# Lecture 16

## Introduction to causal analysis

Julian Reif
Fall 2025

# Final exam schedule

- University has posted the final exam schedule:
  - https://registrar.illinois.edu/courses-grades/final-exam-schedule-public

- 9:30am section: 8am, December 15

- 11:00am section: 8am, December 17

# Midterm summary

- Average: 109/160

- 25th percentile: 94/160

- Curving will be done at the end of the semester. Recall from syllabus:
  - At least 25% receive an A, at least 50% receive an A-, at least 75% receive a B+

# Machine learning vs causal analysis (econometrics)

- Machine learning: use data to **predict** an outcome based on known variables
  - How does $Y$ *correlate* with $X$?

- Econometrics: use data to estimate **causal** effects
  - How does $Y$ change when $X$ is *manipulated*?

# Machine learning vs causal analysis

- Some problems only require prediction
  - Identifying potential customers for targeted sales
  - Recommendation systems (e.g. Netflix, Pandora)
  - Logistics and supply chain management (e.g. anticipating demand)

- Other problems involve **changing** the environment to achieve an effect
  - Does spending more on advertising lead to higher sales?
  - How does a firm's earnings management affect stock price?
  - Does education increase earnings?
  - Will health insurance make Americans healthier?

# Causality is fundamental in economics and computing

## The Prize in Economic Sciences 2021

The Royal Swedish Academy of Sciences has decided to award the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

with one half to

**David Card**
University of California, Berkeley, USA

*"for his empirical contributions to labour economics"*

and the other half jointly to

**Joshua D. Angrist**
Massachusetts Institute of Technology, Cambridge, USA

**Guido W. Imbens**
Stanford University, USA

*"for their methodological contributions to the analysis of causal relationships"*

## JUDEA PEARL

United States – 2011



The "Nobel Prize of computing"

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

# Causal inference tools are in high demand

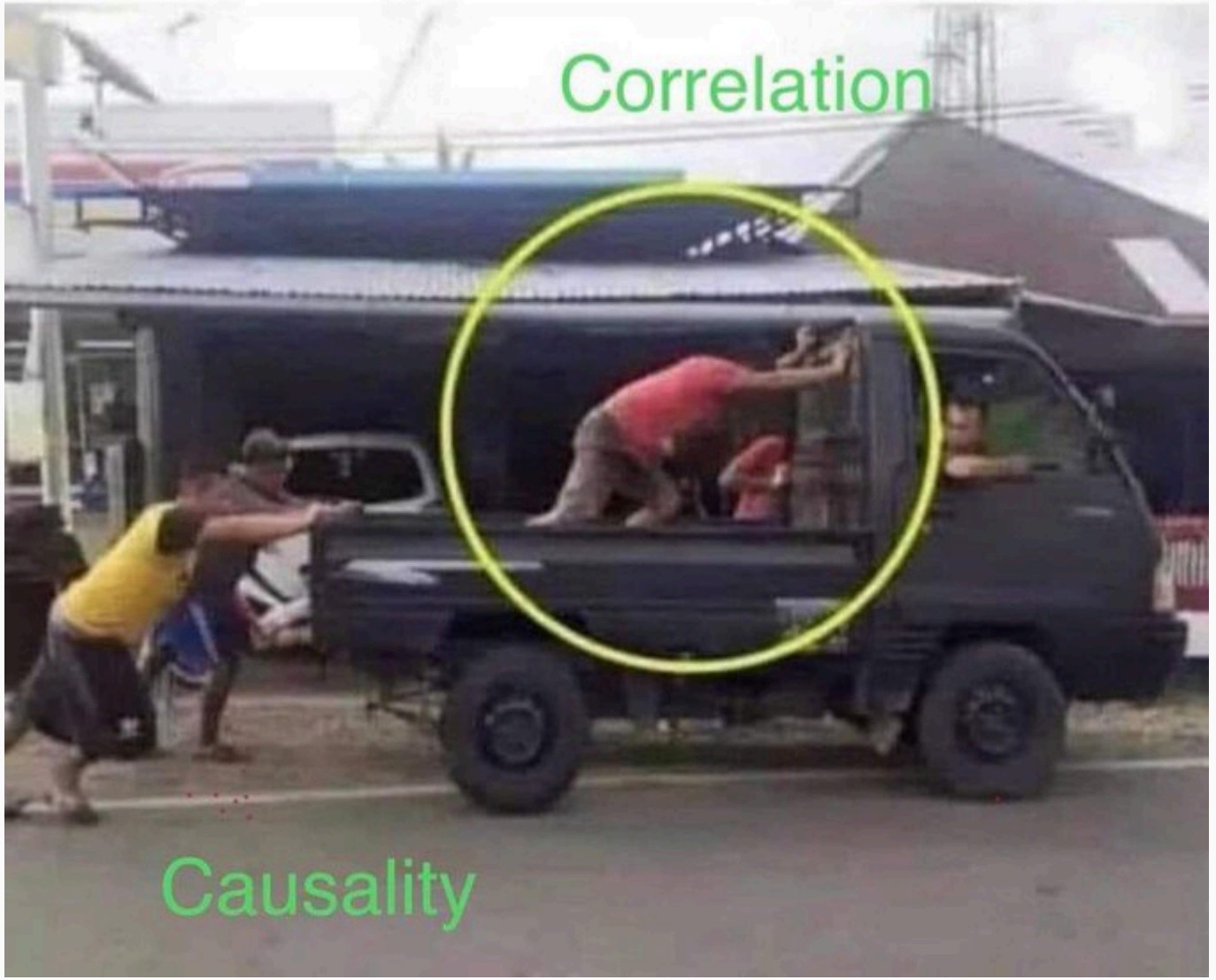
> "At Uber Labs, we apply behavioral science insights and methodologies to help product teams improve the Uber customer experience. One of the most exciting areas we've been working on is causal inference."
>
> --- Uber Blog

