Testing Hypotheses

References:

Chapter 11 - Testing Hypotheses

https://www.inferentialthinking.com/chapters/intro

Overview

- Assessing Models
- Multiple Categories
- Decisions and Uncertainity
- Error Probabilities

Testing Hypotheses

- Data scientists are often faced with yes-no questions about the world.
  - Is drinking tea good for you?
  - Did water from the Broad Street pump cause cholera?
  - Did interventions in a set of villages improve their living conditions?

- Answers to above depends on the availability of relevant data

- An approach to such yes-no questions is to basing our conclusions on random samples and empirical distributions.

Assessing Models

- In data science, a "model" is a set of assumptions about data. Often, models include assumptions about chance processes used to generate data.
- Data scientists have to decide whether or not their models are good
- Need for a framework for testing hypotheses

U.S. Supreme Court, 1965: Swain vs. Alabama

- In the early 1960's, in Talladega County in Alabama, a black man called Robert Swain was convicted for a heinous crime and sentenced to death
- He appealed his sentence, citing among other factors the all-white jury.
- At the time, only men aged 21 or older were allowed to serve on juries in Talladega County.
- In his county, 26% of the eligible jurors were black, but there were only 8 black men among the 100 selected for the jury panel in Swain's trial.
- No black man was selected for the trial jury.

In 1965, the Supreme Court of the United States denied Swain's appeal. In its ruling, the Court wrote "... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes."

Model for selecting jury members

- Jury panels are supposed to be selected at random from the eligible population. Because 26% of the eligible population was black, 8 black men on a panel of 100 might seem low.
- But one view of the data – a model, in other words – is that the panel was selected at random and ended up with a small number of black men just due to chance. This model is consistent with what the Supreme Court wrote in its ruling
- The model specifies the details of a chance process. It says the data are like a random sample from a population in which 26% of the people are black. We are in a good position to assess this model, because:
  - We can simulate data based on the model. That is, we can simulate drawing at random from a population of whom 26% are black.
  - Our simulation will show what a panel *would* look like *if* it were selected at random.
  - We can then compare the results of the simulation with the composition of Robert Swain's panel.
  - If the results of our simulation are not consistent with the composition of Swain's panel, that will be evidence against the model of random selection.

