COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 13: Models

March 4, 2026

Upcoming Labs and HW4

- **There are no labs this week or next week**
  - Labs resume after spring recess (March 25/26)
- **HW 4 is due next week Monday**
  - It was released last week Monday, but we're giving you all a little extra time to work on it
  - HW 5 will not be released until after spring recess

Midterm Logistics

- Midterm is **next week** on Wednesday, March 11
- You are permitted to bring a single formula sheet (5”x8” index card, double-sided) that will be submitted along with the exam
  - Exam is otherwise closed-note and no computers
- **Exam will only cover material up until last lecture**
  - Any new material introduced after this lecture will not be on the exam
- There will be a review session during class on Monday, March 9

Programming

Iteration

Manipulating Arrays & Tables

Data Types

Functions

Building Visualizations

Conditionals

Midterm Exam

Statistics

Probabilities

Confidence Intervals

Correlation

Linear Regression

P-value & Statistical Significance

Residuals

Final Project

Review: Distributions

Distributions

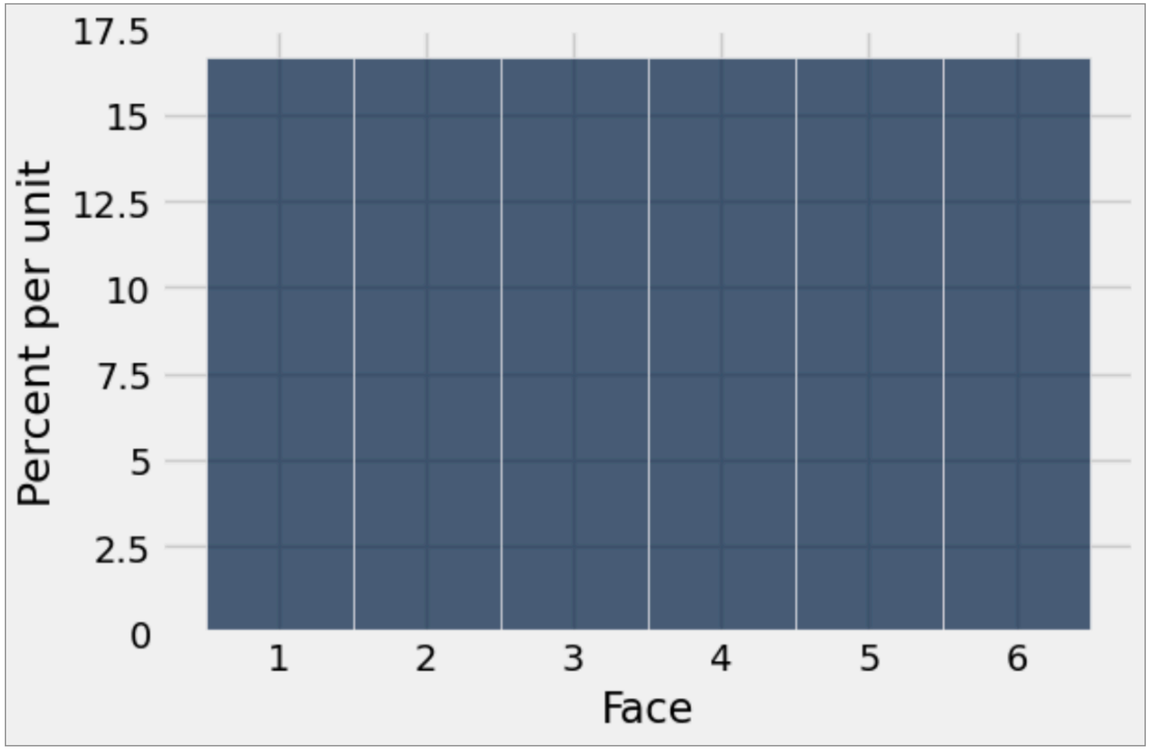
Recall categorical distributions (how often each unique value appears)

- If we have a population, we can get a list of unique values and how often they appear