uber.com/en-AU/blog/causal-inference-at-uber

> "The integration of causal inference into engineering systems can lead to large amounts of new innovation."
>
> --- Netflix TechBlog

netflixtechblog.com/computational-causal-inference-at-netflix-293591691c62

# Correlation does not imply causation



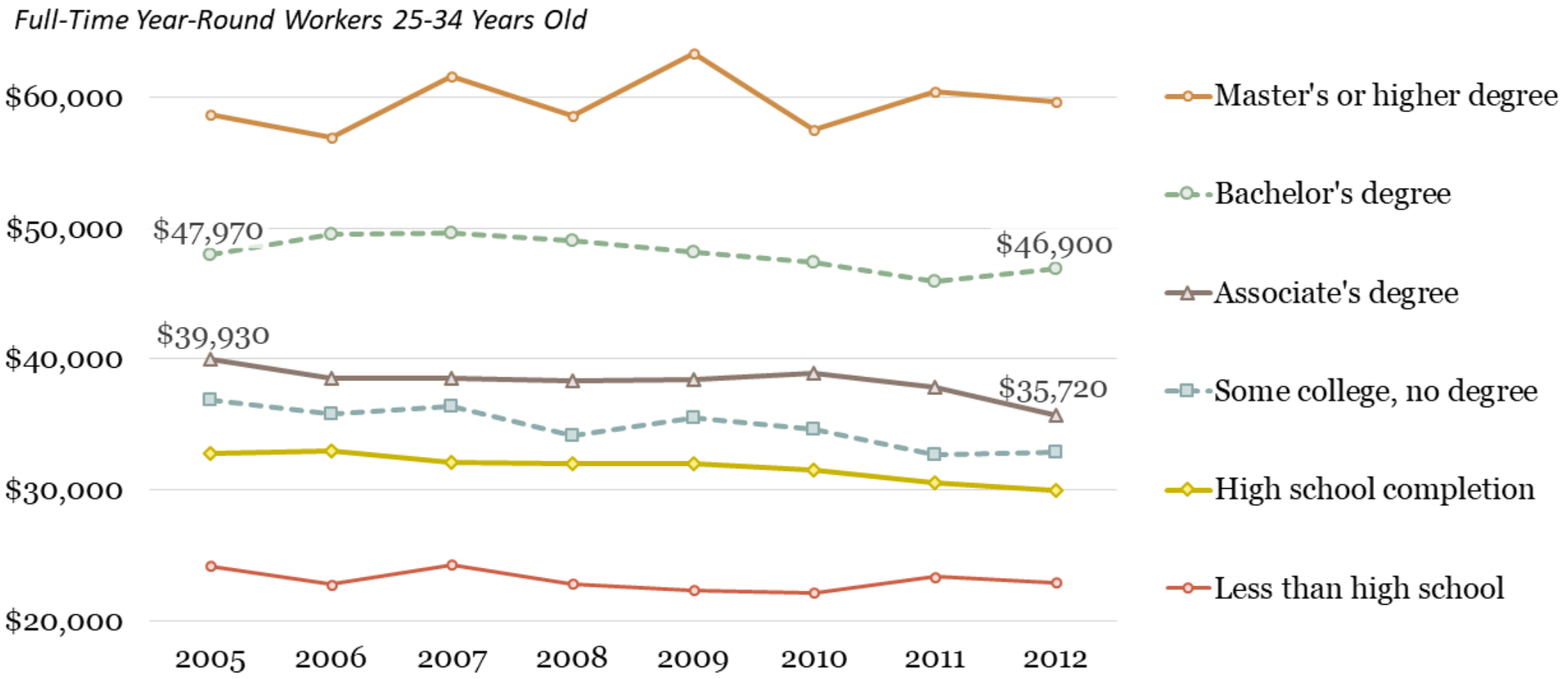
# Are these causal relationships or just correlations?

