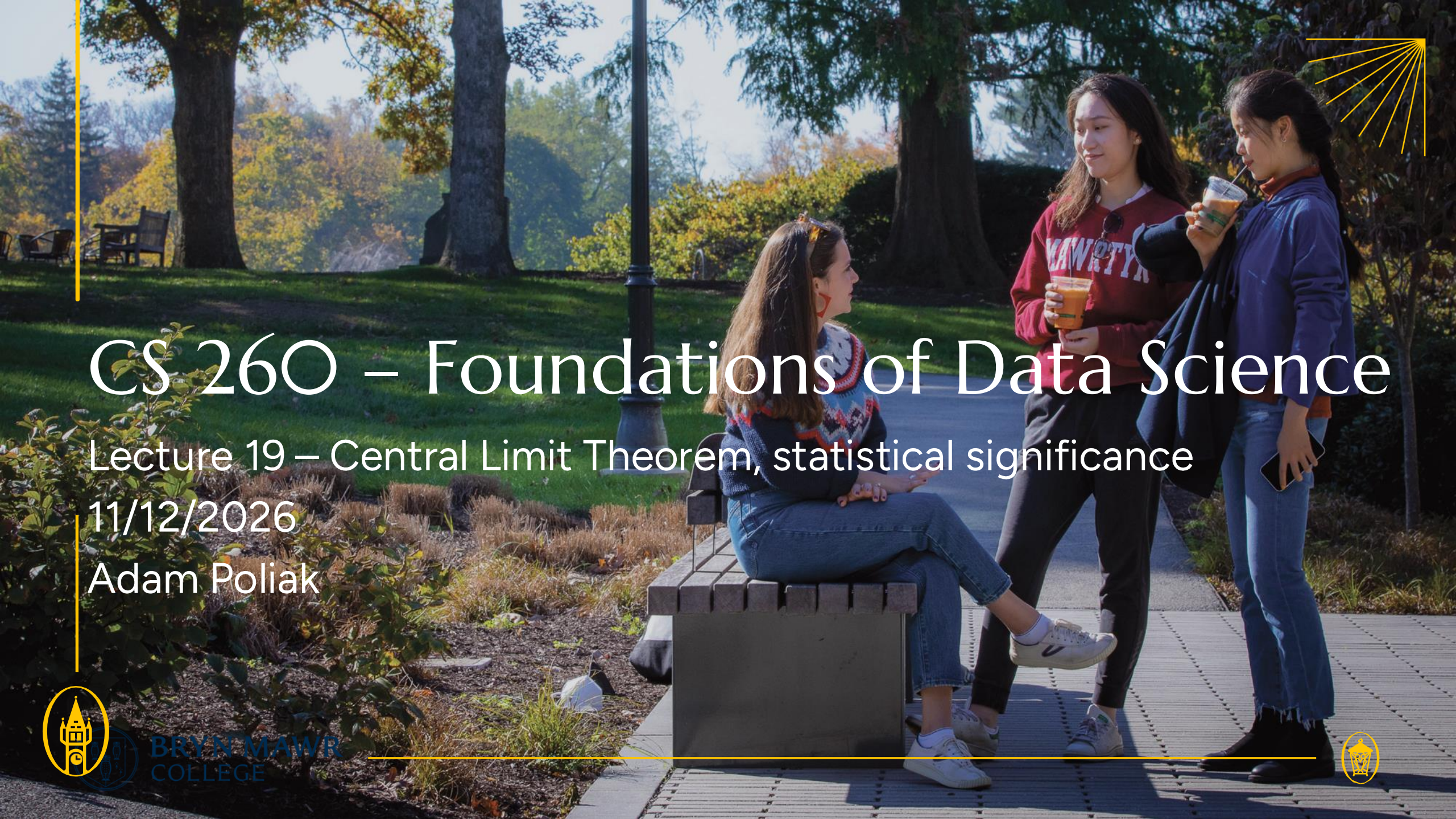


CS 260 – Foundations of Data Science

Lecture 19 – Central Limit Theorem, statistical significance

11/12/2026

Adam Poliak

BRYN MAWR COLLEGE

Announcements

HW08 due 11/17

Midterm 2

Distributions

Probability Distribution

Random quantity with various possible values

"Probability Distribution":

