructed counterfactual plausible? Is it even a good point for comparison?
- Randomization fixes a host of issues (think: confounding!), but not everything (think: attrition, noncompliance, contamination, measurement).

External validity

"Is this relevant? Does this work for us?"

- Is the outcome of interest measured in a way that is relevant for the target population?
- Is the treatment of interest comparable to the policy intervention of interest? (think: what works in mice might not work in humans)
- More generally: How well do the results generalize to other settings, populations, and times?
- Relevance for planning and policy-making requires the generalization of results to a target population of interest.

What is useful causal evidence?

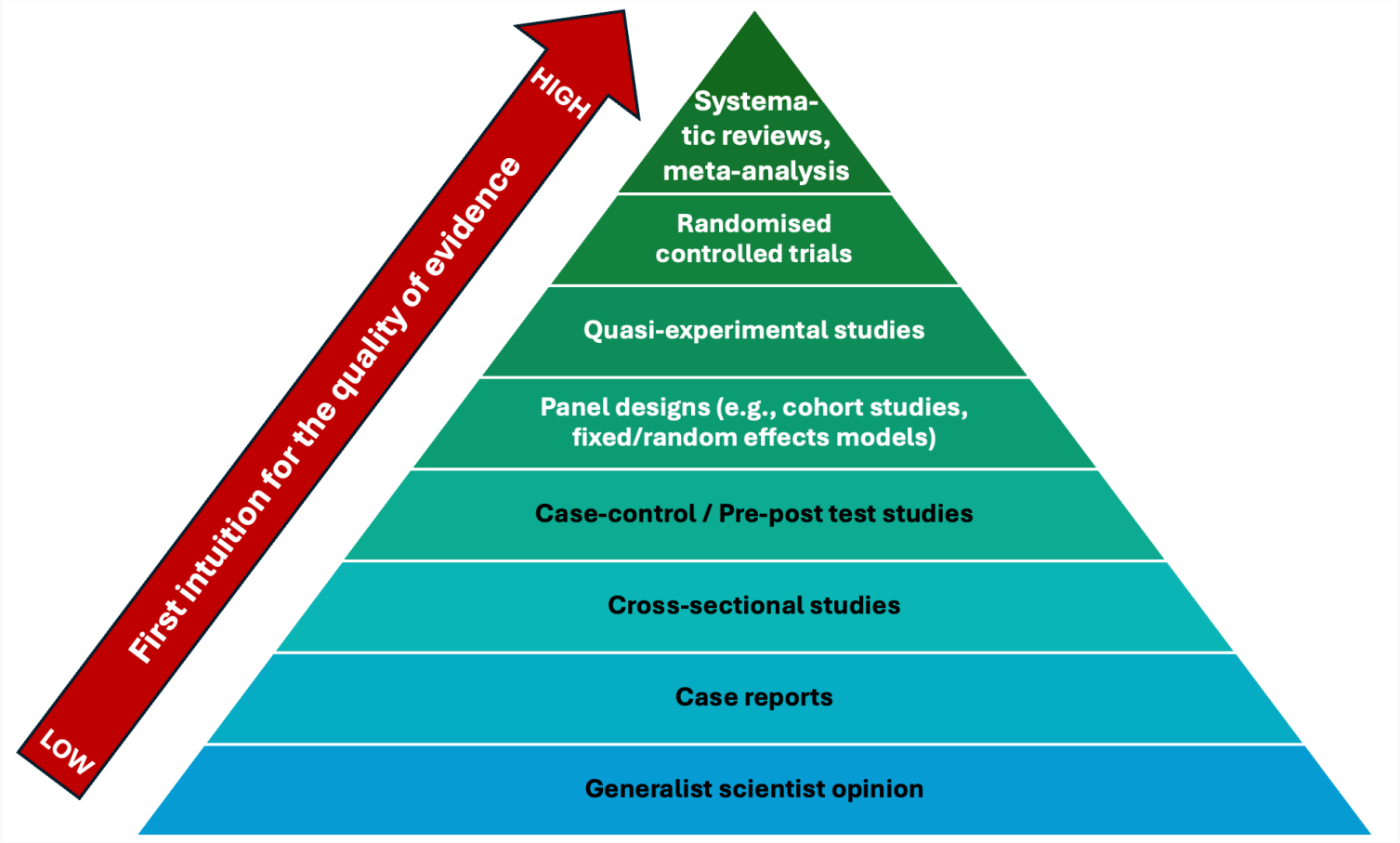
Practical usefulness

- **Internal validity** is about ruling out rival explanations for the observed effects. Importance of balancing study design decisions to maximize both internal and external validity, but scoring high on both is often impossible to achieve.
- Accepting **"practical usefulness" as additional criterion** next to internal and external validity.
- Internal and external bias both exist on a continuous scales (as degrees rather than as dichotomies), and relatively minor violations of internal validity may be tolerable in exchange for greater external validity.

Staying humble

- Different methods come with different pros and cons; not every method fits every context
- Causal inference using observational data is hard; be humble when making causal claims!
- Generalization using experimental data is hard; be humble when making claims about the population of interest!

Hierarchy of evidence*



***Take with a pinch of salt!** First intuition only; conflates internal and external validity and may not account for practical usefulness; lots of within-design variation of quality.