

# BC COMS 1016: Intro to Comp Thinking & Data Science

## Lecture 15 – Causality & Estimation Variability

# Announcements - update



- HW06 - Testing Hypotheses
  - Due Tonight (Monday 11/22)
  - Allowed up to 3 or 4 late days? - your call
- Checkpoint/Project 2 (midterm):
  - Released due Monday 11/30
- Lab05 – moved to this week
  - Due Wednesday 11/23
- Thanksgiving week office hours

# Mid Semester Feedback



- more demos
- Pace picked up quickly
  - HW07 - Confidence Intervals
    - Was due Thursday 11/26
    - Pushed off
- Lots of assignments
  - Learning by doing



# Review: Assessing Models



- A model is a set of assumptions about the data
- In data science, many models involve assumptions about processes that involve randomness:
  - “Chance models”
- **Key question:** does the model fit the data?



The method only works if we can simulate data under one of the hypotheses.

- **Null hypothesis**

- A well defined chance model about how the data were generated
- We can simulate data under the assumptions of this model
  - “Under the null hypothesis”

- **Alternative hypothesis:**

- A different view about the origin of the data

# Approach to Assessing Models

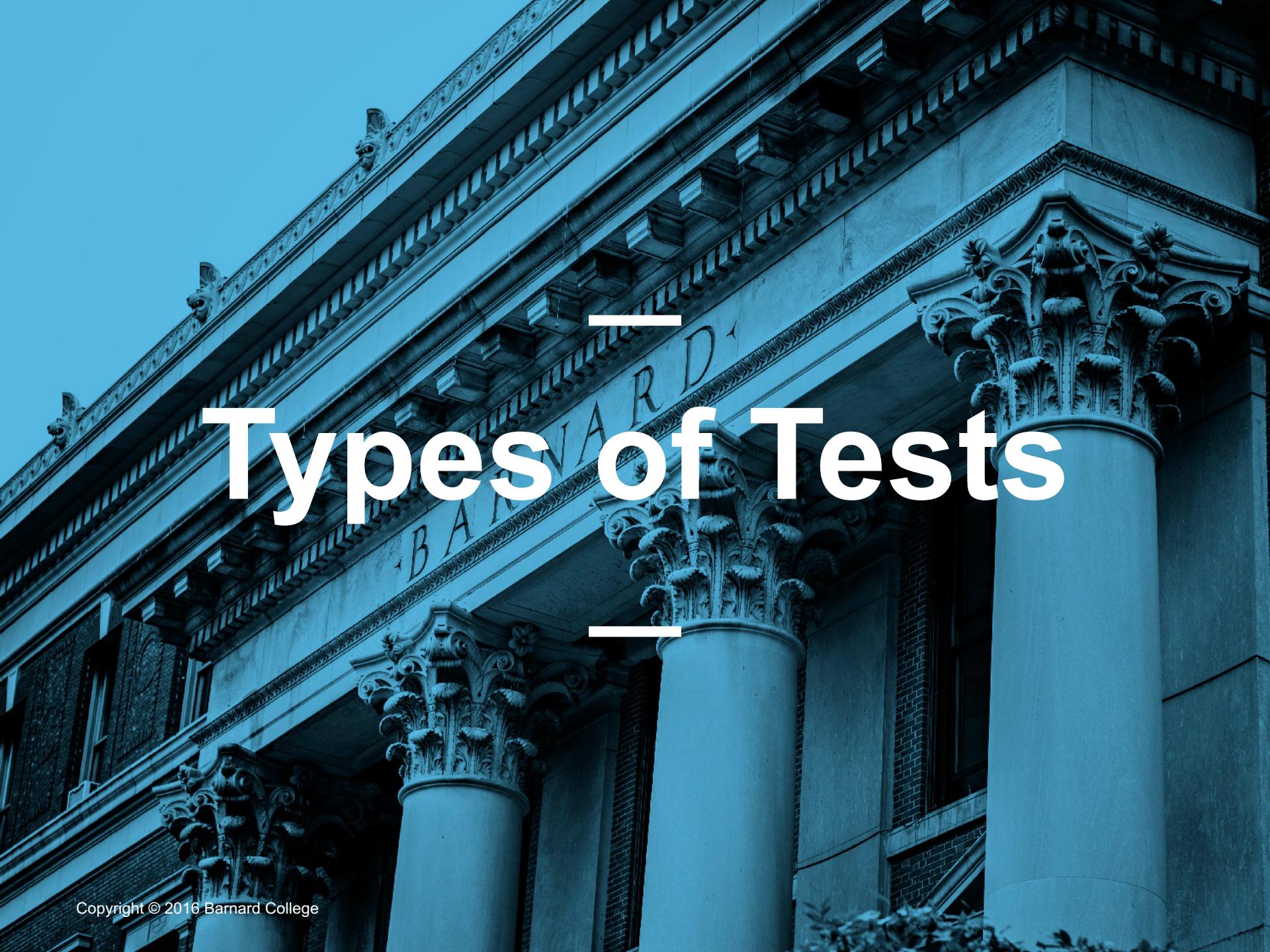


- If we can simulate data according to the assumptions of the model, we can learn what the model predicts
- We can compare the model's predictions (simulations) to the observed data
  - Here, "observed data" == what actually happened
- If the data and the model's predictions are not consistent, that is evidence against the model



