

Political Science 209 - Fall 2018

Observational Studies

Florian Hollenbach

24th September 2018

What is the fundamental problem of causal inference?

What about randomized control trials allows us to credibly estimate a causal effect?

What can induce citizens to vote?

What was the experiment?

What was the experiment?

Letters to randomized households with treatment:

1. Naming and Shaming: your neighbors will know
2. Civic Duty
3. Hawthorne Effect Message
4. Control (no letter)

Let's go to R-studio quick

What is the main problem for observational studies?

What is the main problem for observational studies?

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

What is the Problem with Confounders?

What is the Problem with Confounders?

- If pre-treatment characteristics are associated with treatment and outcome, we can't disentangle causal effect from confounding bias

What is the Problem with Confounders?

- If pre-treatment characteristics are associated with treatment and outcome, we can't disentangle causal effect from confounding bias
- Selection into treatment example: Maybe minimum wage was increased because unemployment was particularly low in NJ, but not PA

Examples of Confounding

- Are incumbents more likely to win elections? Yes, but...

Examples of Confounding

- Are incumbents more likely to win elections? Yes, but...
- Incumbents receive more campaign contributions
- Incumbents have more staff

Examples of Confounding

- Does higher income lead countries to democratize?

Examples of Confounding

- Does higher income lead countries to democratize?
- Higher income countries have more educated populations

What can we do about confounding in observational studies?

What can we do about confounding in observational studies?

- Make *Treatment* and *Control* groups as similar to each other as possible
- Especially on variables that might matter for treatment status and outcome
- Analyze subsets or *statistical control*, such that we compare treated and control units that have same value on confounder

Another problem with observational studies:

- Reverse causality

Another problem with observational studies:

- Reverse causality
- Example: Does economic growth cause democratization or democratization cause growth?

Why do experiments not suffer from the threat of reverse causality?

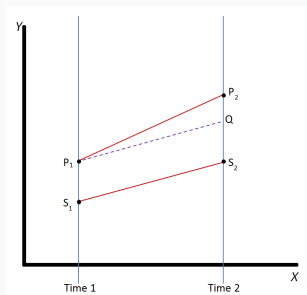
Difference-in-Differences Design

Difference-in-Differences Design

- Compare trends before and after the treatment across the same units
- Takes initial conditions into account

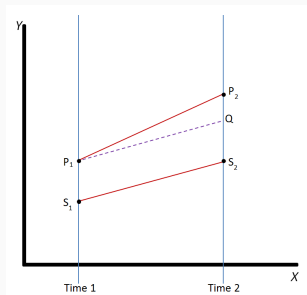
Difference-in-Differences Design

- Need data measured for both treatment and control at two different time periods: before and after treatment



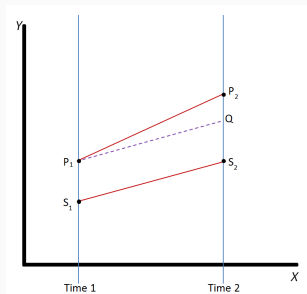
- Total difference between P2 and S2 can not be attributed to treatment. Why?

Difference-in-Differences Design



What might be a necessary condition for Diff-in-Diff to work?

Difference-in-Differences Design



What might be a necessary condition for Diff-in-Diff to work?

Parallel Trends Assumptions

Difference-in-Differences Design

The **difference-in-differences** (DiD) design uses the following estimate of the average treatment effect for the treated (ATT),

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

The assumption is that the counterfactual outcome for the treatment group has a time trend parallel to that of the control group.

Describing numeric variables:

- Mean
- Median
- Quantiles

- splitting observations into equally size groups, e.g., quartiles, quantiles
- 75th percentile is the threshold under which 75% of observations lie
- What percentile is the median?

