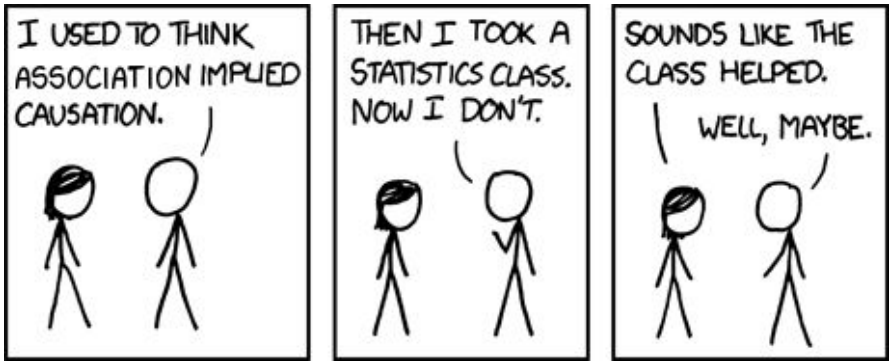
DATA 8
Fall 2022

Lecture 20

Causality



Announcements

- **Homework 7** is due Wednesday, 10/12
- No lab notebook this week
- **Midterm on Friday at 7pm**
  - Midterm Prep Guide, Past Exams
  - **Midterm Review Session** on Thursday 3:30-6:30pm
    - More info to be posted on Ed
- Tutoring worksheets, walkthroughs, etc. available here!

Weekly Goals

- **Today**
  - Causation
  - Randomized Control Experiments
- Wednesday
  - P-Value as an Error
  - Examples
- Friday
  - Midterm review

Recap: A/B Testing

(Demo)

Random Assignment

Importance of Random Assignment

We've concluded that in the population, birth weights of babies whose mothers smoke weigh less than those whose mothers do not

- *Is **lower birth weight** caused by maternal **smoking**?*
- Can't Tell:
  - Moms aren't randomly assigned whether to smoke
  - Other factors contribute to their decision to smoke (e.g. income, geography, diet)

