

# Lecture 12: Causation and Natural Experiments

## Modeling Social Data, Spring 2017

### Columbia University

Cecilia Simas

April 21, 2017

## 1 Random Experiments

### 1.1 Confounds

Reviewing the example from last lecture, if we want to assess the effect of going to hospitals on health, we need to account for confounds. The cause and effect of the outcome may be affected (confounded) by a common cause, and this impacts the result. As seen in Figure 1, a person's health today has an effect on both on whether they go to hospital and their health tomorrow,

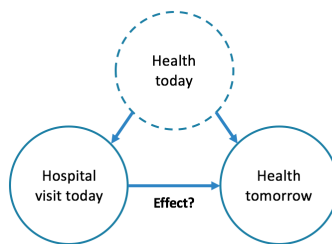


Figure 1: *Source: Lecture 11*

### 1.2 Random Assignment

**Random Assignment** is the process of assigning human participants into different groups using randomization.

Using random assignment breaks the link between confounds and the results. By assigning different treatments randomly, we can account for the confounds.

For example, if the people who would go to the hospital were chosen by a coin flip, we would be able to isolate the effect of going to the hospital without the confounds, as shown in Figure 2.

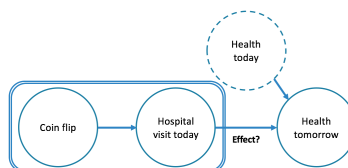


Figure 2: *Source: Lecture 11*

### 1.2.1 Problems with Random Assignment

While it is considered a "gold standard" of causal inference, it may be misleading. There are 4 main issues with random assignment: small sample sizes, researcher degrees of freedom, publication bias, and p-hacking.

**Small sample sizes** make the sample more subject to statistical variation, leading to the potential of observing several flukes.  $\alpha$ , the significance level, and Power ( $1 - \beta$ ), the chance of detecting a real effect if one exists, also have an effect on the experiment. To choose these values for an experiment, it is often useful to run a **pilot study**, a smaller, preliminary version of the experiment.

**Researcher degrees of Freedom** may affect the results, as the flexibility in analysis is greater than the flexibility in the data. A researcher may have decided to run a different test on different data, for example, and they may also decide when to stop collecting samples such that the data fits a desired result.

**Publication bias**

**p-hacking**

### 1.2.2 Other Limitations

- Randomization often is possible or ethical
- Experiments are expensive (in terms of both time and money)
- It may be difficult to create feasible parallel worlds
- People may be non-compliers: they may not follow the rules of the test

### 1.2.3 Measuring Compliance Rates

To measure compliance rates, measure the fraction of people that accept treatment in the treatment group, and the fraction that gets treated anyway in the control group.

## 1.3 Counterfactuals

**Counterfactuals** are events or outcomes that did not happen, "what if?" questions. However, to isolate causal effect, we have to change one and only one thing and compare outcomes, so we are not able to see the counterfactual.

## 2 Natural Experiments

Sometimes, experiments happen naturally, without needing to be formally designed. There are four main kinds of Natural Experiment: as-if random, instrumental variables, discontinuities, and difference in differences.

### 2.1 As-if Random

**Nature randomly assigns conditions, and people comply** An example of this is the Cholera outbreak in London in 1854, in which people were exposed to different water sources.

### 2.2 Instrumental Variables

**An instrument independently shifts the distribution of the treatment** For example, a lottery determines the military draft. It functions like a random assignment, and helps obtain results.

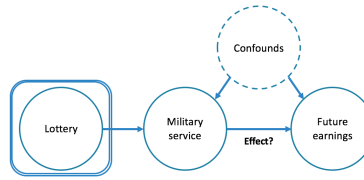


Figure 3: *Source: Lecture 11*

## 2.3 Regression Discontinuities

**Things change around an arbitrarily chosen threshold** Yelp reviews, for example, round the stars arbitrarily, resulting in the discontinuity seen in Figure 4.

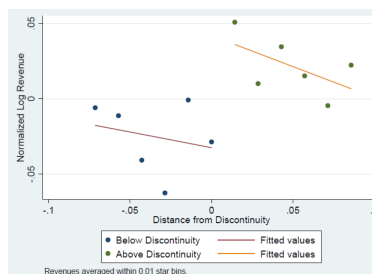


Figure 4: Yelp star ratings

## 2.4 Difference in Differences

**Compare differences after a sudden change with trends in a control group** If minimum wage changes in one state, for example, the effect may be visible in the difference in the trend line, as shown in Figure 5.

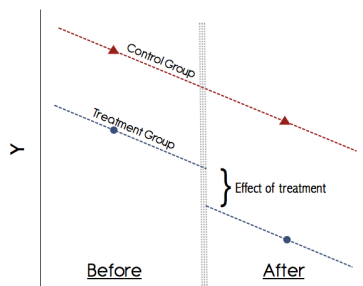


Figure 5: Difference in Treatments can be seen in Trendlines