Describing the spread of numeric variables:

- IQR:

Describing the spread of numeric variables:

- IQR:

Difference between 75th percentile and 25th percentile

Describing the spread of numeric variables:

Standard Deviation

Describing the spread of numeric variables:

Standard Deviation

$$SD = \sqrt{\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Standard Deviation

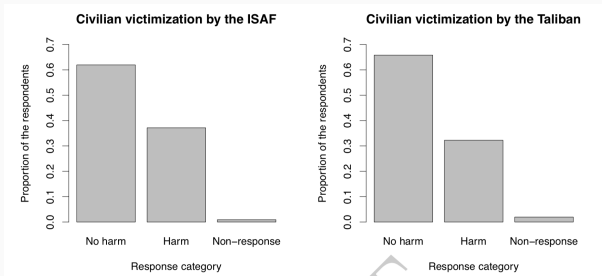
The sample **standard deviation** measures the average deviation from the mean and is defined as,

$$\text{standard deviation} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

where \bar{x} represents the sample mean, i.e., $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and n is the sample size. Few data points lie outside of 2 or 3 standard deviations away from the mean. The square of standard deviation is called **variance**.

Describing single Variables

- Barplots can be used to summarize factor(?) variables
- Proportion of observations in each category as the height of each bar



Histograms

- Histograms look similar to barplots
- Used for numeric variables
- Numeric variables are *binned* into groups

Histograms

- Each bar is for one bin
- Height of each bar is the *density* of the bin

Histograms

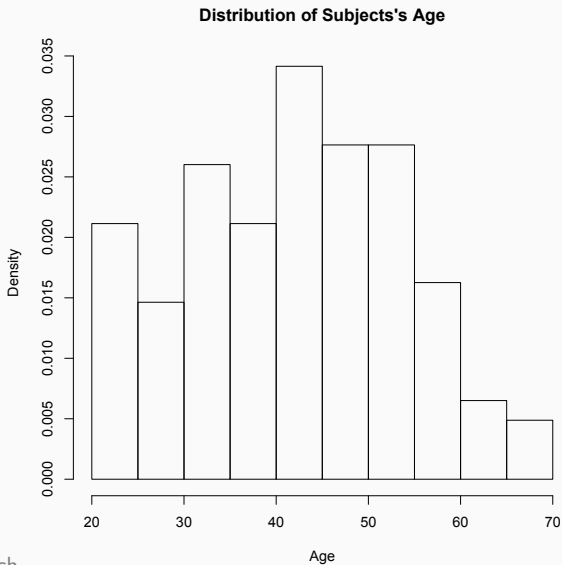
- Each bar is for one bin
- Height of each bar is the *density* of the bin
- Important: Height is share of observations in bin divided by bin size

Histograms

- Each bar is for one bin
- Height of each bar is the *density* of the bin
- Important: Height is share of observations in bin divided by bin size
- Unit of vertical axis (y-axis) is interpreted as percentage per horizontal (x-axis) unit

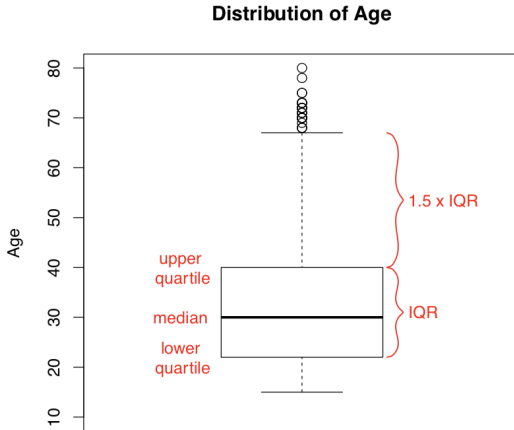
- Area of each bar is the share of observations that fall into that bin
- Area of all bins sum to one

Histograms

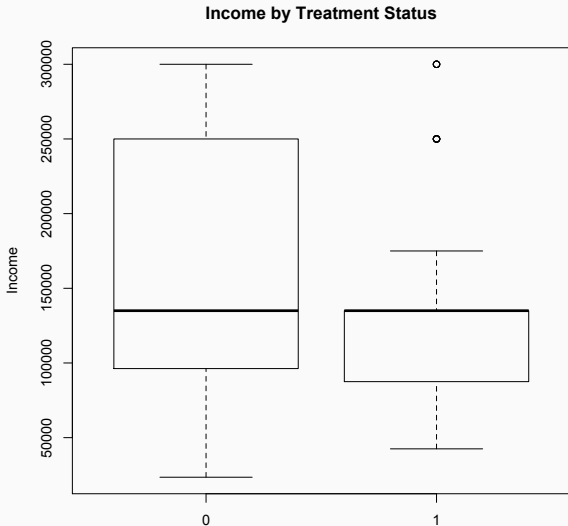


- Boxplots also display the distribution of a numeric variable
- Boxplots show the *median*, *quartiles*, and *IQR*