- There are more shark attacks when ice cream sales are higher
- There are more policemen in neighborhoods with higher crime
- People who spend more time indoors are sicker

# Introduction to causal analysis

- What does it mean for a correlation to have a causal interpretation?
- When do correlations have a causal interpretation?
- We will use three analytical tools for using data to measure causal connections
  1. Randomized experiments
  2. Difference-in-differences
  3. Regression discontinuity
- Note: linear regression can be used to implement all three of these methods

# Annual earnings by education (2015)




Source: January 2015 Salary Survey, National Association of Colleges and Employers

# Understanding these earnings differentials is important

- Policymakers care
  - Should we encourage people to go to college / graduate school?
  - Are students borrowing too much? Too little?
  - Should government facilitate student borrowing and ease the cost of loans?

- People care
  - Does going to college (or graduate school) pay off?
  - What degree should YOU pursue?

# The CAUSAL question for college

- Does college cause individuals to earn more?
  - Yes? (And if so, how much?)
  - Or no?

- This question is surprisingly difficult to answer

# What do these individuals have in common?

# Answering causal questions is often quite difficult

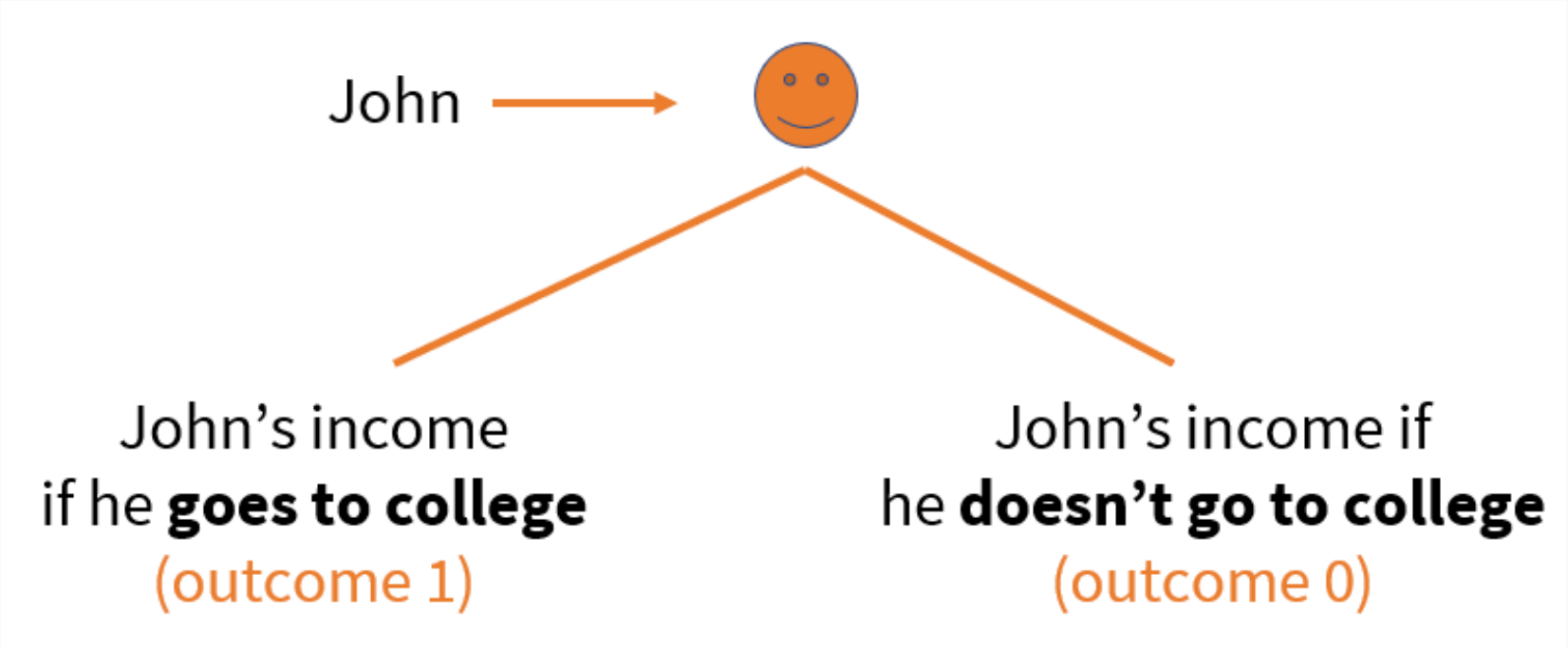
- Does college cause individuals to earn more?