- If we take a sample, the list of unique values may change based on the sample

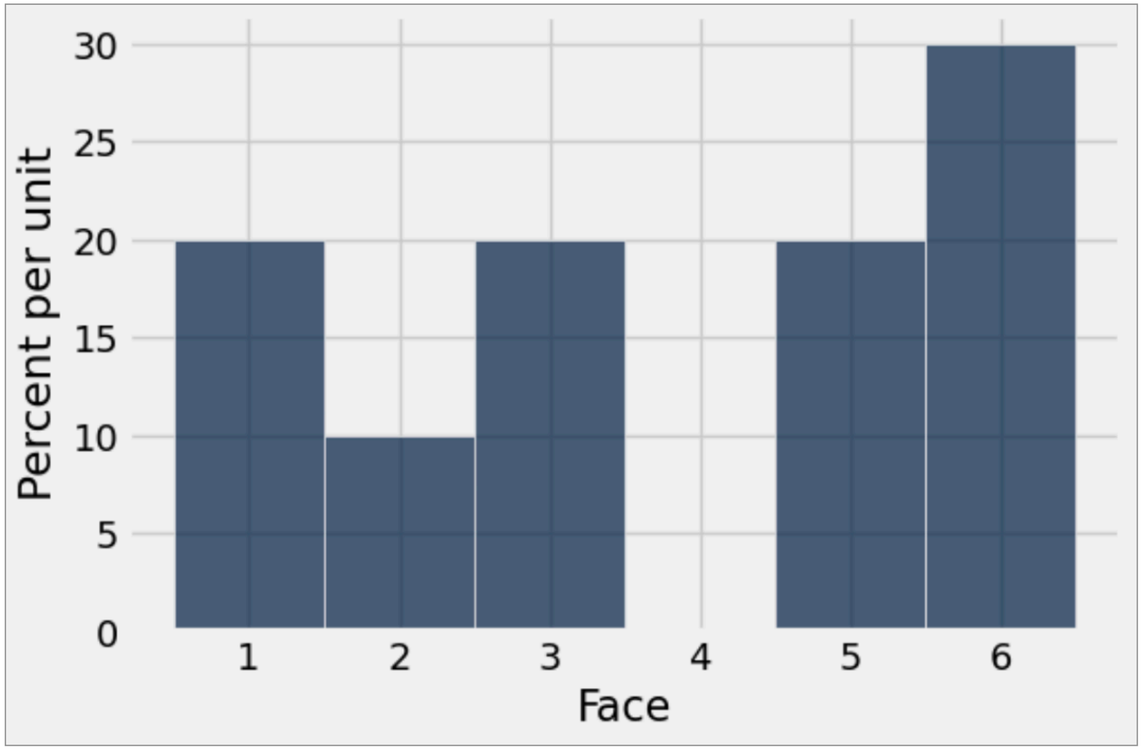
    - There can be differences in what's in the population vs what we see based on taking samples

Distributions



Probability Distribution

Contains all possible values and the probability of each value

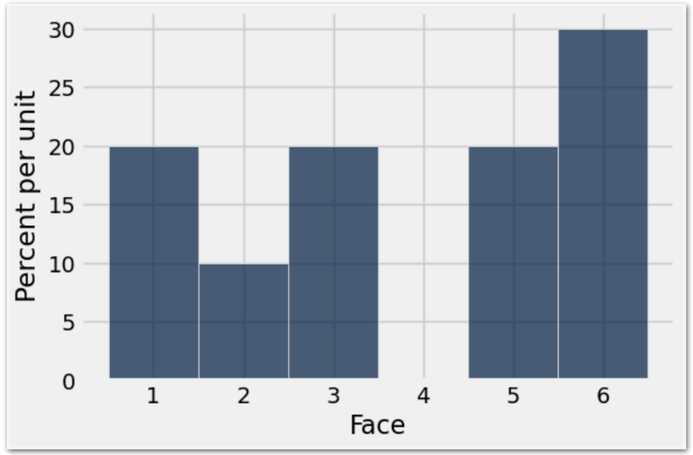


Empirical Distribution

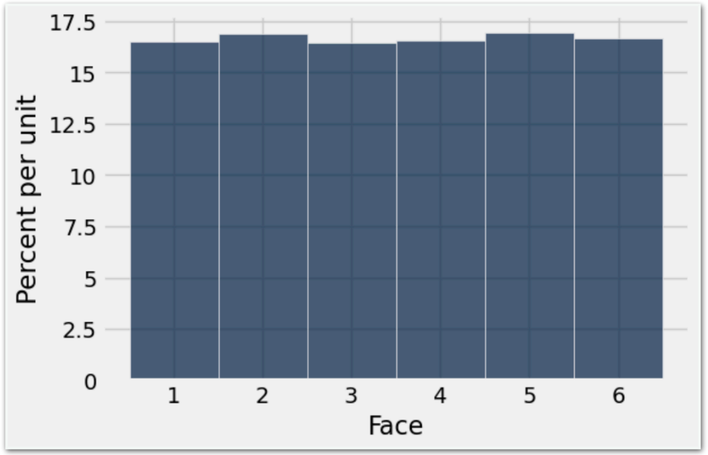
Distribution of **observed** values of a **sample** (in this case, sample = 10 die rolls)