Choosing a Statistic

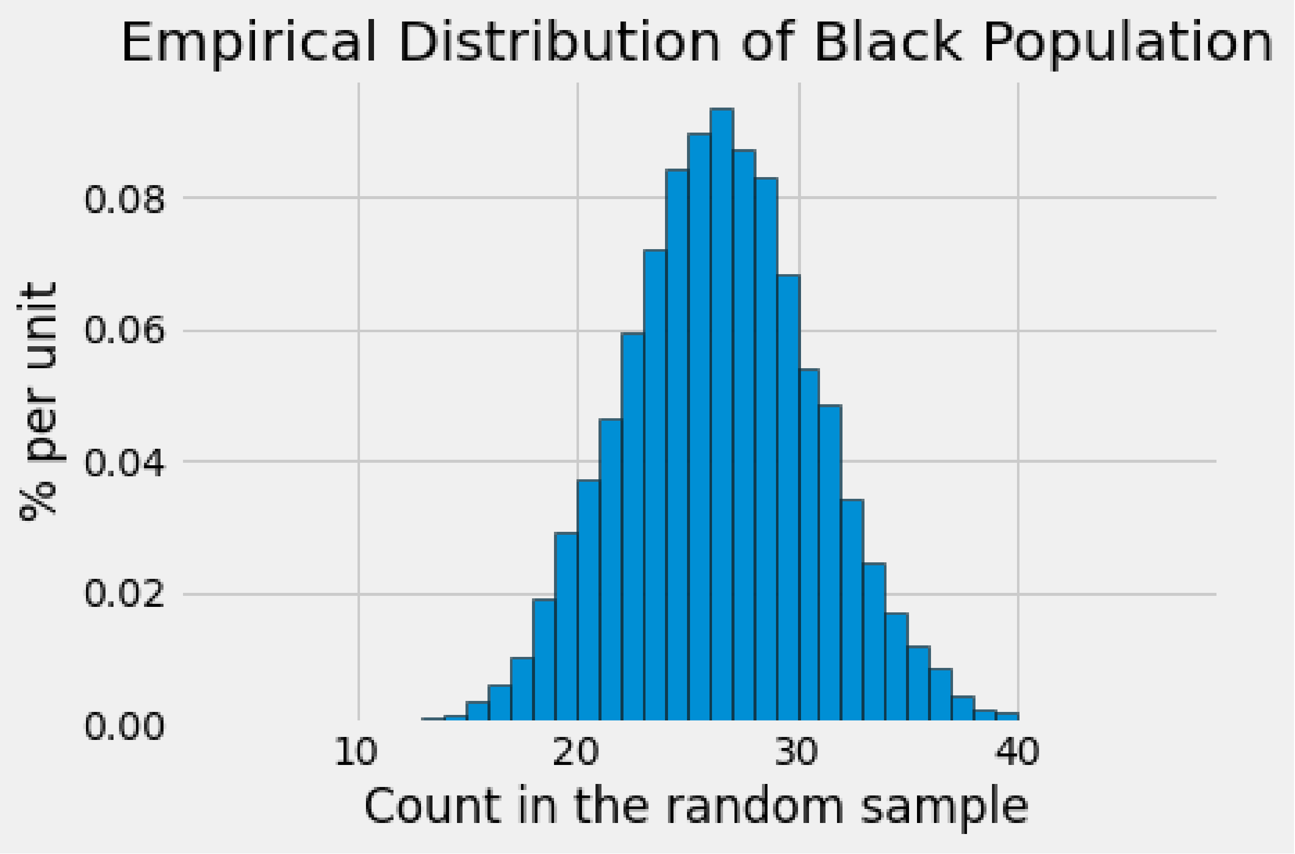
- First, we have to choose a statistic to simulate.
- The statistic has to be able to help us decide between the model and alternative views about the data.
- The model says the panel was drawn at random. The alternative viewpoint, suggested by Robert Swain's appeal, is that the panel was not drawn at random because it contained too few black men.
- A natural statistic, then, is the number of black men in our simulated sample of 100 men representing the panel.
- Small values of the statistic will favor the alternative viewpoint.

Predicting the Statistic Under the Model

- If the model were true, how big would the statistic typically be? To answer that, we have to start by working out the details of the simulation.
- First, to simulate one value of the statistic. For this, we have to sample 100 times at random from the population of eligible jurors and count the number of black men we get.
- To see how to use this, remember that according to our model, the panel is selected at random from a population of men among whom 26% were black and 74% were not. Thus the distribution of the two categories can be represented as the list [0.26, 0.74],
- Now let's sample at random 100 times from this distribution, and see what proportions of the two categories we get in our sample.
- Because there are 100 men in the sample, the number of men in each category is 100 times the proportion