- All the possible values of a quantity
- The probability of each of the values

Computing the probability distribution:

- Math
- Simulation

Empirical Distribution

"Empirical" – based on observations

Observations can be a repeated experiment

"Empirical Distribution":

- All observed values
- The proportion of times each value appears

Large Random Samples



Law of Averages / Law of Large Numbers

If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that event occurs gets closer to the theoretical probability of the event

Example:

- As you increase the number of rolls of a die, the proportion of times you see the face with 5 dots gets closer to 1/6

Empirical Distribution of a Sample

If the sample size is large,
then the empirical distribution of a uniform random sample
resembles the distribution of the population,
with high probability


A Statistic
BRYN MAWR
COLLEGE

Inference

Statistical Inference:

- Making conclusions based on data in random samples

Example:

- Use the data to guess the value of an unknown number

fixed

Depends on the random sample

- Create an **estimate** of an unknown quantity

Terminology

Parameter

- Numerical quantity associated with the population

Statistic

- A number calculated from the sample

A statistic can be used as an **estimator** of a parameter

Probability distribution of a statistic

Values of a statistic vary because random samples vary

"Sampling distribution" or "probability distribution" of the statistic:

- All possible values of a statistic
- and all corresponding probabilities for each possible values

Can be hard to calculate:

- Either have to do math
- Or generate all possible samples and calculate the statistic based on the each sample

Empirical Distribution of a Statistic

Based on simulated values of a statistic

Consists of all observed values of the statistic,
and the proportion of times each value appeared

Good approximation to the probability distribution of a statistic

- If the number of repetitions in the simulation is large



Assessing models

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?

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 to the observed data

If the data and the model's predictions are not consistent, that is evidence against the model



Terminology

brynmawr.edu

Testing Hypotheses

A test chooses between two views of how data are generated

What are these views called?

- Answer: **hypotheses**

The test picks the hypothesis that is better supported by the observed data

What is the method for choosing between the hypotheses?

- Simulate data under one of the hypotheses
- Compare the simulation results and the observed data
- Pick one of the hypotheses based on whether the simulated results and observed data are consistent

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

Test Statistic

The statistic that we choose to simulate, to decide between the two hypotheses

Questions before choosing the statistic:

What values of the statistic will make us lean towards the null hypothesis?

What values will make us lean towards the alternative?

- Preferably, the answer should be just a “high” or just a “low” value
- Try to avoid “both high and low”

Prediction Under the Null Hypothesis

Simulate the test statistic under the null hypothesis

- Draw the histogram of simulated values
- **The empirical distribution of the statistic under the null hypothesis**

It is a prediction about the statistic, made by the null hypothesis

- It shows all the likely values of the statistic
- Also how likely they are (**if the null hypothesis is true)**

The probabilities are approximate, because we can't generate all the possible random samples

Conclusion of the Test

Resolve choice between null and alternative hypotheses

Compare the **observed test statistic** and its empirical distribution under the null hypothesis

If the observed value is not **consistent** with the empirical distribution

- The test favors the alternative
- “data is more consistent with the alternative”

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- If not, there are conventions about “consistency”

Statistical Significance



brynmawr.edu

Tail Areas

Conventions On Inconsistency

"Inconsistent with the null": The test statistic is in the tail of the empirical distribution under the null hypothesis

"In the tail," first convention:

- The area in the tail is less than 5%
- The result is "statistically significant"

"In the tail," second convention:

