Observational Studies

Week 3

Yunkyu Sohn
School of Political Science and Economics
Waseda University

2019 Spring Statistics I

Logistics

- R Tips on https://github.com/ysohn/stats/
- R Exercises (from textbook): https://github.com/kosukeimai/qss

Contents (Book Chapter 2.5 - 2.7)

Descriptive Statistics for a Single Variable (Lab Class)

- Center of Data X
  - Mean ($\overline{X}$): sum(values) / n
  - Median (med(X); robust for outliers than mean) for n observations
    - n is odd: middle value
    - n is even: some of 2 middle values

- e.g. $X = \{1,5,3,7\}$: Mean: Median:

- e.g. $X = \{1,3,5,9,7\}$: Mean: Median:

Descriptive Statistics for a Single Variable (Lab Class)

- Spread of Data X
  - Range: [min(X), max(X)]
  - Quantile: quartile (4), quintiles (5), deciles (10) percentiles (100)
    - 25 percentile = lower quartile (median of X lower than median)
    - 50 percentile = median
    - 75 percentile = upper quartile (median of X higher than median
  - Inter-Quartile Range (IQR): Upper quartile - Lower quartile
  - Standard deviation: $$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
  - e.g. $X = \{1,3,3,3,4,4,4,6,8\}$

Contents (Book Chapter 2.5 - 2.7)

Research Question: Emotional Contagion Hypothesis

- Effect of your friends' FB wall posting on your expressed emotion

Positive post

Great foods!

I love vegetables.

Negative post

I hate vegetables..

I'll not come here again.

Your emotion

- Implication: Large-scale global synchrony/diffusion of emotion

Experimental Version of Facebook Emotion Study

- RCT: 3 mil posts; 155,000 users
- Manipulating FB wall post content exposure probability by sentiment
- 3 page paper with a single figure HOW??
  - Thanks to The POWER of RCT: sole effect of treatment identified

