DATA 8

Spring 2020

Lecture 20

Causality

Announcements: Part 1

- Homework 7 due Thursday, March 12
- Midterm on March 13th, 7:10 - 9:00 pm
  - **Now online on Gradescope.com**
  - Must work COMPLETELY by yourself in a quiet room
    - Proctoring in 2050 VLSB if needed
  - You cannot use any resources outside of:
    - Midterm Reference Sheet
    - Blank scratch paper
    - Online textbook
  - Multiple versions - will receive email at 6:00 pm
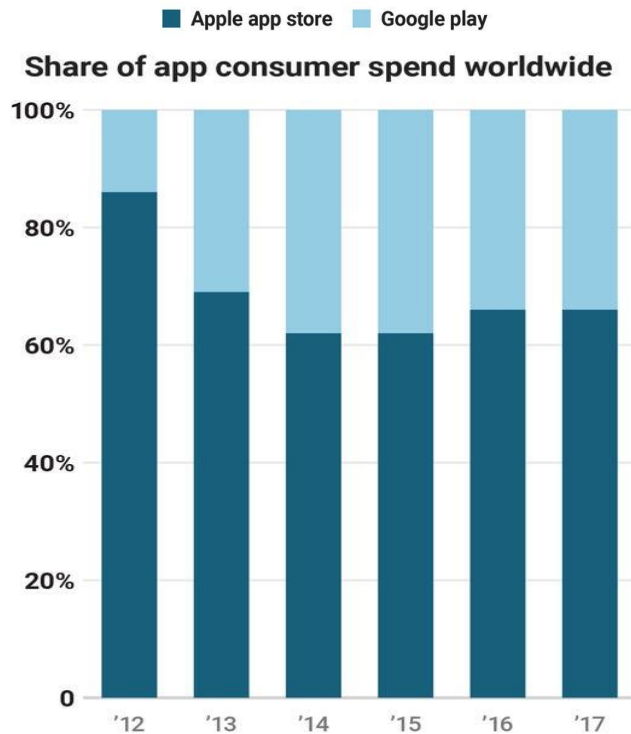
Announcements: Part 2

- In-Person Office Hours:
  - Cancelled today & Tuesday
- Lecture:
  - **Online only, starting this Wednesday**
- Labs:
  - This week:
    - Review worksheet walk-through posted Wednesday
  - **Next week onwards: in-person labs cancelled**
    - Walk-throughs posted Wednesdays
    - Lab notebooks due Fridays at 11:59pm
    - Lab rooms on campus still available but GSIs remote

Review: Comparing Two Samples

- Null:
  - In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)
- Alternative:
  - In the population, the babies of the mothers who smoked weigh less, on average, than the babies of the non-smokers.

(Demo)

Importance of Random Assignment

Apple users more willing to pay for apps

Apple app store ▪ Google play

Share of app consumer spend worldwide

statista | BUSINESS INSIDER

Importance of Random Assignment

- iOS users spend 2x as much as Android users on 3rd party apps
  - *Is **higher spending** <u>caused by</u> users owning **iPhone**?*
  - Can't Tell:
    - Users aren't randomly assigned a phone
    - Other factors contribute to their phone purchasing decisions (e.g. income, geography)
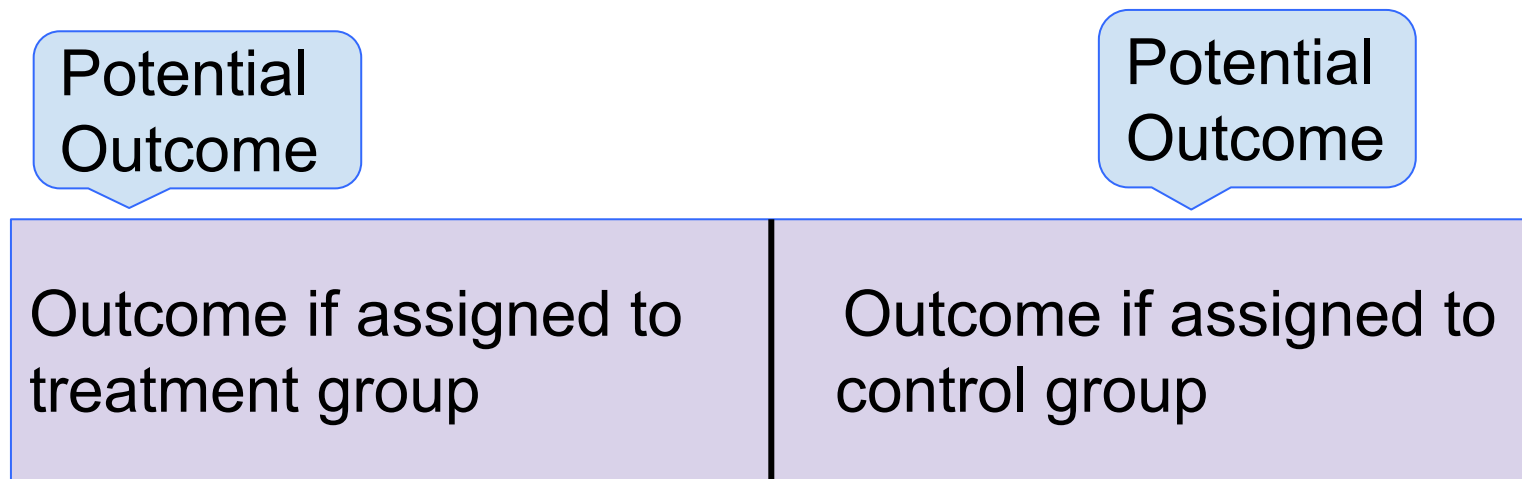
Causality