Causality

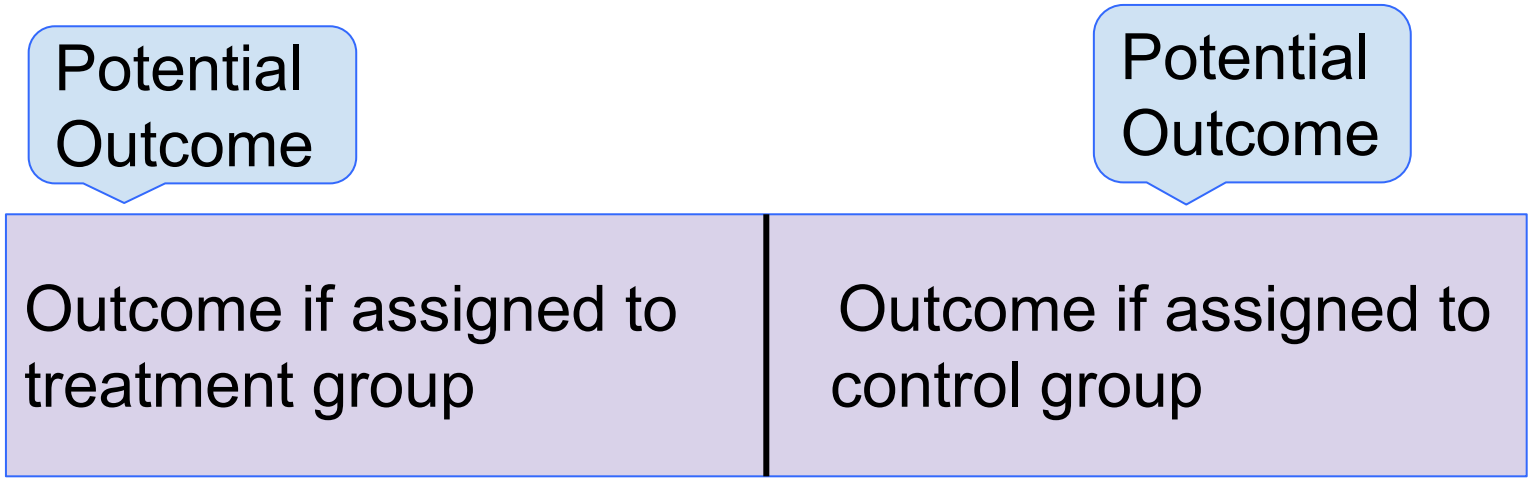
Randomized Controlled Experiment

- 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:

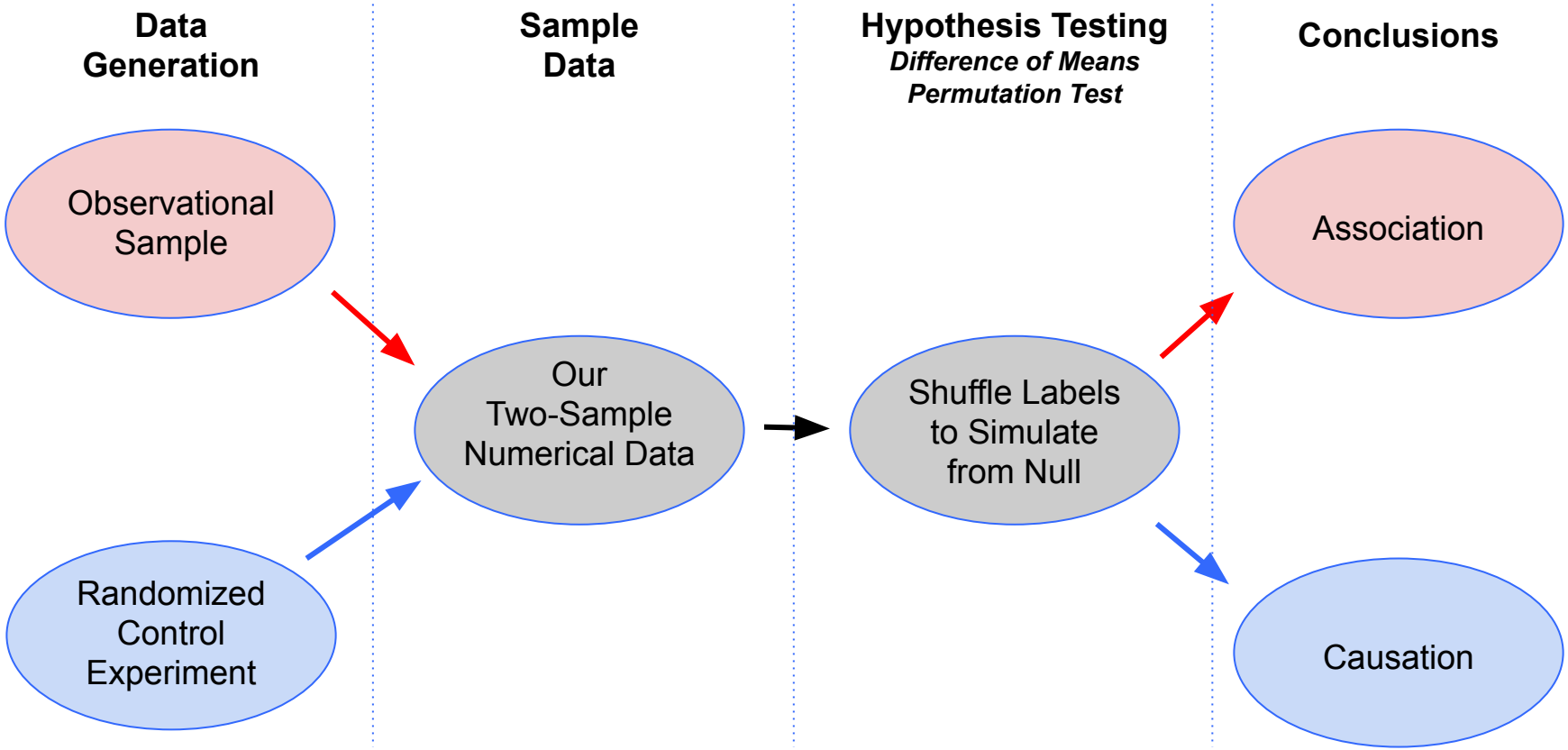
Outcome if assigned to treatment group	

The Hypotheses

- **Null:**
  - 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)

