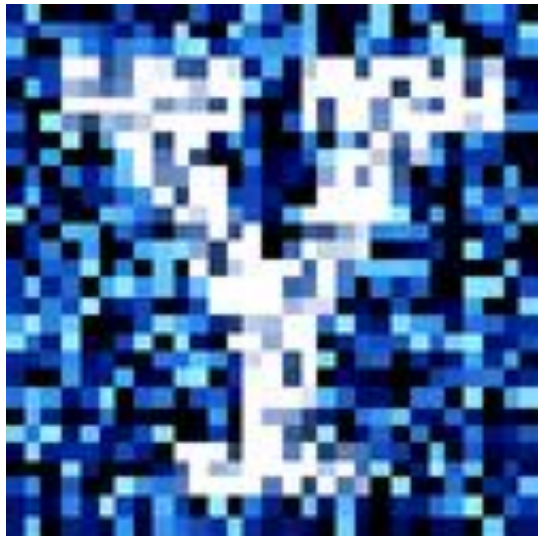
YData: Introduction to Data Science



Lecture 17: Comparing distributions

Overview

Review and continuation of models and model assessment

Example 1: Gregor Mendel and pea flower color

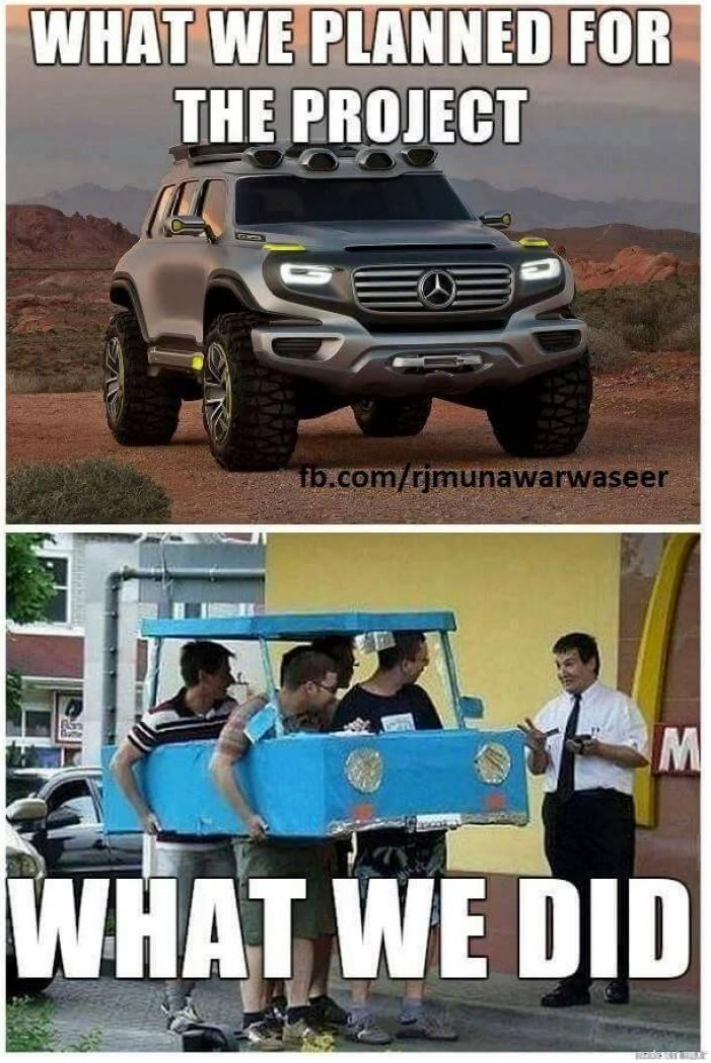
Example 2: Jury selection in Almeda county

Announcements

Project 1 due tonight at 11pm

- For people working together in pairs, only one person should submit the project.
- Be sure to mark both people's names on Gradescope.
  - Instructions on part of the project are incorrect. Only make one submission with both partners' names

Practice 5 and homework 6 have been added to the class calendar page



Review: Models

A model is a description of some underlying phenomenon

- Models are based a set of assumptions about how a particular phenomenon works

In order to assess if a model is capturing key aspects of a phenomenon of interest, we need to compare the predictions of the model to actual data