Law of Averages / Large Numbers

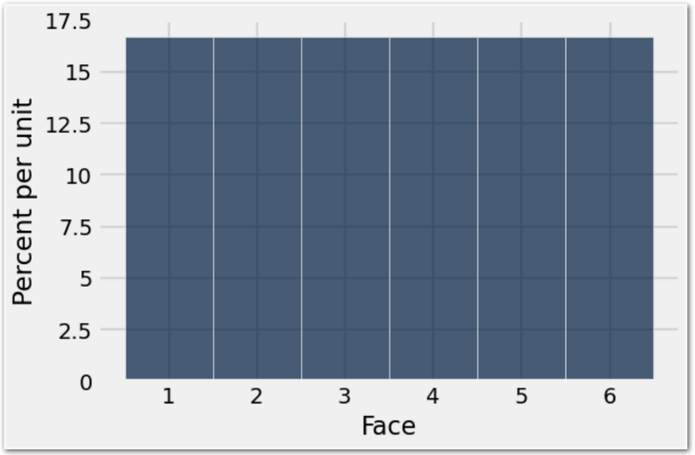
If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the true probability of the event



Empirical Distribution of 10 rolls



Empirical Distribution of 10,000 rolls



Probability Distribution for Rolling Dice

Terminology

Parameter: Number associated with the population

- Example: average, max, min, mean

Statistic: A number calculated from the sample, can be used to describe the distribution

- A statistic can be used as an estimate of a parameter

- Example: sample mean, sample max, sample min

Statistical Inference

Statistical Inference: drawing conclusions based on data in random samples

- Create an **estimate** of an **unknown value** using sample data and statistics

- Inference occurs from not being able to know an entire population

    - Estimates change based on the sample you draw

    - Statistics help you measure how much you expect those differences to vary

Mean vs Median

- Mean is the average
  - Sum of all the elements divided by the number of elements
- Median is the “middle value”
  - Value that separates the lower half and higher half of a sample

1, 3, 3, 6, 7, 8, 9

$$\text{mean} = \frac{1+3+3+6+7+8+9}{7}$$

$$= \frac{37}{7}$$

$$= 5.28$$

$$\text{median} = 6$$

Probability vs Empirical Distribution of a Statistic

For every sample size, a statistic has both a probability and empirical distribution

- **Probability Distribution**: All possible values and their associated probabilities

- **Empirical Distribution**: Based on simulated values of a statistic and include all observed values along with the proportion of times the value appeared

    - Can approximate probability distribution if number of repetition is high

Assessing Models

Models

A model is a set of assumptions about data

- In data science, many models involve assumptions about the processes that involve randomness

- Question: Does the model fit the data?

Assessing Models

- Suppose we have a statistical model that describes how data should behave (based on certain assumptions)
    - If we can use that model to **generate (simulate)** fake data, then we can see what the model “thinks” the data should look like
    - By simulating data, we can see the kinds of outcomes or patterns the model expects, i.e., its **predictions**
- We can then compare the predictions to the data that were observed (irl data)
- If the data and the model’s predictions are not consistent, that is evidence against the model.

1960s Supreme Court Case: *Swain v Alabama*

Amendment VI of the U.S. Constitution:

