DATA 8

Spring 2022

Lecture 16

Empirical Distributions

Announcements

- Lab 5 due today
- Project 1 due tonight
- HW 6 due next Thursday
    - A bit longer, so please start early!
- Midterm on March 11th, 7-9pm PT

Weekly Goals

- Last Week
  - Simulation
  - Chances
- Wednesday
  - Methods of sampling
  - Distributions of large random samples
- **Today**
  - Models that involve chance
  - Assessing the consistency of the data and the model

Review: Distributions

- Any random quantity has a **probability distribution**:
    - All **possible** values it can take
    - The **probability** it takes each value

- After repeated draws, it has an **empirical distribution**:
    - All **observed** values it took
    - The **proportion of times** it took each value

- After **many independent draws**, the empirical distribution looks more and more like the probability distribution

A Statistic

Inference

- **Statistical Inference:**

  Making conclusions based on data in random samples

- **Example**:

  fixed

  Use the data to guess the value of an unknown number

  depends on the random sample

  Create an **estimate** of the unknown quantity

Terminology

- **Parameter**
    - A number associated with the population
- **Statistic**
    - A number calculated from the sample

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

(Demo)

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 the statistic,
    - and all the corresponding probabilities
- Can be hard to calculate
    - Either have to do the math
    - Or have to generate all possible samples and calculate the statistic based on each sample

Empirical Distribution of a Statistic

- Empirical distribution of the statistic:
  - Based on simulated values of the statistic
  - Consists of all the observed values of the statistic,
  - and the proportion of times each value appeared

- Good approximation to the probability distribution of the statistic
  - if the number of repetitions in the simulation is large

(Demo)

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 Assessment

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts.

- We can then compare the predictions to the data that were observed.

- If the data and the model’s predictions are not consistent, that is evidence against the model.

Today’s Examples

Some Goals of Data Science

- Understand the world better
- Help make the world better

For example

- Help expose injustice
- Help counter injustice

The skills that you have gained empower you to do this.

First Example

- U.S. Constitution grants equal protection under the law
- All defendants have the right to due process

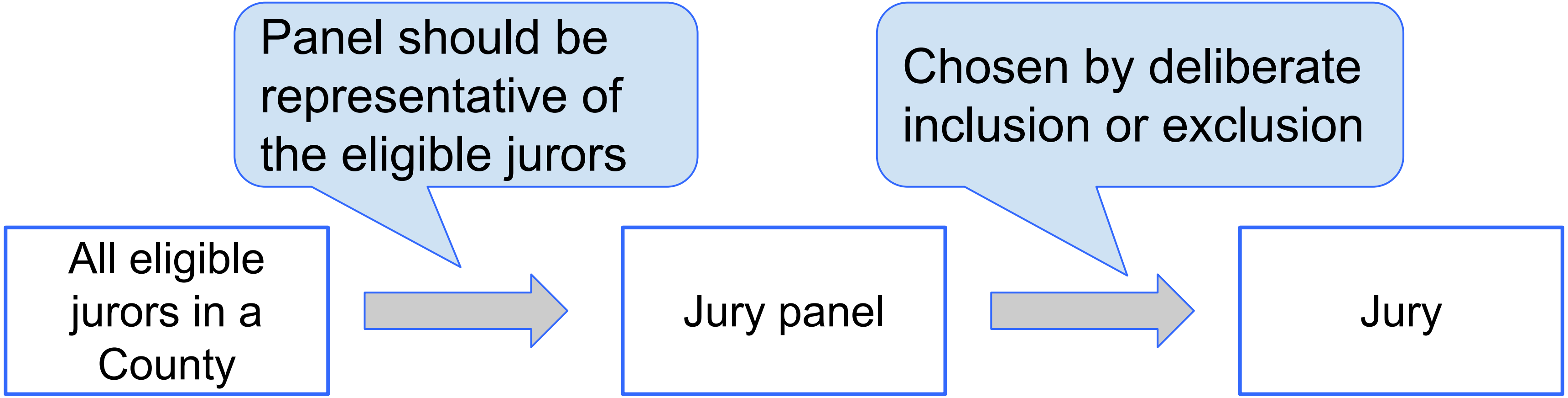
We will study a U.S. Supreme Court case in the 1960s

- A Black defendant was denied his Constitutional right to a fair jury
- The Court made incorrect and biased judgments about
  - the data in the case
  - the legal processes in the defendant’s original trial
- We will discuss errors and racial bias in the Court’s judgment

This case became the foundation of significant reform.

