Political Science 209 - Fall 2018

Observational Studies

Florian Hollenbach

18th September 2018

Review

What is the fundamental problem of causal inference?

Review

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

Get out the Vote Study

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 treament 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?

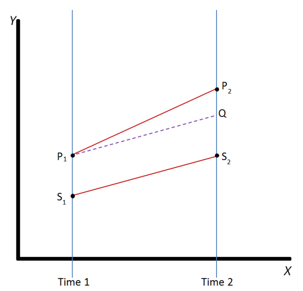
Difference-in-Differences Design

Difference-in-Differences Design

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

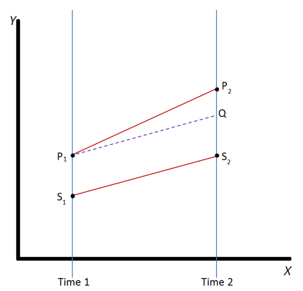
Difference-in-Differences Design

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



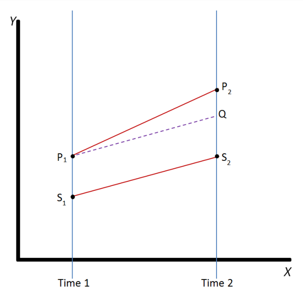
- Total difference between P2 and S2 can not be attributed to treatment. Why?

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?

Parralel 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),

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

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

Quantiles

- splitting observations into equaly 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

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

Standard Deviation

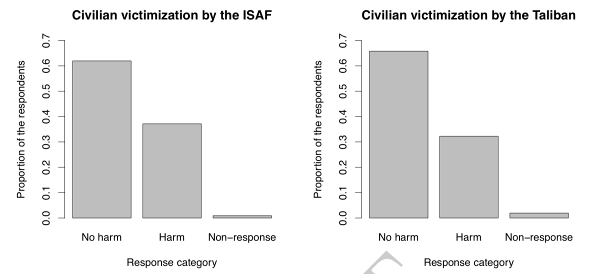
The sample **standard deviation** measures the average deviation from the mean and is defined as,

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

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**.

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

Barplots



Civilian victimization by the ISAF
Proportion of the respondents
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
No harm
Harm
Non−response
Response category
Civilian victimization by the Taliban
Proportion of the respondents
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
No harm
Harm
Non−response
Response category

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

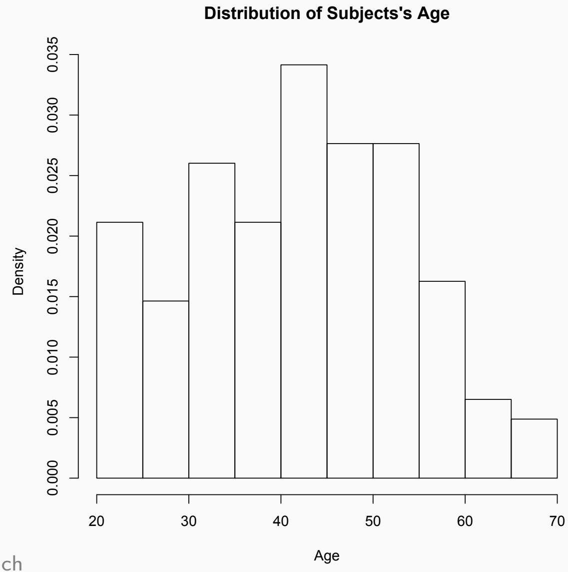
- Important: Height is share of observations in bin divided by bin size

- Unit of vertical axis (y-axis) is interpreted as percentage per horizontal (x-axis) unit

Histograms

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

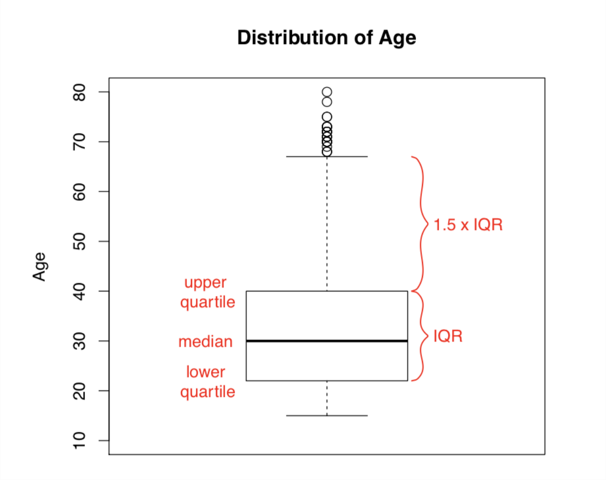
Histograms


Distribution of Subjects's Age
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
20
30
40
50
60
70
Age

Boxplots

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

Boxplots


Distribution of Age
Age
10
20
30
40
50
60
70
80
upper
quartile
median
lower
quartile
1.5 x IQR
IQR

Boxplots can show how two variables covary