DATA 8

Spring 2020

Lecture 16

Models

Weekly Goals

- Monday
  - 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

Announcements

- Project 1 due today!
- Midterm on March 13th, 7PM
  - Scope: up to and including **A/B testing**
  - Review material on Piazza
  - **Review on 03/11, 6-8PM in 10 Evans & 2050 VLSB**
  - Fill out conflict form by **tonight**
- Homework 4 grades out today
- Homework 6 is out today, due next Thursday

Fun fact: Venus is visible in the sky tonight ✨

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

(Demo)

Inference

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 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.

Jury Selection

Swain vs. Alabama, 1965

- Talladega County, Alabama
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Only men 21 years or older were allowed to serve
- 26% of this population were black
- Swain’s jury panel consisted of 100 men
- 8 men on the panel were black

Supreme Court Ruling [in English]

- About disparities between the percentages in the eligible population and the jury panel, the Supreme Court wrote:

"... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes"

- The Supreme Court denied Robert Swain's appeal

Supreme Court Ruling [in Data]

- **Paraphrase:** 8/100 is less than 26%, but not different enough to show Black men were systematically excluded
- **Question:** is 8/100 a realistic outcome if the jury panel selection process were truly unbiased?

Sampling from a Distribution

- Sample at random from a categorical distribution

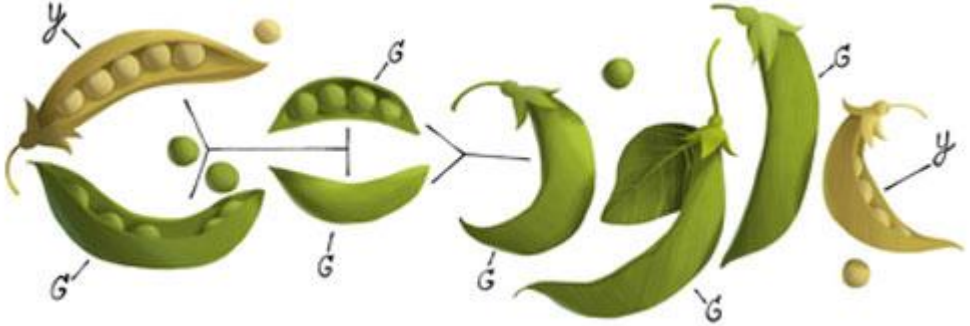
```
sample_proportions(sample_size, pop_distribution)
```

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

(Demo)

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

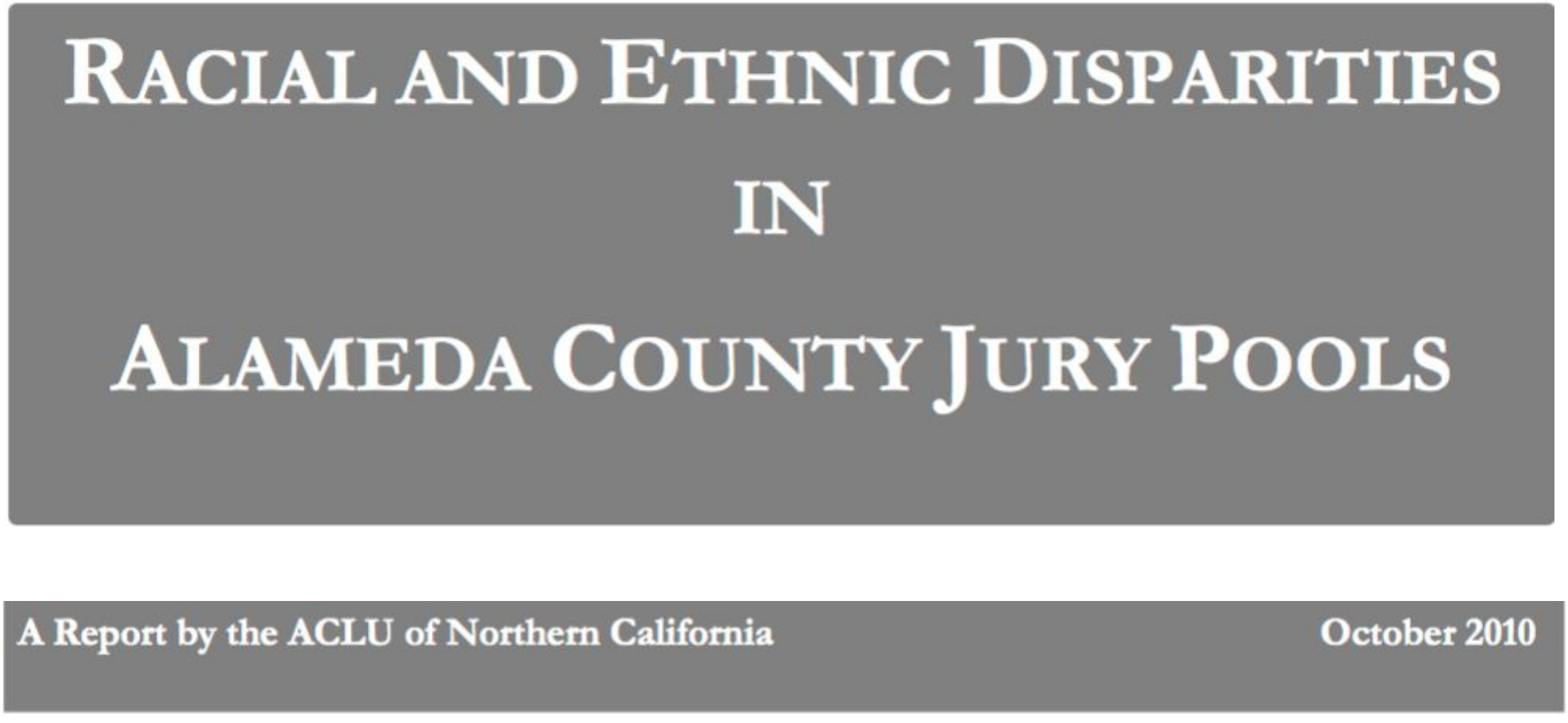
- **Jury selection:**
  - **Model:** The people on the jury panels were selected at random from the eligible population
  - **Alternative viewpoint:** No, they weren't

- **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