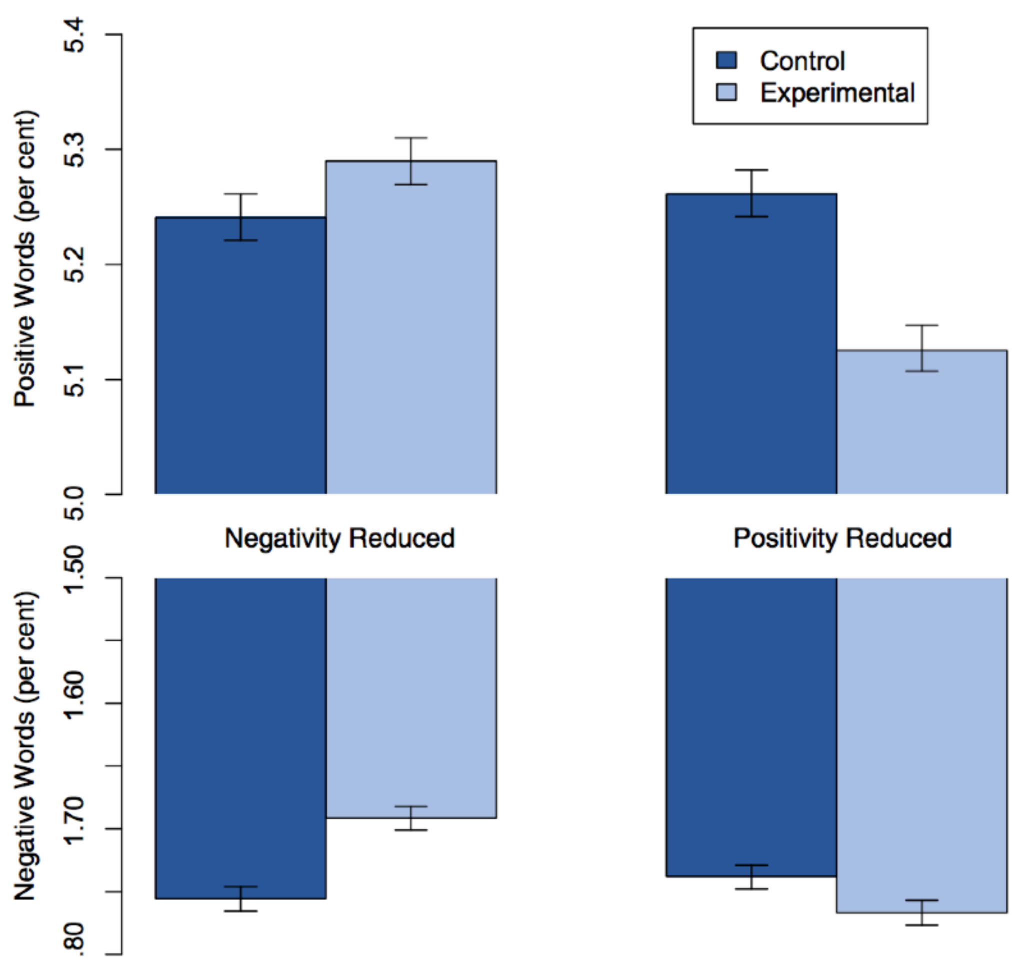
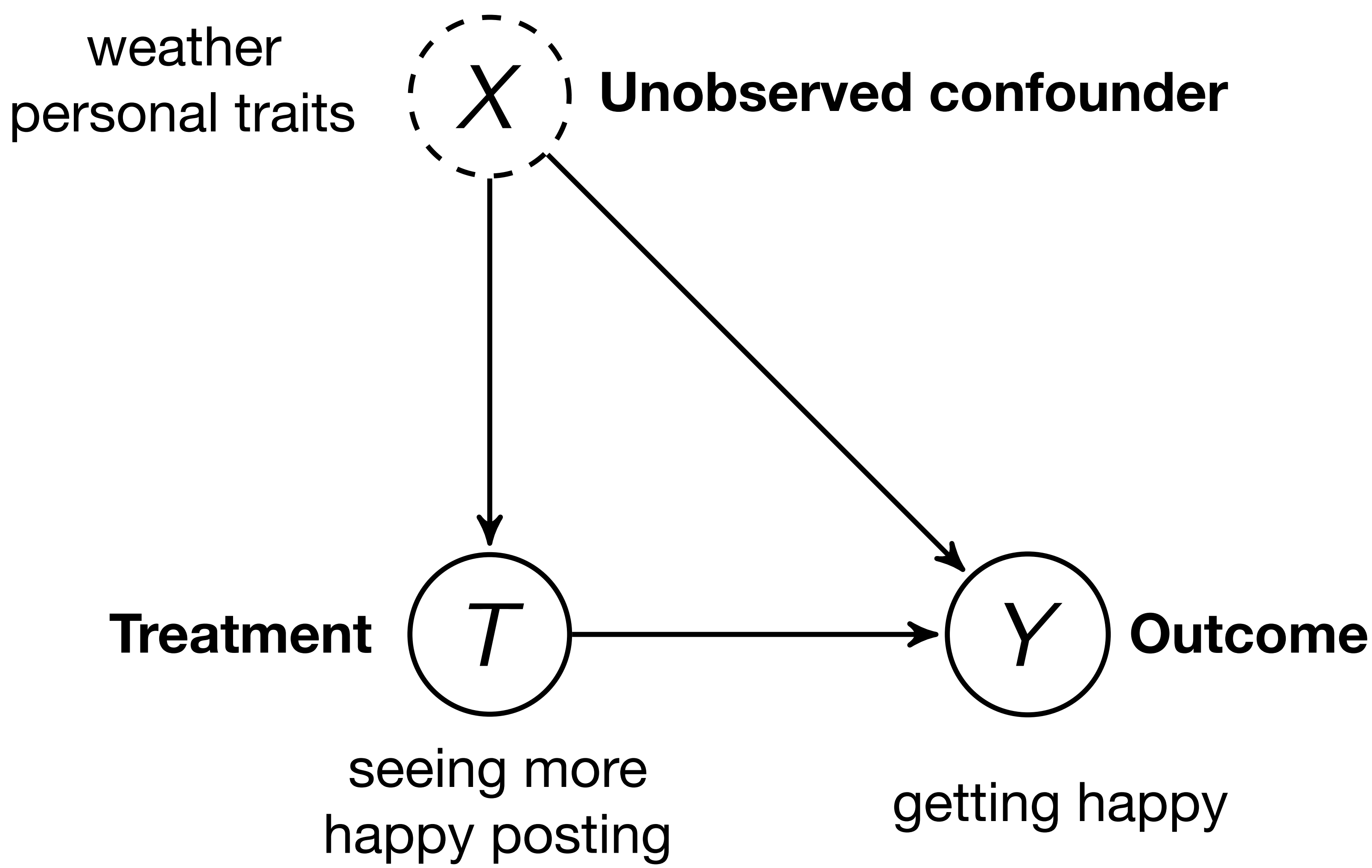


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Kramer, Guillory, and Hancock (2014)

Confounders: Facebook Emotion Study

- Confounders:
  - Pretreatment variables that are associated with both the treatment and outcome variables



Observational Version of Facebook Emotion Study

Detecting Emotional Contagion in Massive Social Networks

Lorenzo Coviello[1], Yunkyu Sohn[2], Adam D. I. Kramer[3], Cameron Marlow[3], Massimo Franceschetti[1], Nicholas A. Christakis[4,5], James H. Fowler[2,6*]

- Observational Studies (things get **super complicated**)
  - Non-experimental study using spontaneous user activities
  - 1,180 days of observation of millions of Facebook users in US
  - Advanced statistical methods to deal with **confounders**
    - **confounders:** both affecting messages you see and your emotion
      - weather: precipitation, temperature, ….
      - user demographic characteristics
      - article length: **44** pages in total.