Jury Selection

US Constitution:
“right to a speedy and public trial, by an impartial jury”



Panel should be representative of the eligible jurors

Chosen by deliberate inclusion or exclusion

All eligible jurors in a County

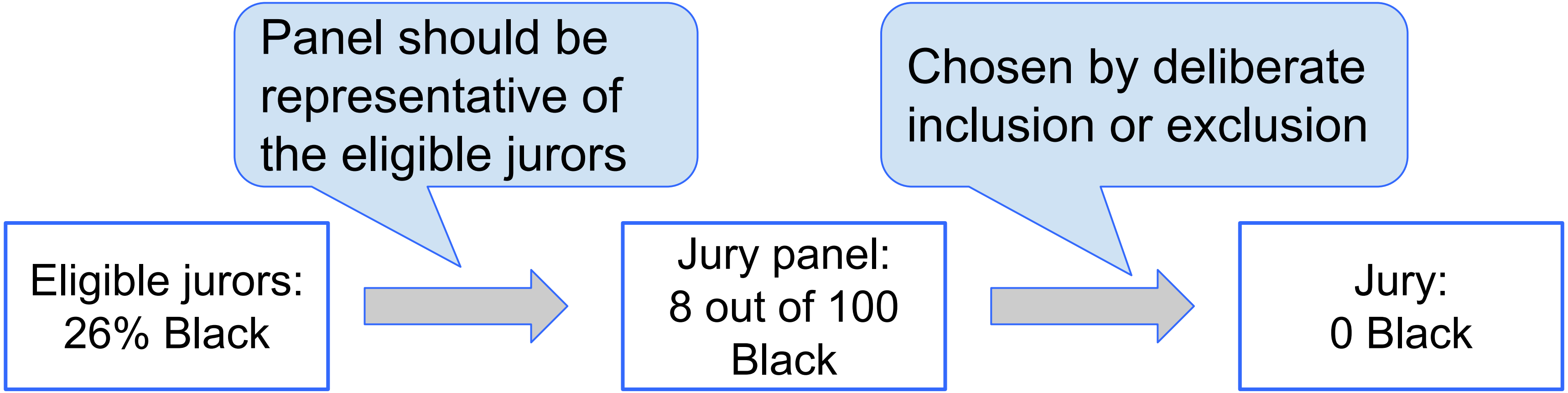
Jury panel

Jury

Supreme Court Case

Robert Swain's Case

- Robert Swain, a Black man, was convicted in Talladega County, AL
- He appealed to the U.S. Supreme Court
- Main reason: Unfair jury selection in the County's trials