Boxplots



Boxplots can show how two variables covary



- A sample is a small share of the population in that we are interested in

Survey Sampling

- A sample is a small share of the population in that we are interested in
- How do we draw samples in such a way that polls accurately reflect what is going to happen?
- How to construct samples that will represent the population?

Survey Sampling

- Example: We want to know the voting intentions of Texans (or Americans)
- We can hardly ask all eligible voters about their intention

Survey Sampling

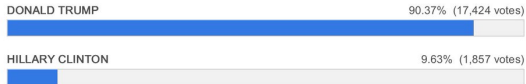
- Example: We want to know the voting intentions of Texans (or Americans)
- We can hardly ask all eligible voters about their intention
- We take a *sample*

Survey Sampling

- The size of the sample is less important than its composition



****DRUDGE POLL** WHO WON THE FIRST PRESIDENTIAL DEBATE?**



Total Votes: 19,281

[Suka](#) [Kongsi](#) [607](#)

[Tweet](#)

Literary Digest Sample

- Mail questionnaire to 10 million people
- Addresses came from phone books and club memberships
- Problems?

Literary Digest Sample

- Mail questionnaire to 10 million people
- Addresses came from phone books and club memberships
- Problems?

- Biased *sample*

Quota Sampling

- Sample certain groups until quota is filled
- Does not mean unobservables are representative

Simple Random Sampling

- Think of all voters sitting in a box, survey firm randomly draws voters
- Random draws without replacement give us an unbiased estimate of the population
- Everybody has the same chance of being in the sample

Simple Random Sampling

- Pre-determined number of units are randomly selected from population
- Sample will be representative of population on observed and unobserved characteristics

Simple Random Sampling

- Not every single sample will be exactly representative
- If we were to take a lot of random samples (say 1000 samples of 1000 respondents), on average the samples would be representative

Simple Random Sampling

- Each single sample can be off and different
- Polls are associated with uncertainty

Ted Cruz leads Beto O'Rourke 54 to 45, new poll says

The new Quinnipiac University poll surveyed likely voters instead of registered voters like it did in past iterations.

BY [PATRICK SVITEK](#) SEPT. 18, 2018 11 AM

Simple Random Sampling

- Each single sample can be off and different
- Polls are associated with uncertainty

Ted Cruz leads Beto O'Rourke 54 to 45, new poll says

The new Quinnipiac University poll surveyed likely voters instead of registered voters like it did in past iterations.

BY [PATRICK SVITEK](#) SEPT. 18, 2018 11 AM

Beto O'Rourke leads Ted Cruz by 2 among likely voters in U.S. Senate race, new poll finds

O'Rourke has been closing the gap over the last several months, but this is the first poll that puts him ahead of Cruz.




BY [KATHRYN LUNDSTROM](#) SEPT. 19, 2018 8 AM

Random Sampling is hard

- How to create sampling frame?
- Random digit dialing? Walking to random houses?
- Multi-stage cluster sampling

Non-response bias

- Unit non-response bias:

 **Nate Cohn** 
@Nate_Cohn [Follow](#) 

We've called thousands of cell phones in VA07 today and we don't have a single new 18-29 year old respondent. Meanwhile, we've got a disproportionate 8 in Colo 6 already, all supporting the Dem

		DEM.	REP.	UND.
Age	18 to 29 n = 8 / 14% of voters	84%	—	16%
	30 to 44 13 / 26%	74%	22%	4%
	45 to 64 16 / 36%	24%	64%	12%
	65 and older 11 / 24%	36%	64%	—

8:13 PM - 12 Sep 2018

- Item non-response bias: *What was the last crime you committed?*
- Sensitive questions: non-response, social desirability bias
Turnout, racial prejudice, corruption

Why could this be a problem in the Afghanistan example?