Review of Casualty and Experimental Studies

- Objective of causal inference
  - Isolating (identifying) the effect of treatment on outcome
- Sample Average Treatment Effect (SATE)
  - Estimating the causal effect of treatment within sample
  - e.g. Impact of social pressure on turnout for n=10

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	1	1	0	1	0	1	1	1	1	0
$Y_i(0)$	0	1	0	0	0	0	1	0	1	1
$Y_i(1)$ - $Y_i(0)$	1	0	0	1	0	1	0	1	0	-1

- **SATE** $= \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}$

Review of Casualty and Experimental Studies

- e.g. Impact of social pressure on turnout for n=10

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	1	1	0	1	0	1	1	1	1	0
$Y_i(0)$	0	1	0	0	0	0	1	0	1	1
$Y_i(1) - Y_i(0)$	1	0	0	1	0	1	0	1	0	-1

- Can we observe both $Y_i(1)$ & $Y_i(0)$ (potential outcomes)?
  - NO!
    - due to Fundamental problem of causal inference
    - =For each i, You only observe ONE among $Y_i(1)$ & $Y_i(0)$
    - What would a real dataset look like? ➠

Review of Casualty and Experimental Studies

- e.g. Impact of social pressure on turnout for n=10

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	X	X	0	X	0	X	X	1	1	0
$Y_i(0)$	0	1	X	0	X	0	1	X	X	X
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

- What would a real dataset look like?
  - = For each i, You only observe ONE among $Y_i(1)$ & $Y_i(0)$
    - Can you calculate SATE?
      - NO! ➠ We should find a way to approximate SATE.
        - What would be a feasible alternative?

Review of Casualty and Experimental Studies

- e.g. Impact of social pressure on turnout for n=10

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	X	X	0	X	0	X	X	1	1	0
$Y_i(0)$	0	1	X	0	X	0	1	X	X	X
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

- (If we can choose to assign treatment & control groups)
- The best possible design: Randomized Control Trials (RCTs)
  - Assign treatment status completely at random
    - Why does this guarantee the best possible estimation?
      - ➠ No sample selection bias

Review of Casualty and Experimental Studies

- e.g. Impact of social pressure on turnout for n=10

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	1	1	0	1	0	1	1	1	1	0
$Y_i(0)$	0	1	0	0	0	0	1	0	1	1
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
…										
$Y_i(1) - Y_i(0)$	1	0	0	1	0	1	0	1	0	-1

- So many **confounding variables** that we do not observe
  - Bias in treatment assignment ➠ Invalid inference

Review of Casualty and Experimental Studies

- Biased assignment scenario 1:

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	X	1	0	X	X	1	X	1	1	X
$Y_i(0)$	0	X	X	0	0	X	1	X	X	1
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
...										
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

- So many confounding variables that we do not observe
  - Bias in treatment assignment ➠ Invalid inference

Review of Casualty and Experimental Studies

- Biased assignment scenario 2:

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	1	X	X	X	X	X	X	1	1	0
$Y_i(0)$	X	1	0	0	0	0	1	X	X	X
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
...										
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

- So many confounding variables that we do not observe
  - Bias in treatment assignment ➠ Invalid inference

