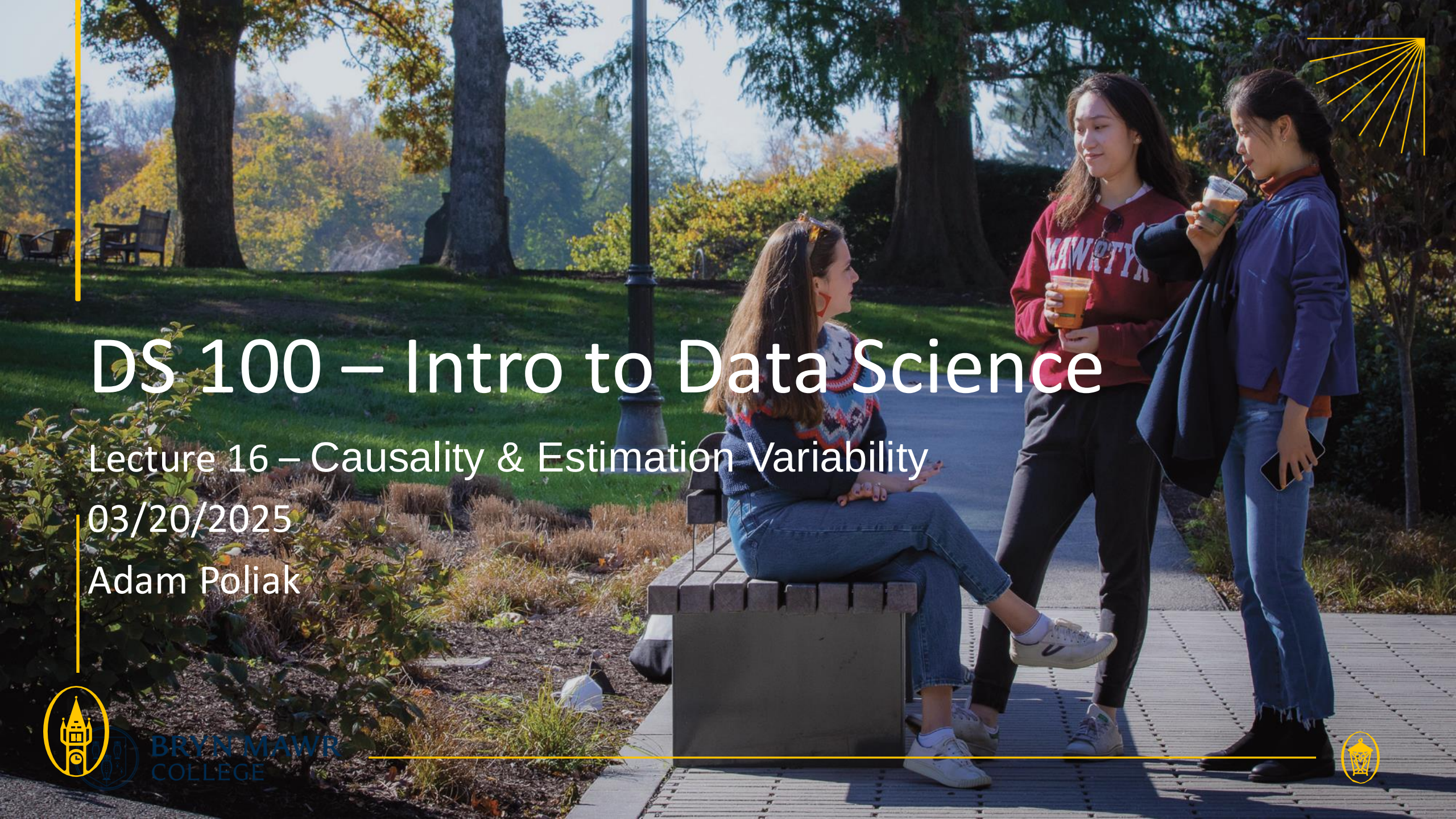


DS 100 – Intro to Data Science

Lecture 16 – Causality & Estimation Variability

03/20/2025

Adam Poliak

BRYN MAWR COLLEGE

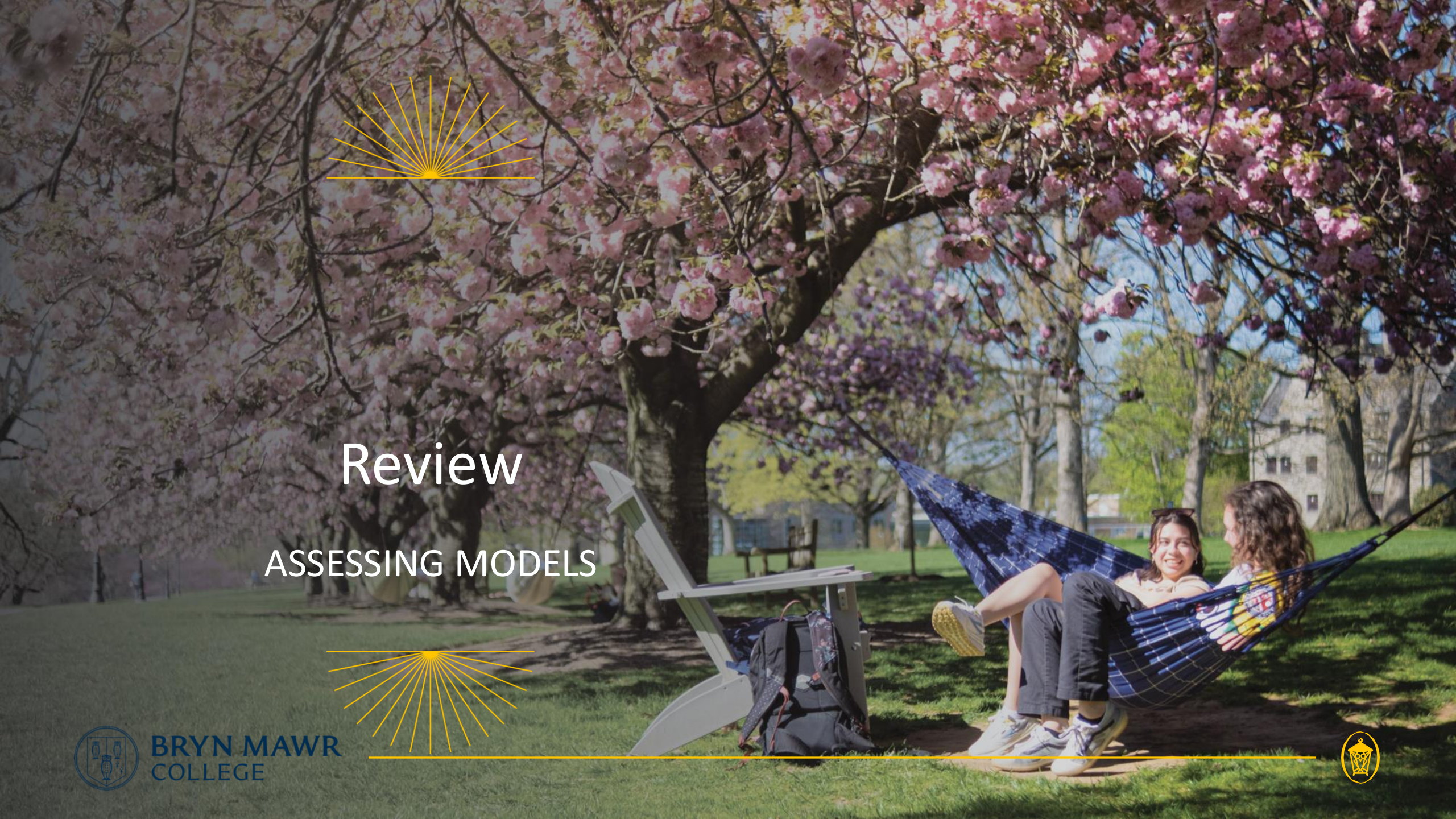
Announcements

Midsemester feedback form - https://forms.gle/M4jVdGTDgQknWAwo6

Midterm – Thursday (03/27)

Lab06 – Great British Bake-Off (A/B Testing)

- Due Friday (03/21)



Review

ASSESSING MODELS

BRYN MAWR COLLEGE

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?