# Steps in Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
- Simulate the statistic under the model’s assumptions
- Compare the data to the model’s predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample
- If the observed statistic is far from the histogram, that is evidence against the model



# Types — of Tests

# Hypothesis Testing Review



## 1 Sample: One Category (e.g. percent of black male jurors)

- Test Statistic: `empirical_percent`, `abs(empirical_percent - null_percent)`
- How to Simulate: `sample_proportions(n, null_dist)`

## 1 Sample: Multiple Categories (e.g. ethnicity distribution of jury panel)

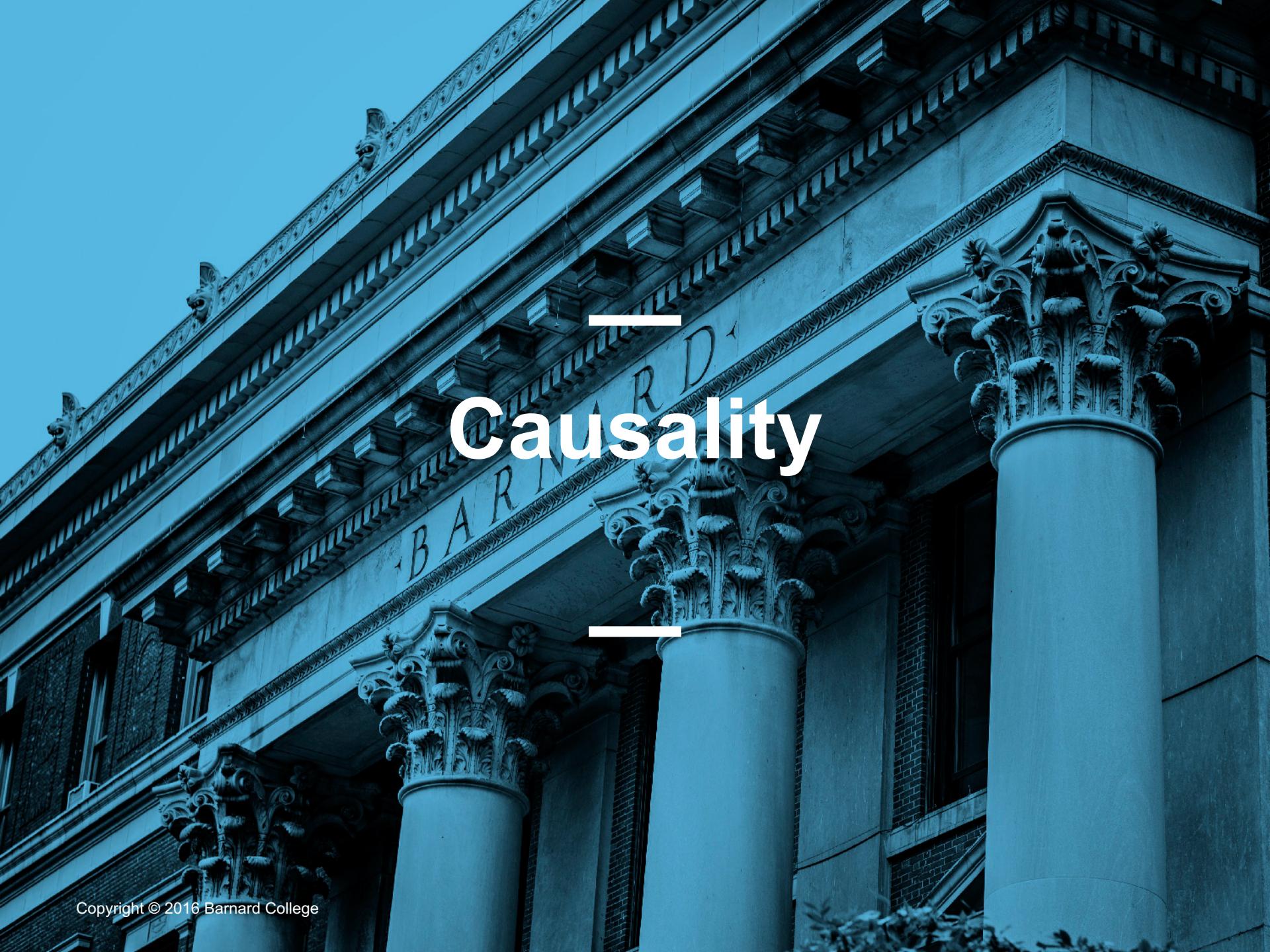
- Test Statistic: `tvd(empirical_dist, null_dist)`
- How to Simulate: `sample_proportions(n, null_dist)`