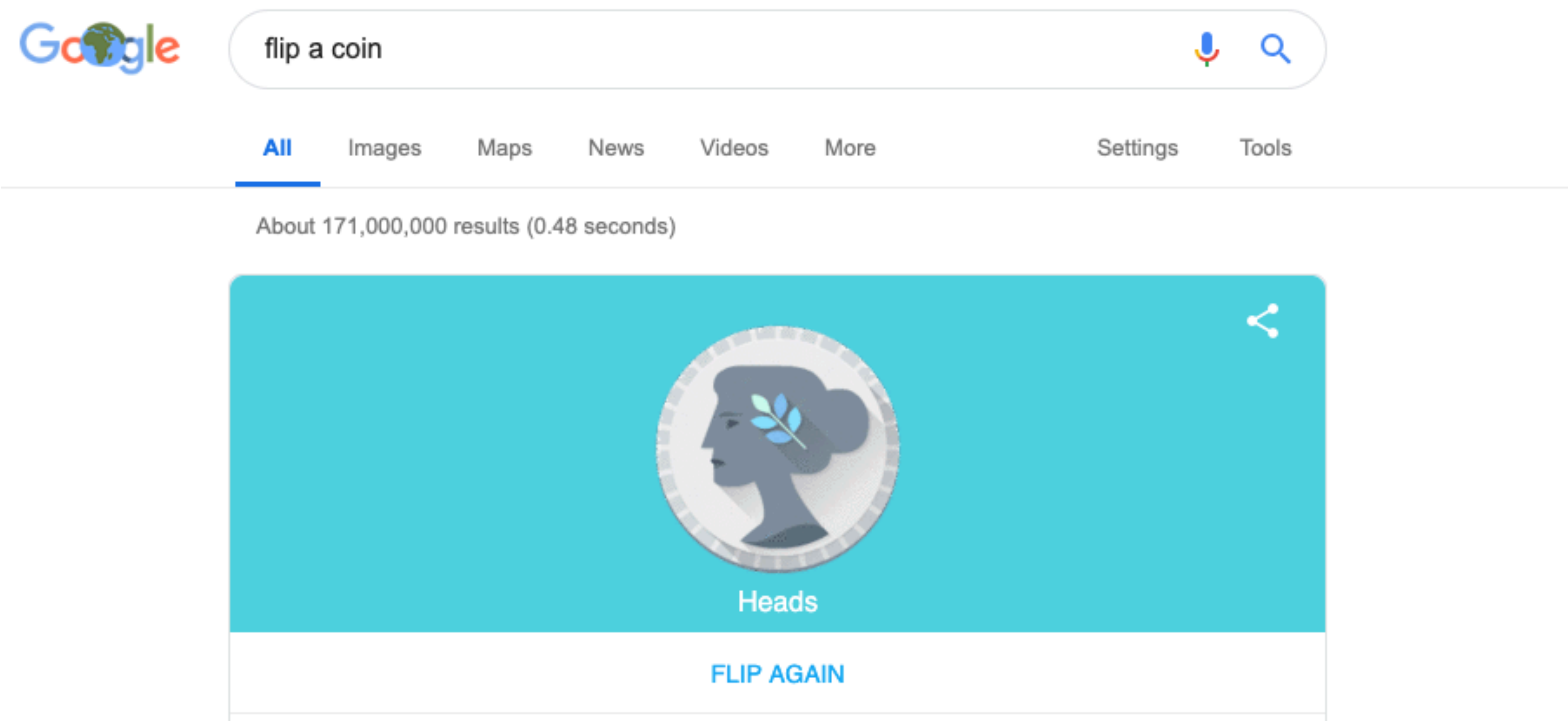
Review of Casualty and Experimental Studies

- What would be the best possible assignment??

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	1	1	0	1	0	1	1	1	1	0
$Y_i(0)$	0	1	0	0	0	0	1	0	1	1
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
...										
$Y_i(1)$ - $Y_i(0)$	1	0	0	1	0	1	0	1	0	-1

Review of Casualty and Experimental Studies

- What would be the best possible assignment??
  - Assign treatment status completely at random



  - or **sample()** or **rbinom()** in R: link

Review of Casualty and Experimental Studies

- Flipped coin 10 times: Head (1) -> Treated; Tail (0) -> Control
  - e.g. Say you got 0 1 0 0 1 0 0 1 1 1

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	X	1	X	X	0	X	X	1	1	0
$Y_i(0)$	0	X	0	0	X	0	1	X	X	X
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

- What does this **guarantee**? ➠ Now we can trust using
- Difference in the sample means estimator (size of treated: $|\{T_i=1\}|$)

$$D = \frac{1}{|\{T_i = 1\}|} \sum_{i \in \{T_i=1\}} Y_i - \frac{1}{|\{T_i = 0\}|} \sum_{i \in \{T_i=0\}} Y_i$$