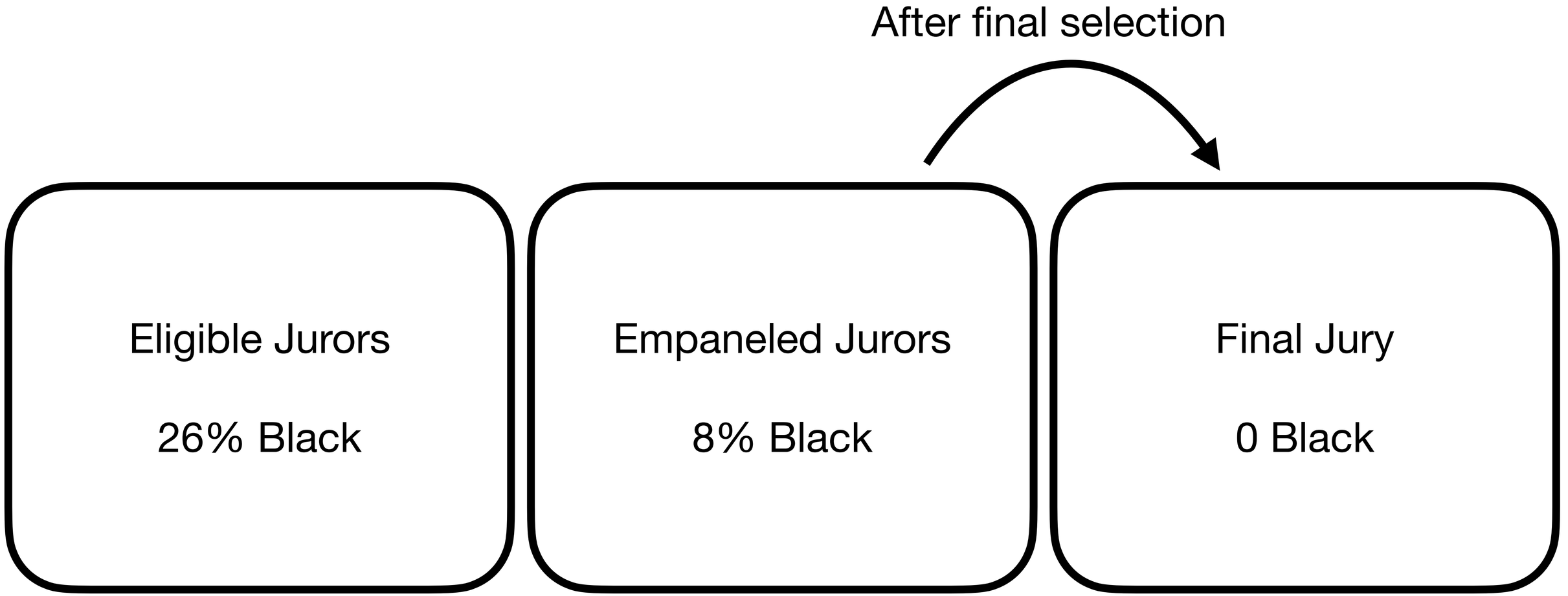
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed"

- An *impartial* jury should be selected from a panel that is representative from the population of the relevant region

Robert Swain was a Black man indicted and convicted by an all white jury in Talladega County, Alabama

- He appealed to the Supreme Court due to the lack of representation on juries in Talladega County

Swain vs. Alabama Jury



The Appeal

- Swain’s argument: the juries in Talladega County were not representative of the population and were therefore unfair

- Supreme Court Decision:

  - “The overall percentage disparity has been small” (between 26% and 8%) - deemed not different enough to indicate that Black panelists were systematically excluded

Our Question

- Would an 8% black jury be a realistic outcome if jury section were truly unbiased?

- In this case - the distribution we care about is categorical

  - % black vs % non-black

Sampling from a Categorical Distribution

- Sample at random from a categorical distribution

  – `sample_proportions(sample_size, pop_distribution)`

- `pop_distribution` is a list or array that adds up to 1

- Function returns an array containing the empirical distribution of the categories in the sample

Steps in Assessing a Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values
   - This is the model's prediction for how the statistic should come out

4. Compute the statistic from the sample in the study
   - If the two are not consistent: evidence against the model
   - If the two are consistent: data supports the model *so far*

Assessing *Swain v Alabama*

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Panelists were selected at random and the small number of Black panelists is by chance

Alternative view: too few Black panelists for it to have been a random sample

Statistic: Number (count) of Black panelists

Notebook Demo: *Swain vs. Alabama*

Our Conclusion

- Percent of Black panelists (8%) was **highly unlikely** under random sampling
  - It's *possible*, but extremely unlikely
  - Suggests the assumptions about the model are wrong
- We observed **statistical bias**, when differences between the parameters and the statistics are systematically in one direction
- The verdict was eventually overruled in *Baton v. Kentucky* (1986)
  - "The dismissal of jurors without stating a valid cause for doing so—may not be used to exclude jurors based solely on their race."

Next Class

- Today
  - Assessing Models
- Monday
  - Midterm Review
- Wednesday
  - Midterm!