Panel should be representative of the eligible jurors

Chosen by deliberate inclusion or exclusion

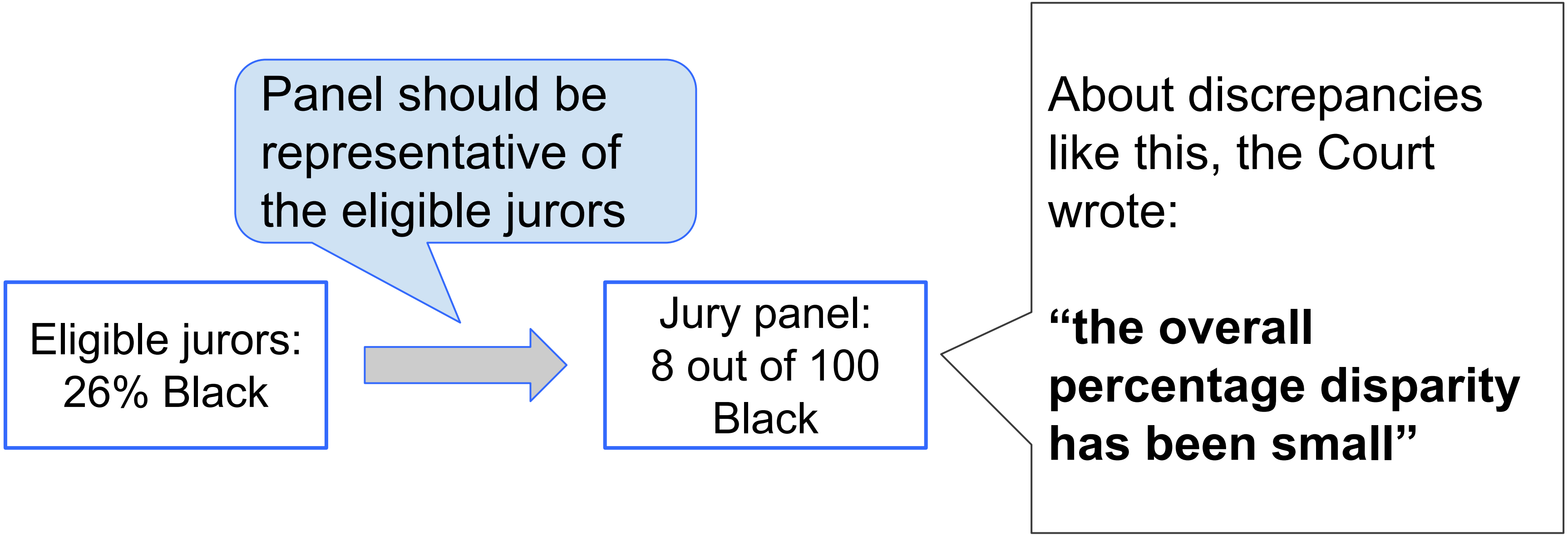
Eligible jurors: 26% Black

Jury panel: 8 out of 100 Black

Jury: 0 Black

Supreme Court Ruling, 1965

- The Court denied Robert Swain’s appeal.



Panel should be representative of the eligible jurors

Eligible jurors:
26% Black

Jury panel:
8 out of 100
Black

About discrepancies like this, the Court wrote:

“the overall percentage disparity has been small”

Discussion Question

- **Court’s view:** 8/100 is less than 26%, but not different enough to show Black panelists were systematically excluded
- **Question:** Would 8/100 be a realistic outcome if the jury panel selection process were truly unbiased?

Sampling from a Distribution

- Sample at random from a categorical distribution

```
sample_proportions(sample_size, pop_distribution)
```

- Samples at random from the population
  - Returns an array containing the empirical distribution of the categories in the sample

(Demo)

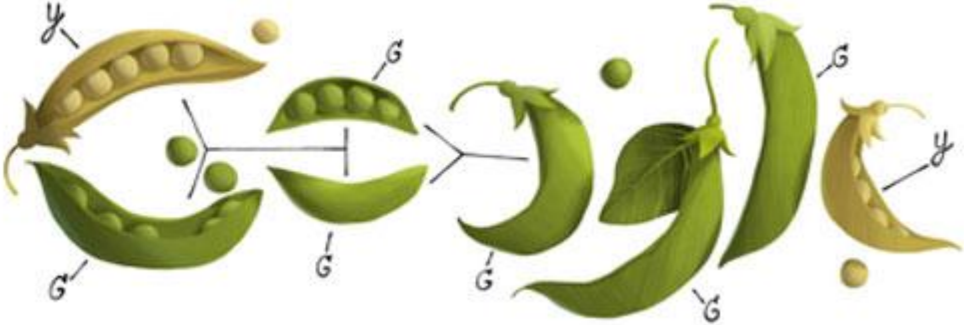
Statistical Bias

- Evidence provided by Robert Swain:

  “only 10 to 15% of … jury panels drawn from the jury box since 1953 have been [Black], there having been only one case in which the percentage was as high as 23%”
- Percent of Black panelists was always lower than expected under random sampling
- *Bias*: when errors are systematically in one direction

A Genetic Model

Gregor Mendel, 1822-1884



A Model

- Pea plants of a particular kind
- Each one has either purple flowers or white flowers

- Mendel’s model:
    - Each plant is purple-flowering with chance 75%,
    - regardless of the colors of the other plants

- Question:
    - Is the model good, or not?