Observational Studies: Getting Extremely Complicated

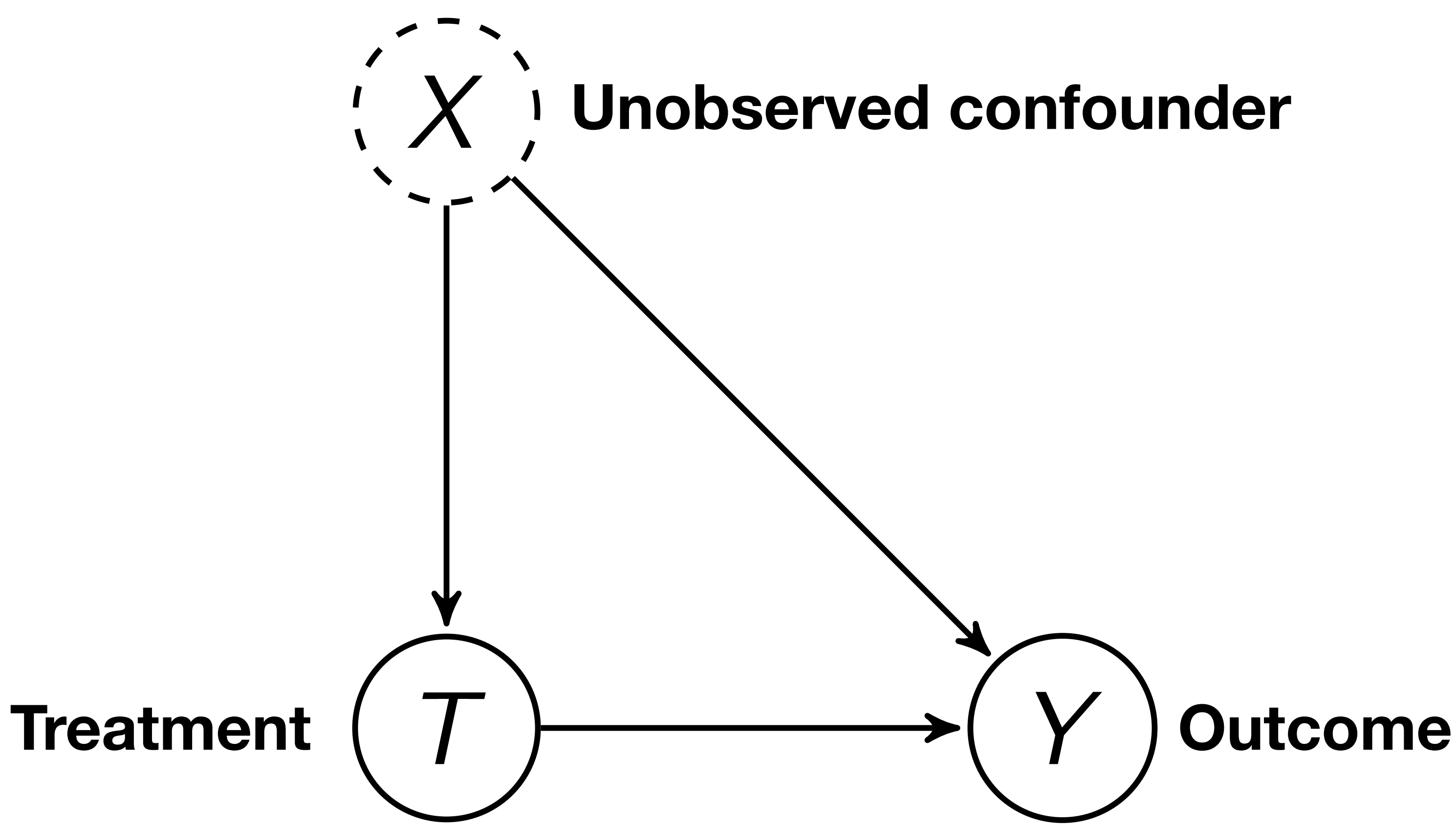
- Challenge: Can we randomly assign as the previous example?

unit	1	2	3	4	5	6	7	8	9	10
$Y_i(1)$	X	1	X	X	0	X	X	1	1	0
$Y_i(0)$	0	X	0	0	X	0	1	X	X	X
Education	C	H	E	E	H	H	H	C	C	C
Race	W	W	B	B	A	W	W	B	W	A
Gender	F	M	M	F	F	M	F	M	M	F
$Y_i(1)$ - $Y_i(0)$	?	?	?	?	?	?	?	?	?	?

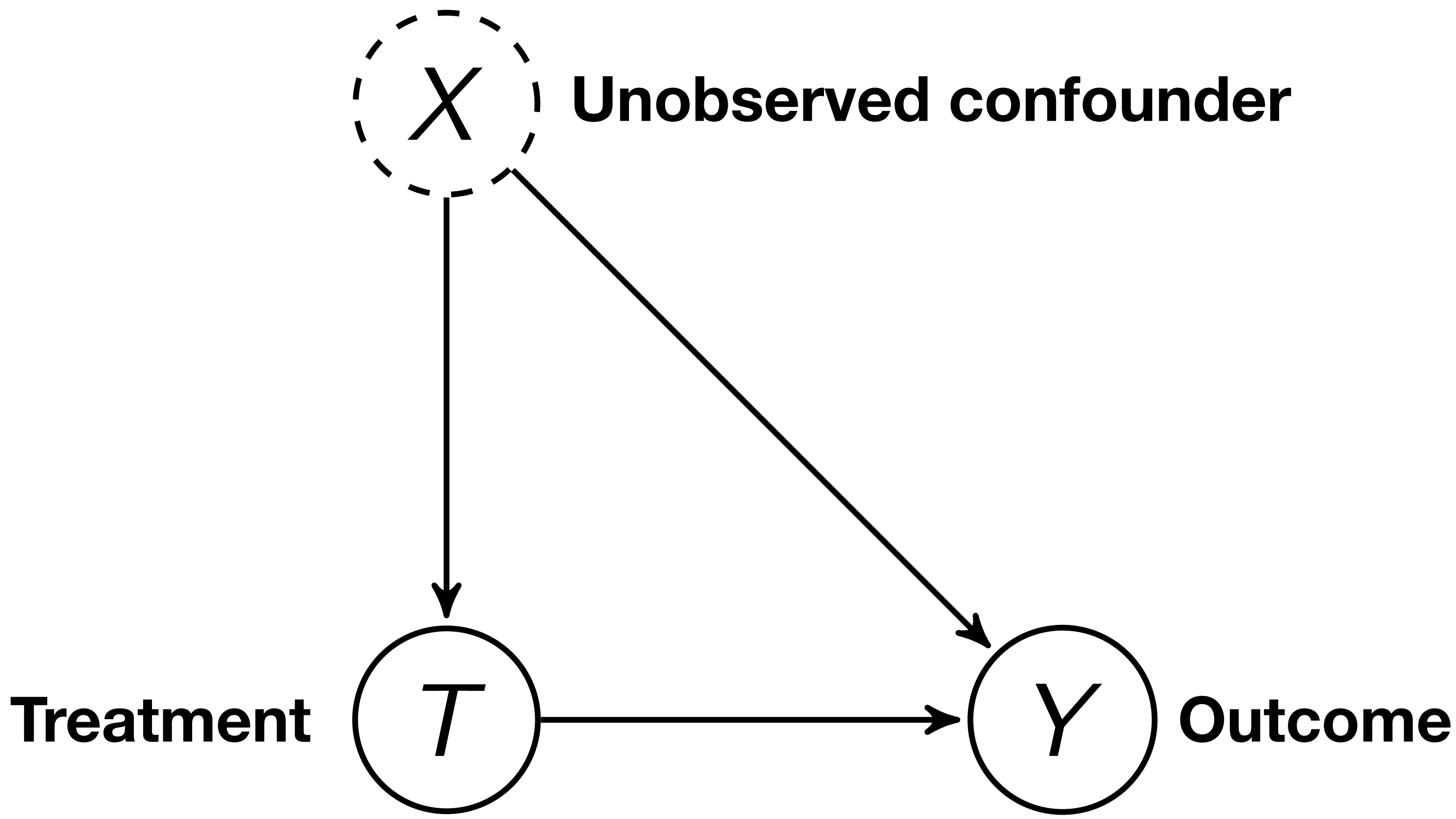
- Real life: Not conducting experiment but mostly observations
  - Some will get pressures and some will not get pressures
    - ➠ No guarantee in balanced **pre-treatment variables**