## 1 Sample: Numerical Data (e.g. scores in a lab section)

- Test Statistic: `empirical_mean`, `abs(empirical_mean - null_mean)`
- How to Simulate: `population_data.sample(n, with_replacement=False)`

## 2 Samples: Numerical Data (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: `empirical_data.sample(with_replacement=False)`



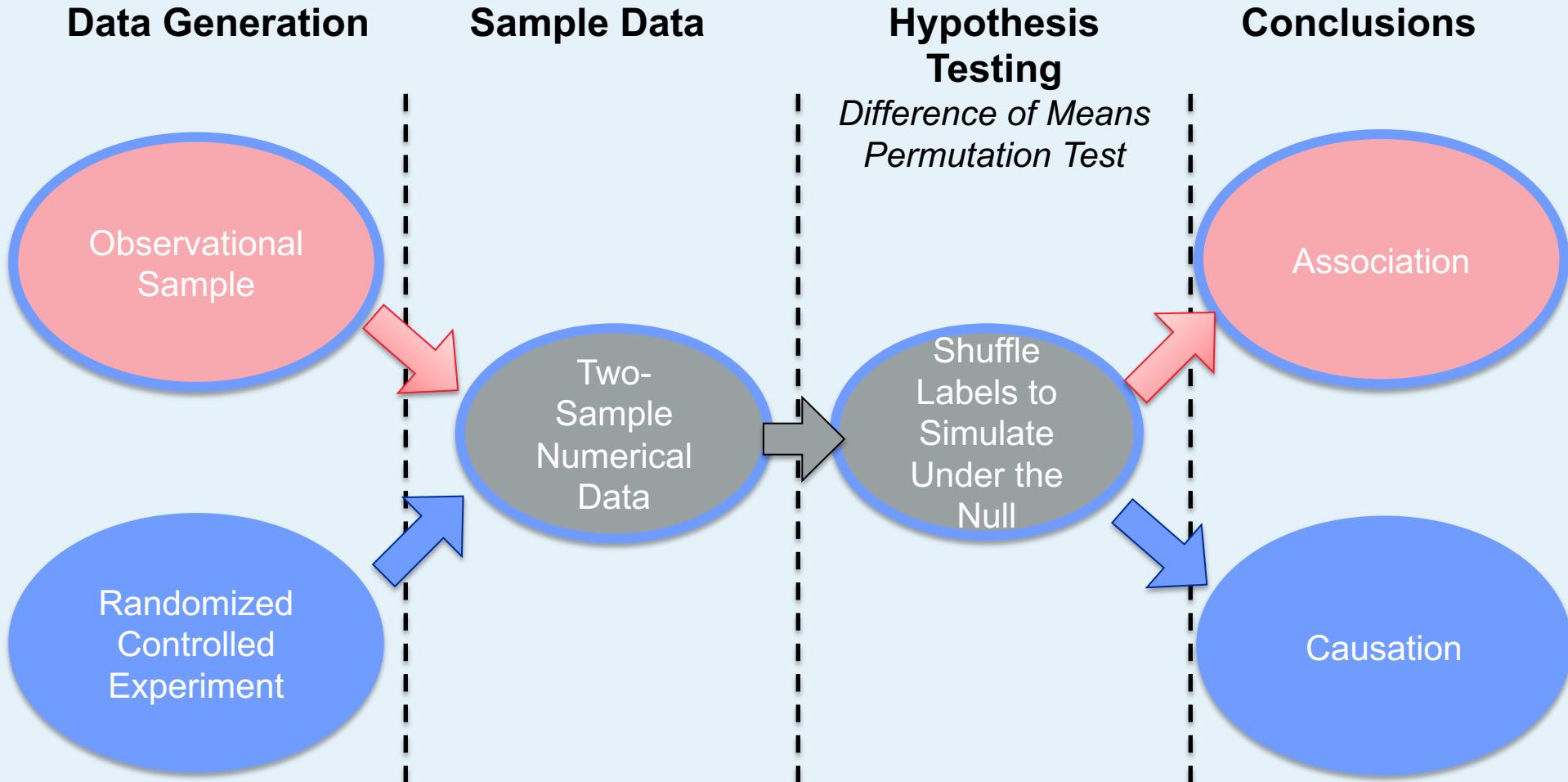
# Causality

# 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

# Randomized Assignment & Shuffling



# — Confidence Intervals

# Computing Percentiles



The Xth percentile is first value on the sorted list that is at least as large as X% of the elements

Example:

$s = [1, 7, 3, 9, 5]$

$s_{sorted} = [1, 3, 5, 7, 9]$

$\text{percentile}(80, s) = ?$

The 80th percentile is ordered element 4:  $(80/100) * 5$

For a percentile that does not exactly correspond to an element, take the next greater element instead