Null and Alternative

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 the model and the data

Simulate statistic under the assumptions of the model

Compare the data to the model’s predictions:

- Draw a histogram of the simulated values
- Compute the observed statistic from the real sample

If the observed statistic is far from the histogram, that is evidence again the model


Types of Tests
BRYN MAWR
COLLEGE

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)

P-value

Definition of the P-value

Formal name: **observed significance level**

The *P*-value is the chance,

- Under the null hypothesis,
- That the test statistic
- Is equal to the value that was observed in the data
- Or is even further in the direction of the tail

It is not the probability of a Type I error


Causality
BRYN MAWR
COLLEGE

Randomized Controlled Experiment

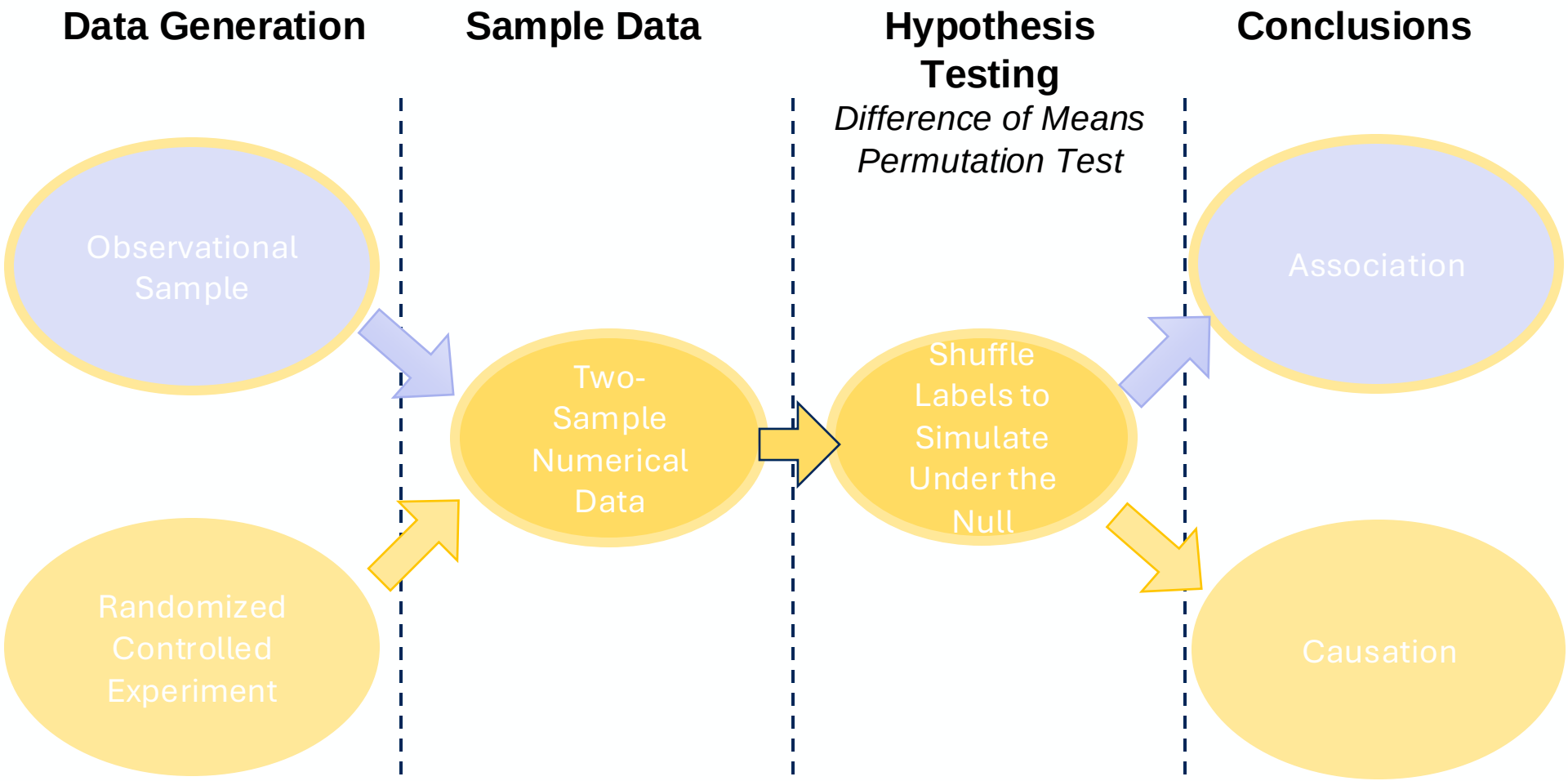
Sample A: **control group**
Sample B: **treatment group**