Observational Studies: Sources of Bias

- Experiment: Merit of conducting RCTs between 2 groups
  - ➠ No difference on average except treatment status
- Observations ➠ Unbalanced pre-treatment variables
  - These variables affect both treatment status & outcome



Observational Studies: Sources of Bias



- Examples: why selection bias matters
  - 1) X: demographic traits; T: happy neighbors; Y: emotion
  - 2) X: colonial history; T: democratic; Y: wealth
    - Selection bias in real life (observational studies)
- What does RCTs guarantee? ➟ Unconfoundedness
- Observations: Confounding bias ➟ needs Statistical Control

Observational Study Designs and Statistical Control

- Learn several forms of observational study designs through example
- Important question in labor economics:
  - How does increase in minimum wage affect fulltime employment?
  - Theory: "raising minimum wage will encourage employers to replace full time employees with part-timers to recoup the increased cost in wages."
  - Center of debate in multiple countries
  - Extremely difficult to conduct experiments: Why?
- Our (longitudinal/panel) data set for a case study
  - 1992: New Jersey minimum wage increased from \$4.25 to \$5.05
  - PA located right next to NJ remained at \$4.25
  - PA and NJ are similar
  - wage/#employees of ff chains in PA and NJ before/after 1992

Observational Study Designs and Statistical Control

- Complete data (please check https://github.com/kosukeimai/qss)

Name	Description
`chain`	name of fastfood restaurant chain
`location`	location of restaurants (`centralNJ`, `northNJ`, `PA`, `shoreNJ`, `southNJ`)
`wageBefore`	wage before the minimum wage increase
`wageAfter`	wage after the minimum wage increase
`fullBefore`	number of fulltime employees before the minimum wage increase
`fullAfter`	number of fulltime employees before the minimum wage increase
`partBefore`	number of parttime employees before the minimum wage increase
`partAfter`	number of parttime employees before the minimum wage increase