Model and alternative

Often we are evaluating between a model's prediction and an alternative viewpoint

Example of jury selection:

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

Steps in assessing a model

1. Create a statistic that will help you decide whether the model is consistent with observed data

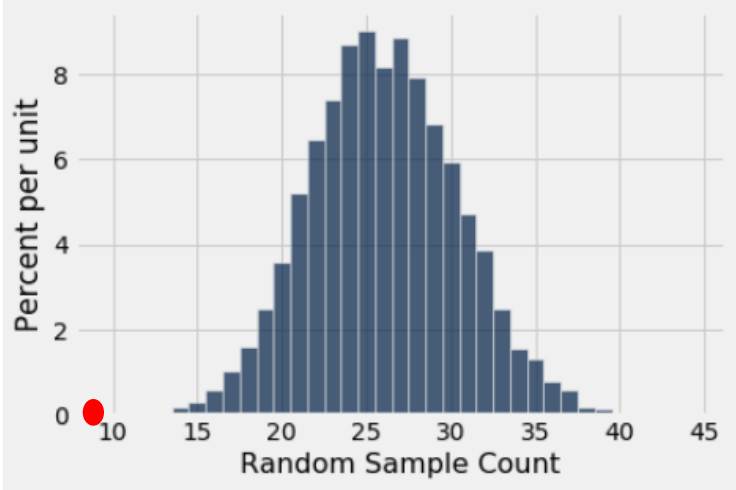
2. Simulate the statistic under the assumptions of the model.

3. Compare the data to the model's predictions

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

    b. Compute the observed statistic from the sample in the study.

    c. If the observed statistic is far from the histogram, that is evidence against the model



Discussion questions: which statistic to choose?

Data: the results of 400 tosses of a coin

Choose a statistic that will help you decide between the two viewpoints

Scenario 1:

- "This coin is fair."
- "No, it's biased towards tails."

Answer:

- Small values of the number of heads suggest "biased towards tails"
- Statistic: number of heads

Discussion questions: which statistic to choose?

Data: the results of 400 tosses of a coin

Choose a statistic that will help you decide between the two viewpoints

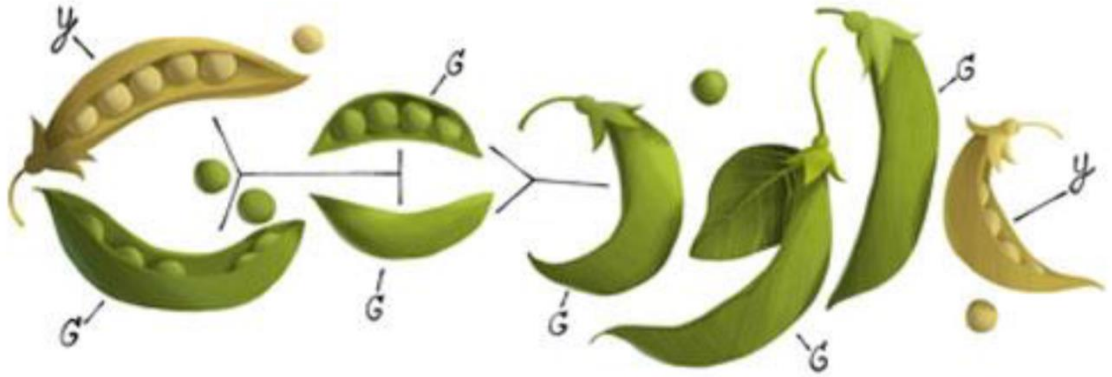
Scenario 2:

- "This coin is fair."
- "No, it's not fair." (as opposed to being just "biased toward tails")

Answer:

- Very large or very small values of the number of heads suggest "not fair."
  - The distance between number of heads and 200 is the key
- Statistic: | number of heads - 200 |
  - Large values of the statistic suggest the coin "not fair"

