GENERAL ASSEMBLY

INTRODUCTION TO STATISTICAL INFERENCE

Matt Brems

Data Science Immersive, GA DC

DATA SCIENCE WORKFLOW

1. Define the problem.
2. Obtain the data.
3. Explore the data.
4. Model the data.
5. Evaluate the model.
6. Answer the problem.

POPULATIONS

- Most data science problems have to do with studying **populations** in some form or another.

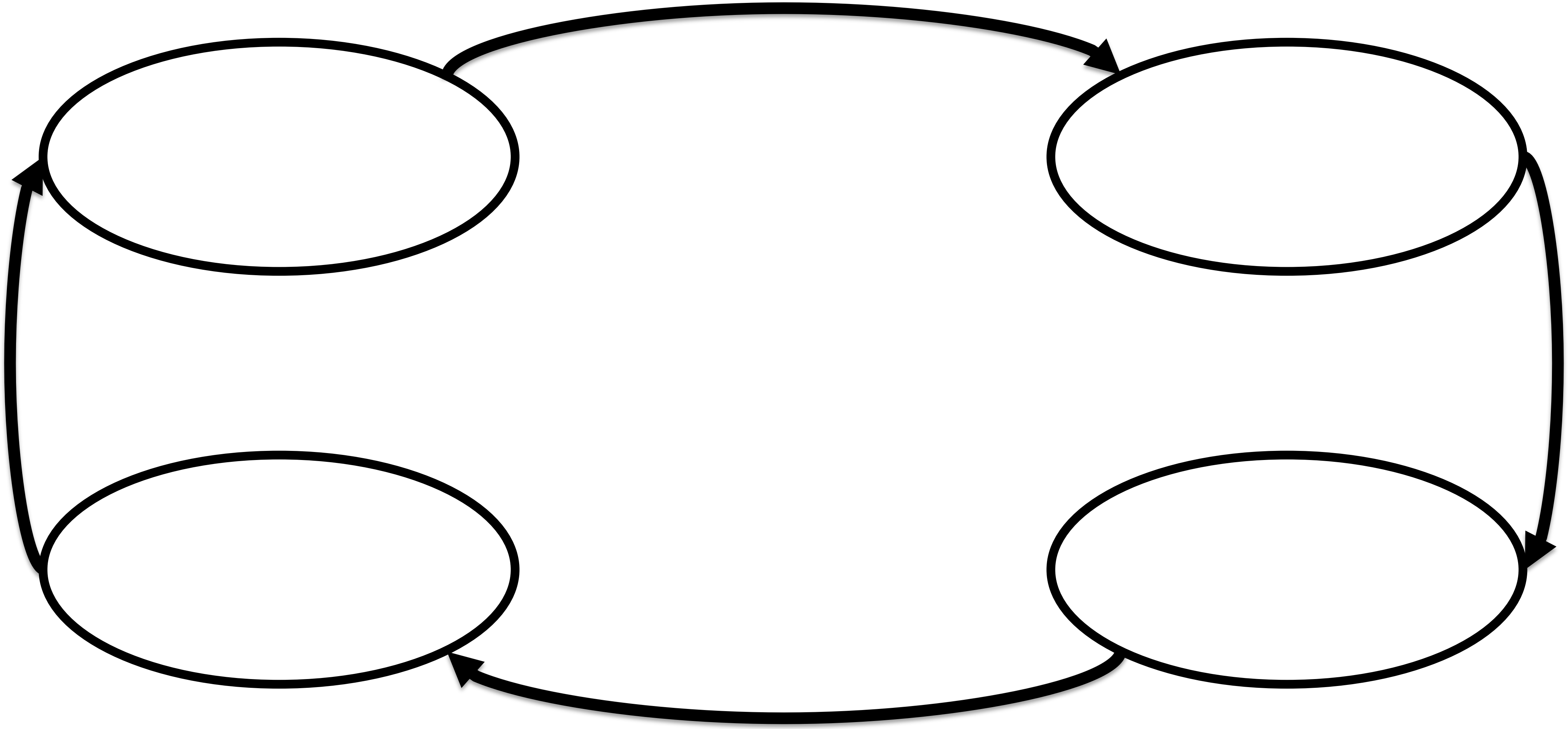
- Examples:
  - All undergraduates currently at Ohio State.
  - All microwaves constructed at my factory this year.
  - All hurricanes to enter the Gulf of Mexico.
  - All people who will vote in the 2020 election.
  - All states (and their average standardized test scores).

POPULATIONS

- If we're interested in learning about populations, why don't we just measure the population directly?

- What might we do instead?

GOAL: LEARNING ABOUT A POPULATION



EXAMPLE

- I want to see who will win the California U.S. Senate election in 2020. I call 1,000 registered voters and ask who they will support.

- Population:

- Sample:

- Statistic(s):

- Parameter(s):

EXAMPLE

- I developed a new drug (“New Drug”) that I believe reduces the diastolic blood pressure of adults over 50. I lead a clinical trial of 100 patients, where I compare my drug to the standard drug (“Old Drug”).
- Population:
- Sample:
- Statistic(s):
- Parameter(s):

STEPS

1. We identify our **population**.

2. We gather a **random sample** of data from the population.

3. We calculate some **statistic(s)** based on our sample.

4. Using statistics, we conduct inference on the **parameters**.

5. We use our understanding of **parameters** to make conclusions about the population.

STATISTICAL INFERENCE

- Today, we are going to discuss the process of **statistical inference**.
  - That is, how do we get from our **statistics** (measures of samples) to our **parameters** (measures of populations)?

- In frequentist statistical inference, there are two main ways to generalize from a sample to a population:
  - Confidence Intervals
  - Hypothesis Tests