# The percentile Function



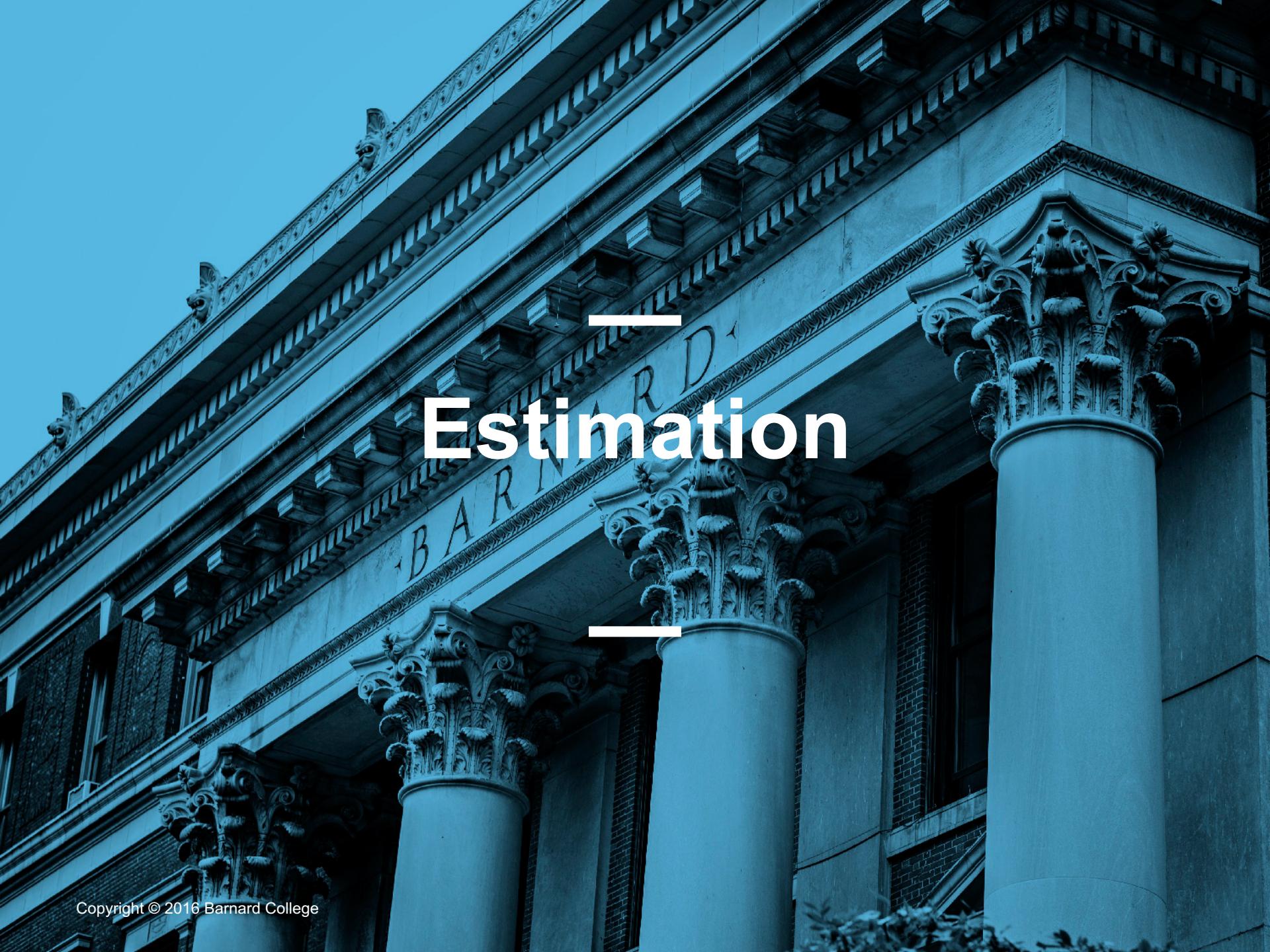
- The  $p$ th percentile is the **smallest value** in a set that is **at least as large as  $p\%$**  of the elements in the set
- Function in the datascience module:  
 $\text{percentile}(p, \text{values})$
- $p$  is between 0 and 100
- Returns the  $p$ th percentile of the array



## Discussion Question?

Which are True, when  $s = [1, 7, 3, 9, 5]$ ?

1.  $\text{percentile}(10, s) == 0$
2.  $\text{percentile}(39, s) == \text{percentile}(40, s)$
3.  $\text{percentile}(40, s) == \text{percentile}(41, s)$
4.  $\text{percentile}(50, s) == 5$



# BARNARD

# Estimation

# Inference: Estimation



- How do we calculate the value of an unknown parameter?
- If you have a census (that is, the whole population):
  - Just calculate the parameter and you're done
- If you don't have a census:
  - Take a random sample from the population
  - Use a statistic as an **estimate** of the parameter



# — Estimation Variability —

# Variability of the Estimate



- One sample → One estimate
- But the random sample could have come out differently
- And so the estimate could have been different
- Big question:
  - How different would it be if we did it again?

# Quantifying Uncertainty



- The estimate is usually not exactly right:

$$\text{Estimate} = \text{Parameter} + \text{Error}$$

- How accurate is the estimate, usually?
- How big is a typical error?
- When we have a census, we can do this by simulation

# Where to Get Another Sample?



- We want to understand errors of our estimate
- Given the **population**, we could simulate
  - ...but we only have the **sample!**
- To get many values of the estimate, we needed many random samples
- Can't go back and sample again from the population:
  - No time, no money
- Stuck?



# The Bootstrap