YData: An Introduction to Data Science

Lecture 20: Causality

Jessi Cisewski-Kehe
Statistics & Data Science, Yale University
Spring 2020

Credit: data8.org

Announcements

A/B Testing

- **Null**: The two samples are drawn randomly from the same underlying distribution.

- If the null is true, all rearrangements of the variable values among the two samples are equally likely. So:
  - compute the observed test statistic
  - then shuffle the values and recompute the statistic; **repeat**; compare with the observed statistic

Deflategate

2015 AFC Championship Game

Deflategate

Wikipedia:
The 2015 AFC Championship Game football tampering scandal, commonly referred to as Deflategate, or Ballghazi

...

Null hypothesis

The 4 Colts footballs are like a sample drawn at random without replacement from all 15 balls.

- To test this hypothesis, repeat this process:
  - Randomly permute all 15 balls
  - Label 11 of them "Patriots" and the remaining 4 "Colts"
  - Compare the averages of the two groups

(DEMO)

Causality

Randomized Controlled Experiment

- Sample A: **control group**
- Sample B: **treatment group**

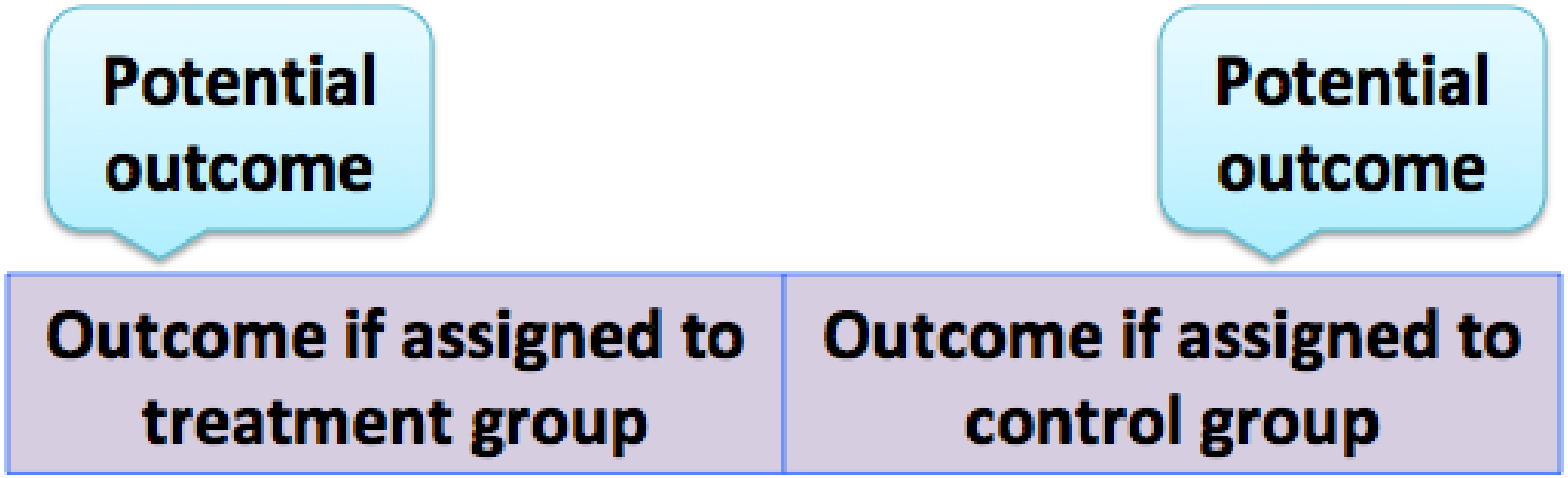
- **If the treatment and control groups are selected at random, then you can make causal conclusions.**

- Any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

(DEMO)

Before the Randomization

- In the population there is one imaginary ticket for each of the 31 participants in the experiment.
- Each participant's ticket looks like this:



The Data

16 randomly picked tickets show:

	Outcome if assigned to control group

The remaining 15 tickets show:

Outcome if assigned to treatment group	

The Hypotheses

- **Null**:
  - The distribution of all 31 potential control scores is the same as the distribution of all 31 potential treatment scores.

- **Alternative**:
  - The distribution of all 31 potential control scores is different from the distribution of all 31 potential treatment scores.

(DEMO)