- Why is answering this question hard?
  - Because **people who finish college are different from those who do not finish**

- College graduates might have had higher incomes even if they hadn't finished college!

# Potential outcomes: a framework for causal analysis

- Consider different possible worlds for an individual
- We refer to the outcomes in these different worlds as **potential outcomes**



# Consider a case with two potential outcomes

- Let $Y$ be an outcome variable of interest (e.g., earnings)
  - Actual outcome for a specific individual is $Y_i$, the value recorded in the data

- To talk about potential outcomes, we add a 0/1 subscript
  - $Y_{0i}$ ("y-zero-i"): the outcome for person $i$ with no treatment
  - $Y_{1i}$ ("y-one-i"): the outcome for person $i$ with treatment

# Definition: causal effect

- Causal effect is the difference between two potential outcomes:

$$Y_{1i} - Y_{0i}$$

- The challenge: only one potential outcome is observed per person!
  - The **actual (realized) outcome** is the one we observe
  - The **counterfactual outcome** is the one we do not observe

# How can we measure the causal effect?

- Suppose John attends college and then we measure $Y_{1i}$, his income after graduation
- How can we measure $Y_{0i}$, income in counterfactual where he never attended college?
- Unless we have access to a parallel universe, we cannot

# Solution: use data from multiple individuals

- Measure average income for a group of college graduates
- Measure average income for a group of **otherwise similar** non-college graduates
- We still cannot identify causal effects for any particular individual
- But, income difference between the two groups is the **average causal effect** of college

# Ceteris paribus ("other things equal" or "apples to apples")

- Finding a group that is **otherwise similar** is the key challenge in causal analysis
- This similarity requirement is sometimes referred to as a "ceteris paribus" condition
- Comparisons under ceteris paribus conditions have a causal interpretation
- A major theme of this class is to understand the conditions under which this occurs

# Example: randomized experiment

- Consider a group of 10,000 18-year-olds
- Flip a coin 10,000 times
  - Heads: individual goes to college
  - Tails: individual does not go to college
- **On average**, those who attend college are **otherwise identical** to those who did not
- Four years later, calculate average incomes for these two groups
- This difference is equal to the average causal effect of college

# Causal inference terminology comes from medical trials

- A causal inference analysis includes the following components:
  - **Outcome**: a measure we are interested in studying
  - **Treatment**: a variable measuring the causal relation of interest
  - **Treatment group**: subjects receiving the treatment
  - **Control group**: subjects not receiving the treatment

- Example: what is the effect of college on income?
  - Outcome: income
  - Treatment: attending college
  - Treatment group: individuals who attended college
  - Control group: individuals who did not attend college

- A good control group describes fate of the treated group, if they had not been treated

A good control group shows what would have happened to the treated group, if they had not been treated

# Random assignment

- Our first tool of analysis will be experimental random assignment
  - Straightforward framework for answering causal questions
  - Benchmark for judging results from other analytical methods