Choosing a Statistic

- Take a sample, see what percent are purple-flowering
- If that percent is much larger or much smaller than 75, that is evidence against the model
- ***Distance*** from 75 is the key

- Statistic:

  | sample percent of purple-flowering plants - 75 |

- If the statistic is large, that is evidence against the model

(Demo)

Two Viewpoints

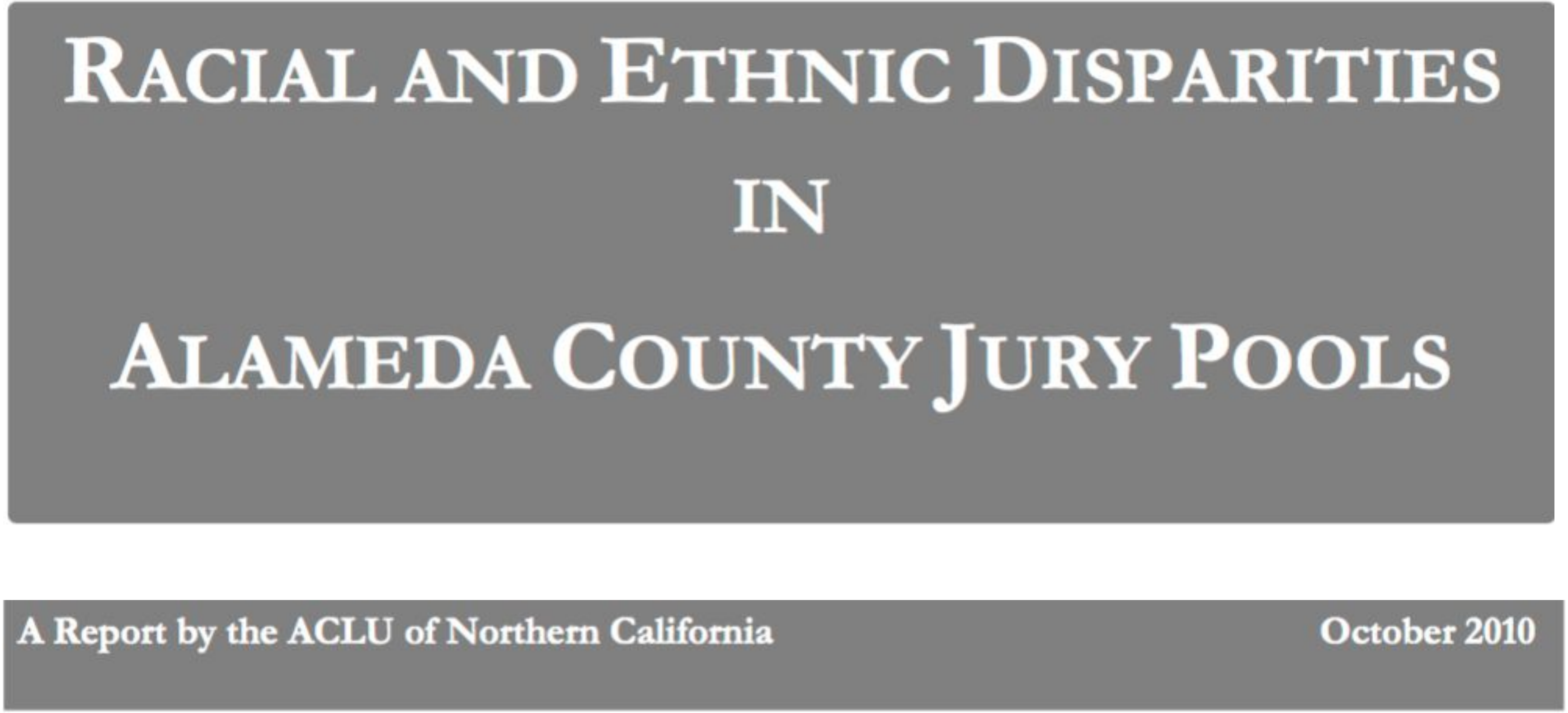
Model and Alternative

- **Jury selection:**
  - **Model:** The people on the jury panels were selected at random from the eligible population
  - **Alternative viewpoint:** No, they were biased against black men
- **Genetics:**
  - **Model:** Each plant has a 75% chance of having purple flowers
  - **Alternative viewpoint:** No, it doesn’t

Steps in Assessing a Model

- **Choose a statistic** to measure 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

Next time



RACIAL AND ETHNIC DISPARITIES

IN

ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California

October 2010