Random Assignment & Shuffling



P-Values and Error Probabilities

Discussion Question

There are 2000 students in Data 8. Each student tests

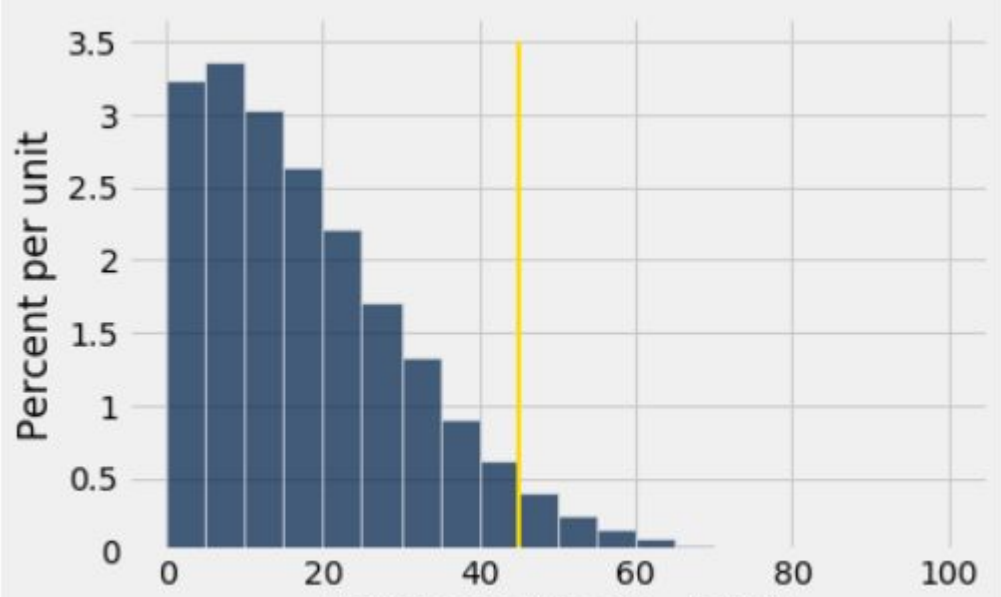
Null: The coin is fair

Alternative: The coin is unfair

- based on 1000 tosses of a coin,
- the statistic | number of heads - 500 |,
- and the 5% cutoff for the P-value.

Suppose all 1000 coins are fair. About how many students will conclude that their coins are unfair?

Statistic Simulated Under the Null



About 5% of the area is to the right of the gold line

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

How We've Tested Thus Far

Hypothesis Testing Review

- **1 Sample: One Category** *(e.g. percent of flowers that are purple)*
  - Test Statistic: `observed_proportion`, `abs(observed_proportion - null_proportion)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **1 Sample: More Than 2 Categories** *(e.g. ethnicity distribution of jury panel)*
  - Test Statistic: `tvd(observed_dist, null_dist)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **1 Sample: Numerical Data** *(e.g. scores in a lab section)*
  - Test Statistic: `observed_mean`, `abs(observed_mean - null_mean)`
  - How to Simulate: `population_data.sample(n, with_replacement=False)`
- **2 Samples: Underlying Values** *(e.g. birth weights of smokers vs. non-smokers)*
  - Test Statistic: `group_a_mean - group_b_mean`, `group_b_mean - group_a_mean`, `abs(group_a_mean - group_b_mean)`
  - How to Simulate: `observed_data.sample(with_replacement=False)`