- But first, it is helpful to understand why naive analysis is so often flawed

- We will illustrate in the context of the effects of health insurance

# Does health insurance improve health?

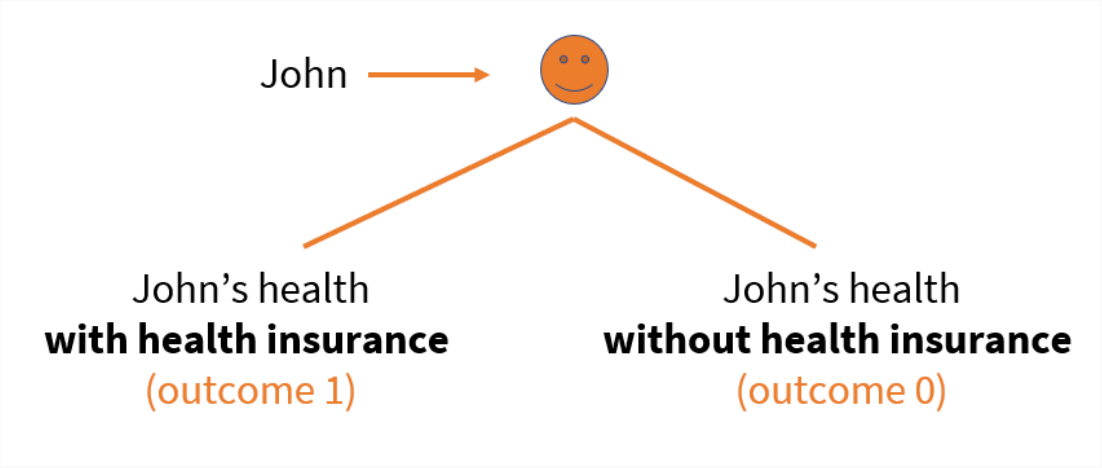
- Background: US vs. other developed countries
  - Americans spend more on health care, yet are less healthy
  - The US has no universal health care coverage (most other countries do)

- The Affordable Care Act increased coverage of health insurance in America

- Is there a causal connection between health insurance and health?
  - Many of the working poor have no health insurance
  - May count on hospital emergency rooms, who cannot turn them away
  - Is the emergency room the best place to receive care for common illnesses?

# Asking the causal question

- Does health insurance make people healthier?
- The ceteris paribus comparison contrasts the health of the same individual with and without insurance
- Challenge: we do not observe both outcomes



# Obtain data from the National Health Interview Survey

- The NHIS surveys Americans about their health and insurance status
- "Would you say your health in general is excellent, very good, good, fair, or poor?"
- Create an index 1-5 to measure health status
  - `5 = excellent`
  - `4 = very good`
  - `3 = good`
  - `2 = fair`
  - `1 = poor`
- Consider men and women separately

# Design of our causal analysis

- Outcome: health index (1-5)
- Treatment: has health insurance (yes/no)
- Treatment group: men (or women) with health insurance
- Control group: men (or women) without insurance

# Men with health insurance (HI) vs men without HI

- What is the estimated treatment effect?
- Is this a good analysis? Are men with no HI a good control group for men with HI?

| | Some Health Insurance | No Health Insurance | Difference |
|---|---|---|---|
| Health Index | 4.0 | 3.7 | **0.3** |
| Age | 44.0 | 41.3 | **2.7** |
| Education | 14.3 | 11.6 | **2.8** |
| Family size | 3.5 | 4.0 | **-0.5** |
| Employed | 0.9 | 0.9 | **0.1** |
| Family income | 106,467 | 45,656 | **60,811** |

# Consider an imaginary table

- Fill in the missing values
  - What are the true treatment effects for John and Dan?
  - What treatment effect would you estimate by comparing actual outcomes?

| | John | Dan |
|---|---|---|
| $Y_{1i}$: Potential outcome, with insurance | 4 | 5 |
| $Y_{0i}$: Potential outcome, no insurance | 3 | 5 |
| $T_i$: Treatment (has insurance) | 1 | 0 |
| $Y_i$: Actual (observed) health outcome | | |
| $Y_{1i} - Y_{0i}$: Treatment effect | | |

# Consider an imaginary table

- Fill in the missing values
  - What are the true treatment effects for John and Dan?
  - What treatment effect would you estimate by comparing actual outcomes?

| | John | Dan |
|---|---|---|
| $Y_{1i}$: Potential outcome, with insurance | 4 | 5 |
| $Y_{0i}$: Potential outcome, no insurance | 3 | 5 |
| $T_i$: Treatment (has insurance) | 1 | 0 |
| $Y_i$: Actual (observed) health outcome | **4** | **5** |
| $Y_{1i} - Y_{0i}$: Treatment effect | **1** | **0** |

# Treatment effects and selection bias

- John has health insurance $(T_i = 1)$ and Dan does not $(T_i = 0)$
- The difference in health is:

$$Y_{John} - Y_{Dan} = -1$$

- This result suggests that health insurance makes you sick!
- In fact, this comparison is flawed because Dan is not a good control group:

$$\begin{aligned} Y_{John} - Y_{Dan} &= Y_{1,John} - Y_{0,Dan} \\ &= \underbrace{Y_{1,John} - Y_{0,John}}_{\text{causal effect } (=1)} + \underbrace{Y_{0,John} - Y_{0,Dan}}_{\text{selection bias } (=-2)} \end{aligned}$$

# Summary

- Correlation is not causation
- Causal analysis requires finding a treatment group and a **good** control group
- Average causal effect is equal to difference in means between treatment and control