Cross-Sectional Comparison

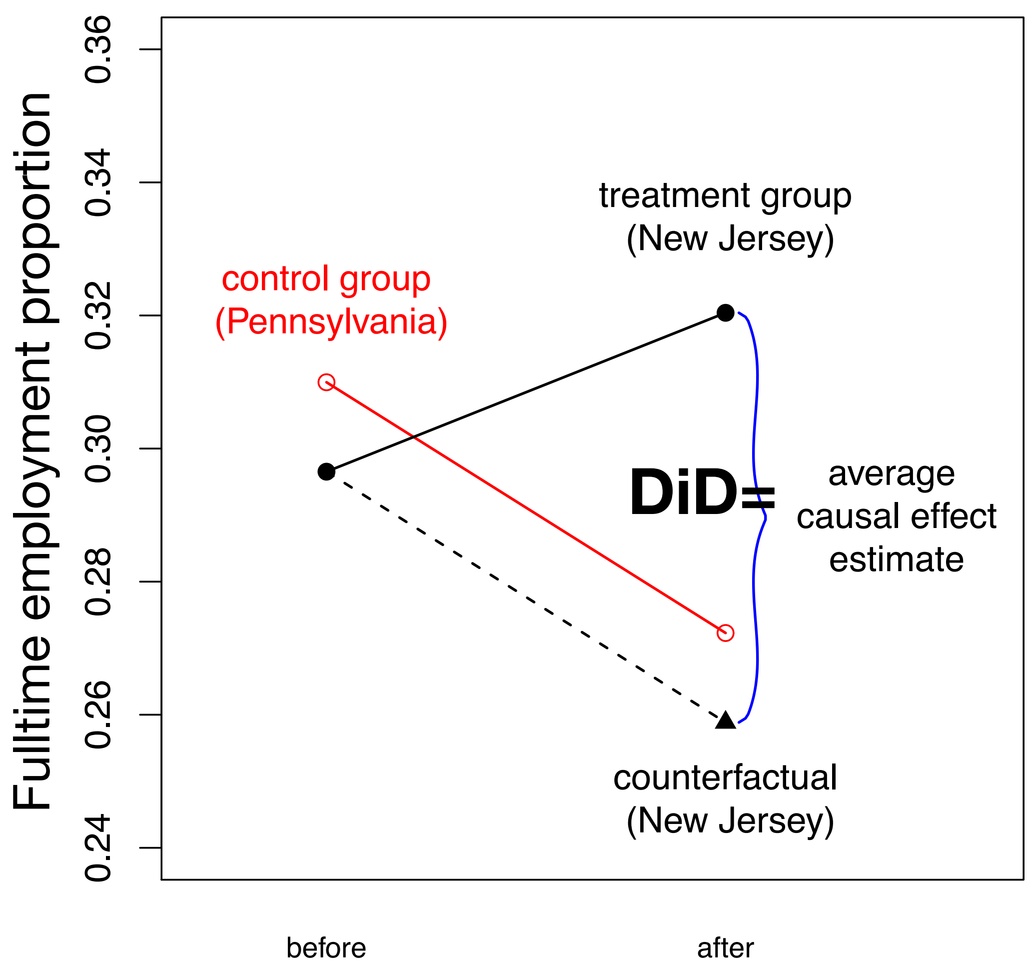
- Calculate difference in sample means (approximation of SATE)
  - Assumption: NJ and PA are very similar except the treatment
    - ➠ We can use PA as a control
  - Estimate SATE using difference in means estimator
    - $$D = \frac{1}{|\{T_i = 1\}|} \sum_{i \in \{T_i = 1\}} Y_i - \frac{1}{|\{T_i = 0\}|} \sum_{i \in \{T_i = 0\}} Y_i$$
    - Y_i = proportion of fulltime employment for chain i
    - Treated units: NJ employee income after the reform
    - Control units: PA employee income after the reform

Before-and-After Design

- Before-and after design
  - In case X (confounder) is very different between NJ and PA
  - Assumption: time-constant confounder ➠ NJ before/after?
  - Compare only NJ before and after the treatment
  - Difference in means estimator
    - $D = \frac{1}{|\{T_i = 1\}|} \sum_{i \in \{T_i=1\}} Y_i - \frac{1}{|\{T_i = 0\}|} \sum_{i \in \{T_i=0\}} Y_i$
    - Y_i = proportion of fulltime employment for chain i
    - Treated units: NJ employee income before the reform
    - Control units: NJ employee income after the reform

Difference-in-Differences Design

- Difference-in-Differences design:
  - Controlling for Time-varying confounders (e.g. US economy)
    - with the parallel time trend assumption
  - Sample Average Treatment Effect for the Treated (SATT)
    - Difference-in-Differences (DiD) estimate using counterfactual Y =



$$\overline{Y}^{\textbf{after}}_{\textbf{treated}} - \{\overline{Y}^{\textbf{before}}_{\textbf{treated}} - (\overline{Y}^{\textbf{before}}_{\textbf{control}} - \overline{Y}^{\textbf{after}}_{\textbf{control}})\}$$

$$= \underbrace{\left(\overline{Y}^{\text{after}}_{\text{treated}} - \overline{Y}^{\text{before}}_{\text{treated}}\right)}_{\text{difference for the treatment group}} - \underbrace{\left(\overline{Y}^{\text{after}}_{\text{control}} - \overline{Y}^{\text{before}}_{\text{control}}\right)}_{\text{difference for the control group}}$$