Example: Gregor Mendel, 1822-1884



A model of pea flowers

Mendel examined the flowers of the common pea plant *Pisum sativum*

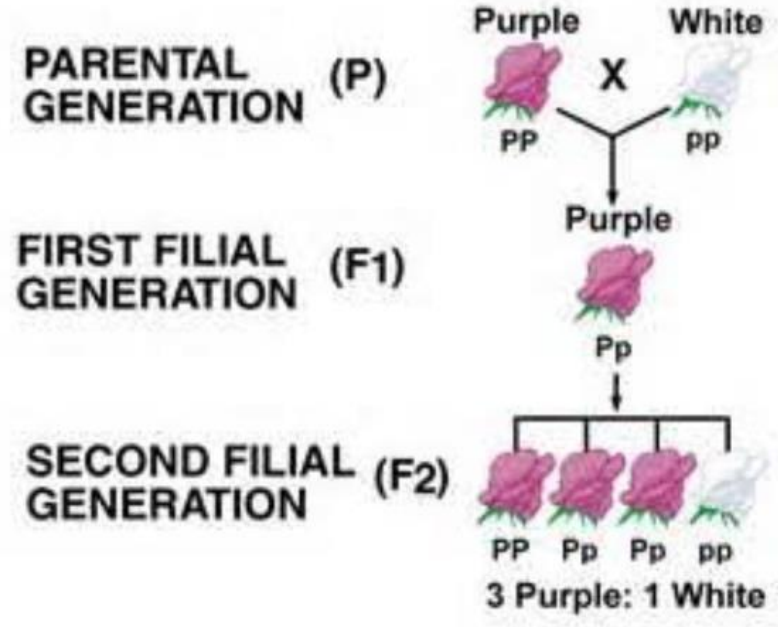
- Each plant has either purple flowers or white flowers

Mendel's model:

- Each plant is purple-flowering with chance 75%

Mendel grew 929 plants and 705 out of them had purple flowers

- 705/929 = 0.76



Choosing a statistic

A statistic we can use to assess our model:

| sample percent of purple-flowering plants - 75 |

We can simulate many statistics from random samples of 929 flowers where 75% are purple

The observed statistics from Mendel's data is: | 100 * 705/929 - 75 |

If the observed statistics from Mendel's data looks much different than the statistics generated from the model, we can reject the model

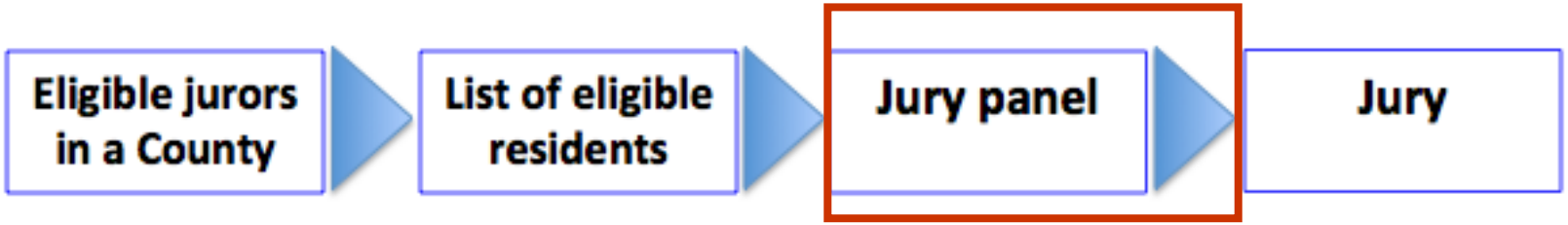
Let's explore this in Jupyter!

Comparing distributions

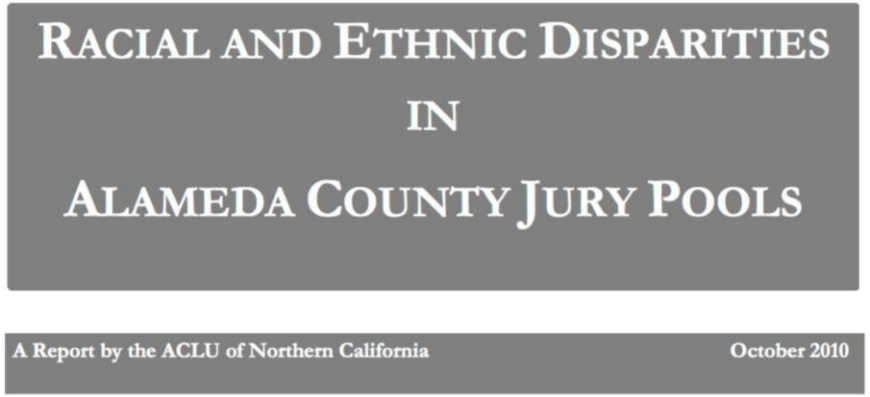
Jury selection in Alameda county

Section 197 of California's Code of Civil Procedure says:

" All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court."



