DATA 8
Fall 2016

Lecture 37, November 28

Causality

Slides created by Ani Adhikari and John DeNero

Announcements

- Homework due today.
- The last (short) homework will be assigned today and is due on Saturday Dec 3 at the end of this week.
- We will post some practice problems on this week's material, but they will not be due.
- Project 3 due tomorrow Tuesday 11/29.
- GSI/Tutor office hour locations: Mondays 3106 Etcheverry, Wednesdays 3108 Etcheverry; no change on other days; no change to times.

A/B Testing

- Two random samples:
  - Sample A
  - Sample B

- Question: Are they drawn from the same underlying distribution?

- Answer by **A/B testing**

Permutation Test

- **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 attribute values and recompute the statistic; **repeat**; compare with the observed statistic

How Big is the Difference?

If you think that the two underlying population means might be different, you'll want to know how different they are.

- So instead of just running a "same/different" test, don't make any hypotheses. Just estimate the difference between the two population means.

- You can do this by bootstrapping the sample and constructing a confidence interval for the parameter: "difference between the population means".
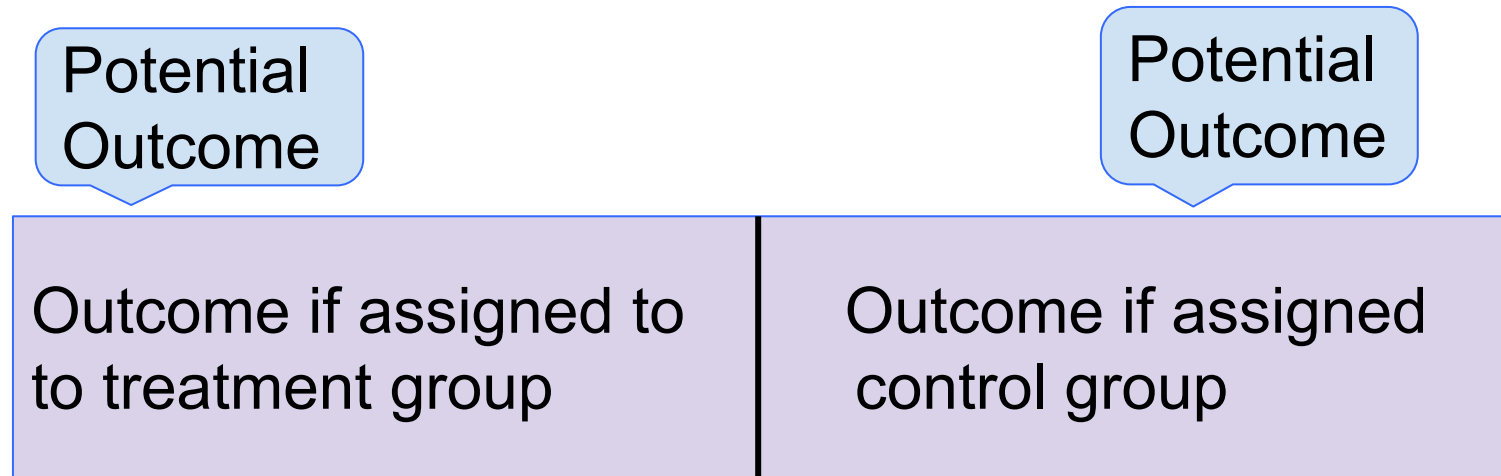
Causality

- 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:

Potential Outcome

Potential Outcome

| Outcome if assigned to to treatment group | Outcome if assigned control group |

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)

Estimating the Effect

- The distributions are numerical.

- So construct an approximate 95% confidence interval for the difference between the means of the two groups.

- No hypotheses needed. Just bootstrap the sample.

(Demo)