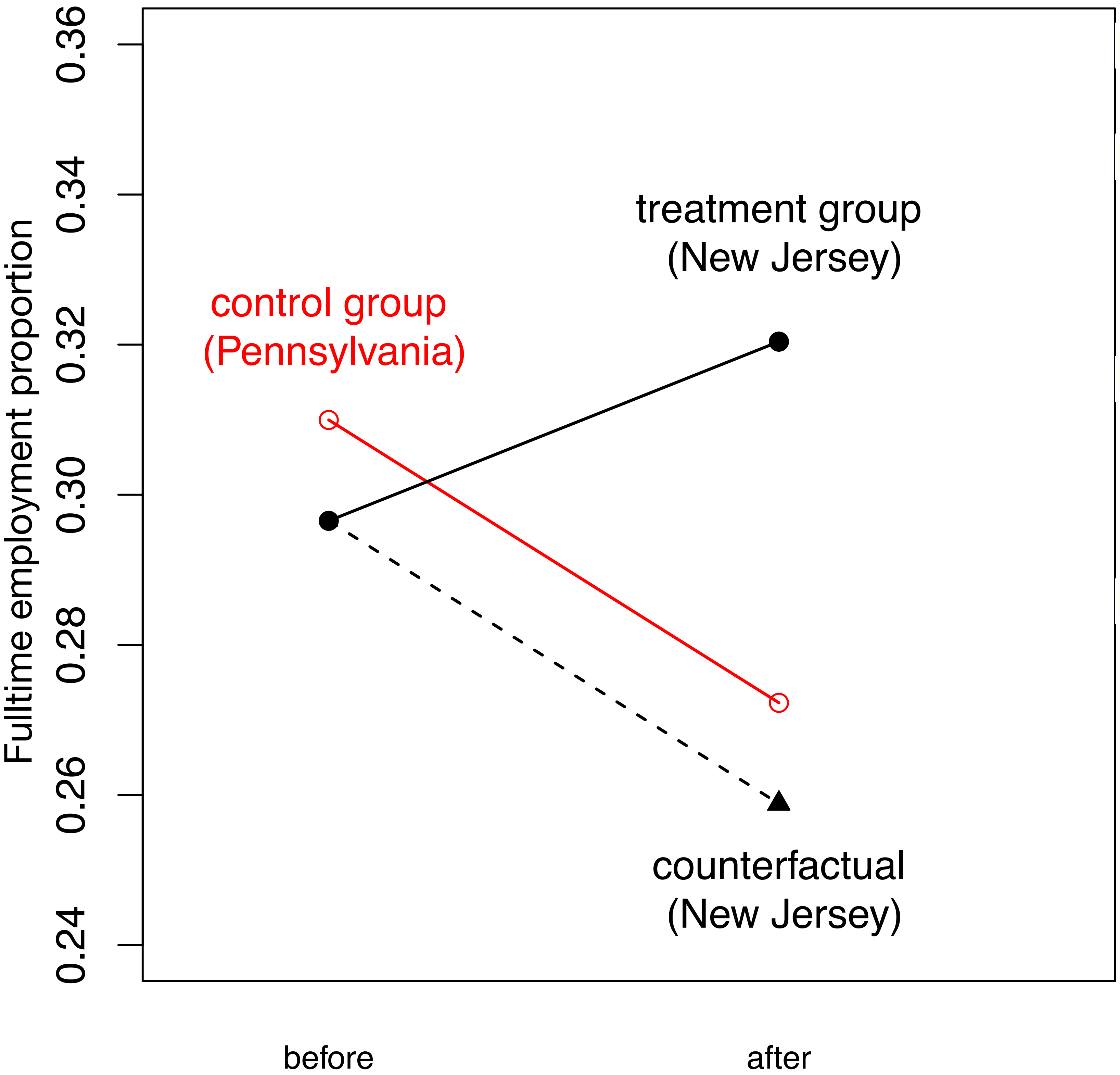
Parallel time trend assumption for

Time-varying confounders:

What would have happened if NJ was not

treated?: Following the same path of PA

The Three Identification Strategies



- Draw lines on the graph
- **Cross-sectional design**
  - Difference in Means
- **Before-and-after design**
  - Difference in Means
- **Difference in Differences**
  - Diference in Differences

The Three Identification Strategies

- **Cross-sectional design**

```r
mean(minwageNJ$fullPropAfter) -
  mean(minwagePA$fullPropAfter)
```

```
## [1] 0.0481
```

- **Before-and-after design**

```r
NJdiff <- mean(minwageNJ$fullPropAfter) -
    mean(minwageNJ$fullPropBefore)
NJdiff
```

```
## [1] 0.0239
```

- **Difference in Differences**

```r
PAdiff <- mean(minwagePA$fullPropAfter) -
  mean(minwagePA$fullPropBefore)
NJdiff - PAdiff
```

```
## [1] 0.0616
```

Summary

- Descriptive statistics for a single variable
- Review of casualty and experimental studies
- Observational studies
  - Confounding bias
  - Cross-section design
  - Before-and-after design
  - Difference-in-differences design

Next Next Week

- Next week: break!
  - Hope you have a great time.
- 2 weeks later
  - Measurement and survey sampling
  - Base Graphics in R

See you 2 weeks later.