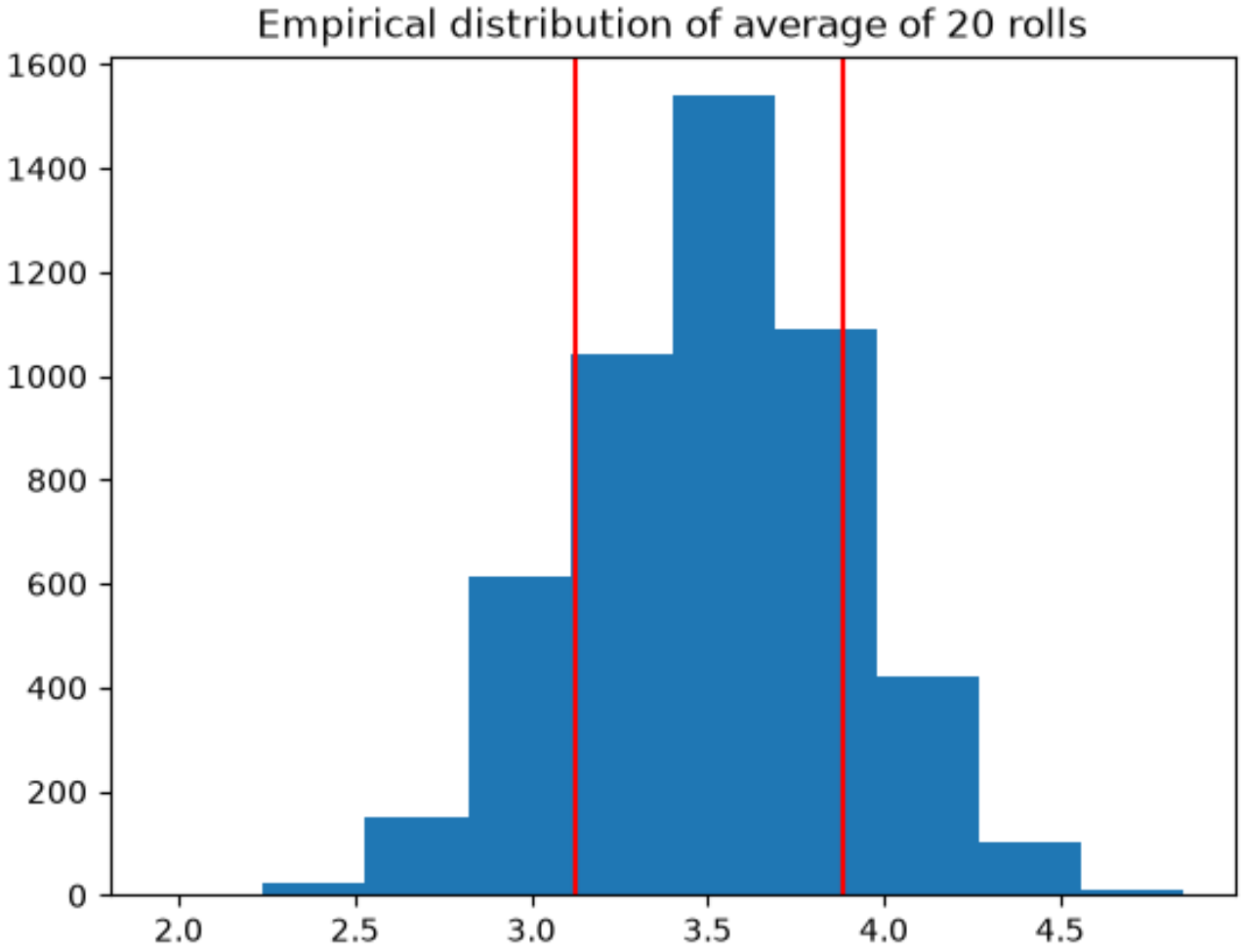
- The area in the tail is less than 1%
- The result is "highly statistically significant"

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



Mean is 3.5,
Std 0.379
67% of the empirical distribution lies between the red lines