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:

Potential Outcome

Potential Outcome

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

The Data

16 randomly picked tickets show:

	Outcome if assigned to control group

The remaining 15 tickets show:

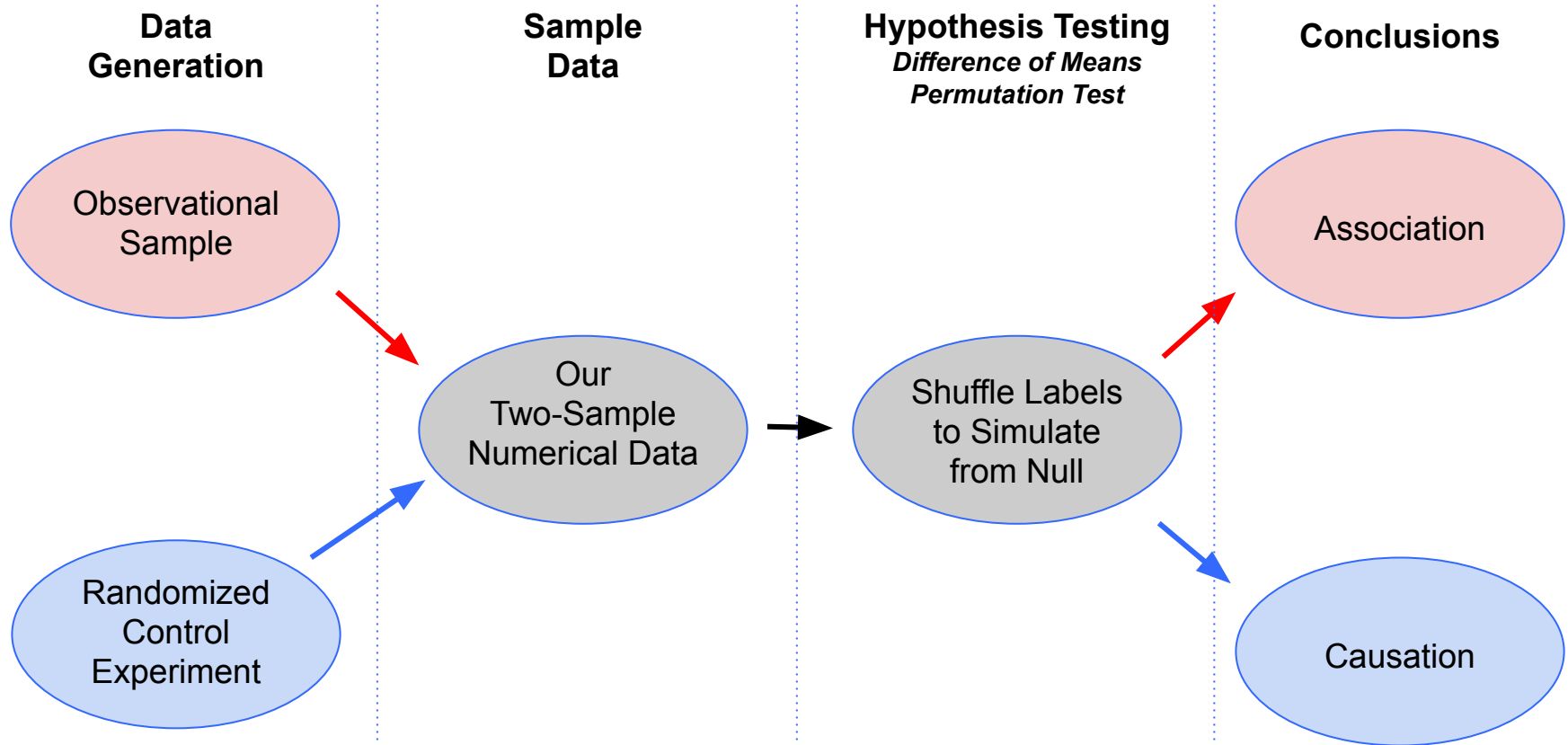| Outcome if assigned to treatment group	

The Hypotheses

- **Null:**
    - In the population, the distribution of all potential control scores is the same as the distribution of all potential treatment scores.
    - tl;dr the treatment has no effect

- **Alternative:**
    - In the population, more of the potential **treatment** scores are 1 (pain improves) than the potential **control** scores.

(Demo)

An Error Probability

Can the Conclusion be Wrong?

Yes.

	Null is true	**Alternative is true**
Test favors the null	✓	✗
Test favors the alternative	✗	✓

An Error Probability

- The cutoff for the *P*-value is an error probability.

- If:
  - your **cutoff is 5%**
  - and the **null hypothesis happens to be true**

- then there is about a **5% chance** that **your test will reject the null hypothesis**.

P-value cutoff vs P-value

- P-value cutoff
  - Does not depend on observed data or simulation
  - Decide on it before seeing the results
  - Conventional values at 5% and 1%
  - Probability of hypothesis testing making an error
- P-value
  - Depends on the observed data and simulation
  - Probability under the null hypothesis that the test statistic is the observed value or further towards the alternative