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



Percentiles

BRYN MAWR COLLEGE

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]
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:

```
percentile(p, values)
```

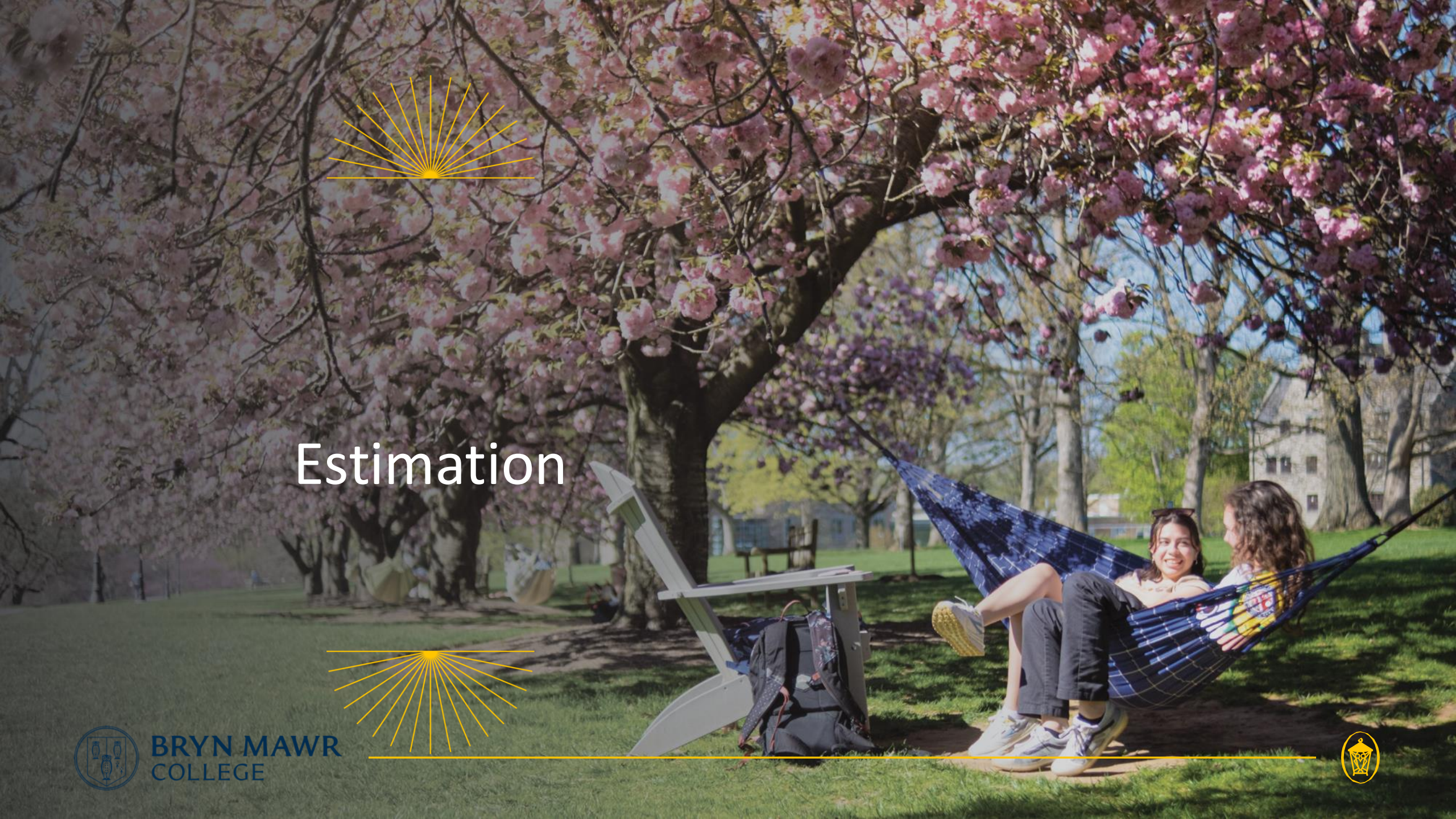
p is between 0 and 100

Returns the *p*th percentile of the array

Discussion Question

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

1. **`percentile(10, s) == 0`**
2. **`percentile(39, s) == percentile(40, s)`**
3. **`percentile(40, s) == percentile(41, s)`**
4. **`percentile(50, s) == 5`**



Estimation

BRYN MAWR COLLEGE

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