In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report that concluded that certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.



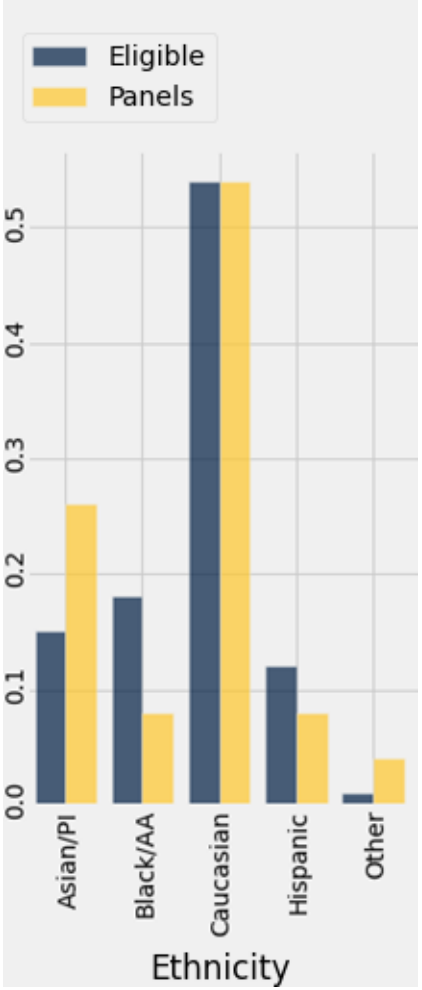
A new statistic to measure the distance between distributions

The ACLU compiled data on the composition of 1453 people who were on jury panels from in the years 2009 and 2010.

People on the panels are of multiple ethnicities

- Distribution of ethnicities is categorical

To see whether the distribution of ethnicities of the panels is close to that of the eligible jurors, we have to measure the distance between two categorical distributions

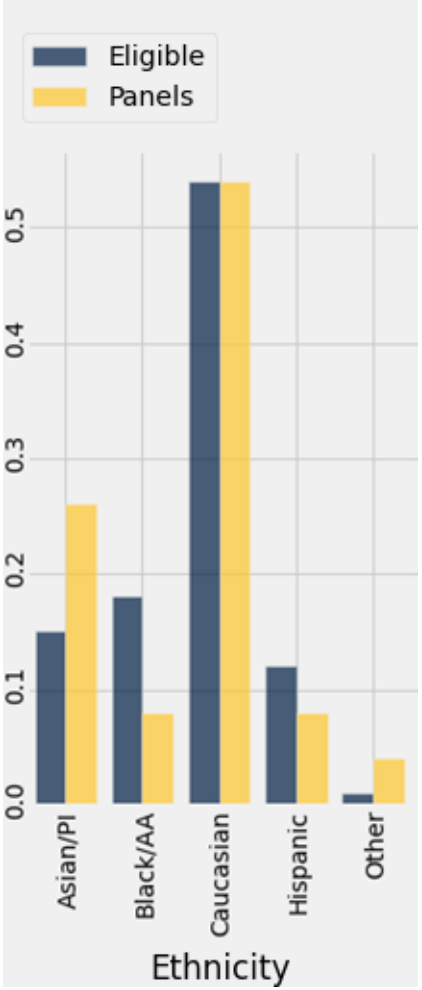


Total variation distance

Every statistic has a computational recipe

Total Variation Distance (TVD) statistic:

- For each category, compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum the values and then divide the sum by 2



Let's explore this in Jupyter!

Summary of the method

To assess whether a sample was drawn randomly from a known categorical distribution:

- Use TVD as the statistic because it measures the distance between categorical distributions
- Sample at random from the population and compute the TVD from the random sample; repeat numerous times
- Compare:
  - Empirical distribution of simulated TVDs
  - Actual TVD from the sample in the study