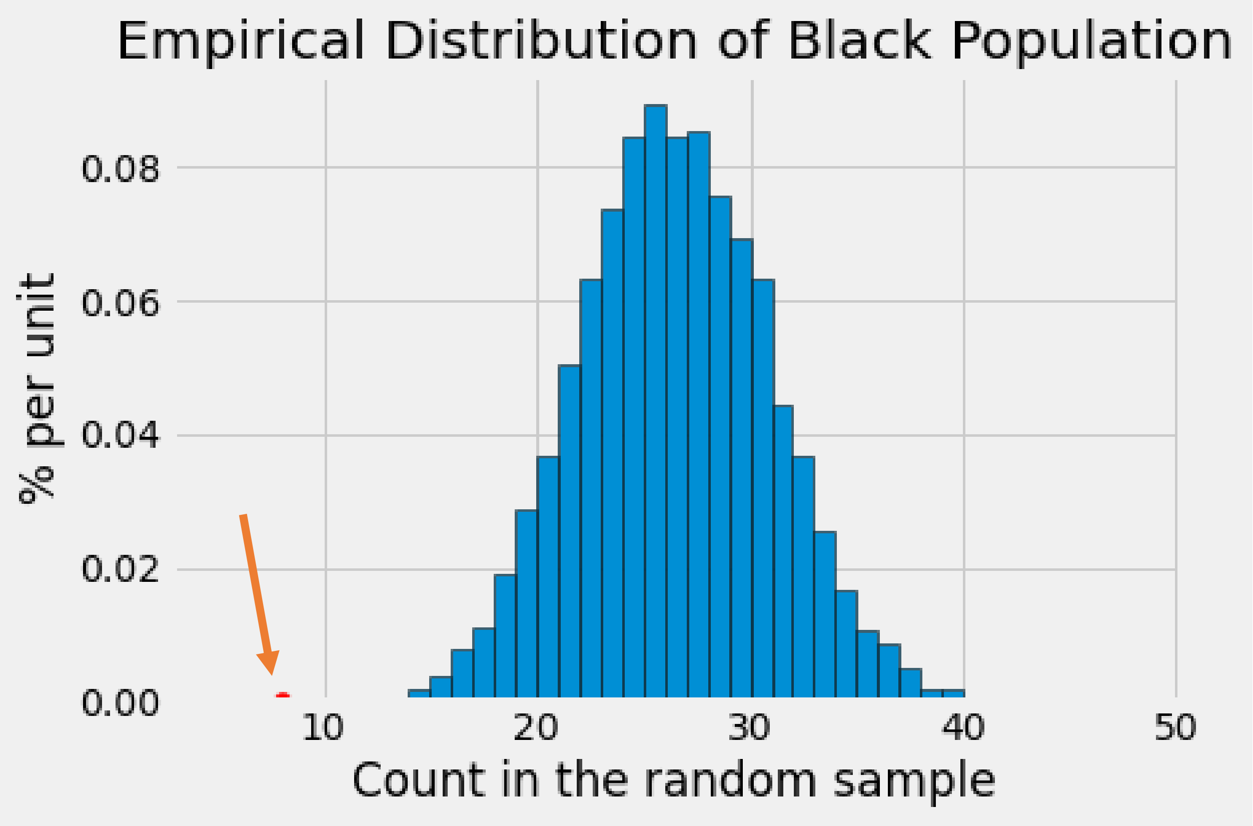
Running the Simulation and making the Prediction

- To get a sense of the variability without running the cell over and over, let's generate 10,000 simulated values of the count.
- The histogram tells us what the model of random selection predicts about our statistic, the count of black men in the sample.
- To generate each simulated count, we drew at 100 times at random from a population in which 26% were black. So, as you would expect, most of the simulated counts are around 26. They are not exactly 26: there is some variation. The counts range from about 10 to about 45.



Comparing the Prediction and the Data

- Though the simulated counts are quite varied, very few of them came out to be eight or less. The value eight is far out in the left hand tail of the histogram. It's the red dot on the horizontal axis of the histogram.
- When the data and a model are inconsistent, the model is hard to justify. After all, the data are real.
- The model is just a set of assumptions. When assumptions are at odds with reality, we have to question those assumptions



Mendel's Pea Flowers

- Gregor Mendel (1822-1884) was an Austrian monk who is widely recognized as the founder of the modern field of genetics.
- Mendel performed careful and large-scale experiments on plants to come up with fundamental laws of genetics.
- Many of his experiments were on varieties of pea plants. He formulated sets of assumptions about each variety; these were his models.
- He then tested the validity of his models by growing the plants and gathering data.
- Here we shall analyze the data from one such experiment to see if Mendel's model was good

Mendel's Model

- In a particular variety, each plant has either purple flowers or white. The color in each plant is unaffected by the colors in other plants. Mendel hypothesized that the plants should bear purple or white flowers at random, in the ratio 3:1.

For every plant, there is a 75% chance that it will have purple flowers, and a 25% chance that the flowers will be white, regardless of the colors in all the other plants.

Defining the Statistic

- To go about assessing Mendel's model, we can simulate plants under the assumptions of the model and see what it predicts. Then we will be able to compare the predictions with the data that Mendel recorded.
- Our goal is to see whether or not Mendel's model is good. We need to simulate a statistic that will help us make this decision.
- If the model is good, the percent of purple-flowering plants in the sample should be close to 75%. If the model is not good, the percent purple-flowering will be away from 75%. It may be higher, or lower; the direction doesn't matter.
- The key for us is the *distance* between 75% and the percent of purple-flowering plants in the sample. Big distances are evidence that the model isn't good.
- Our statistic, therefore, is the **distance between the sample percent and 75%**:

$$|\text{sample percent of purpleflowering plants} - 75|$$

Predicting the Statistic Under the Model

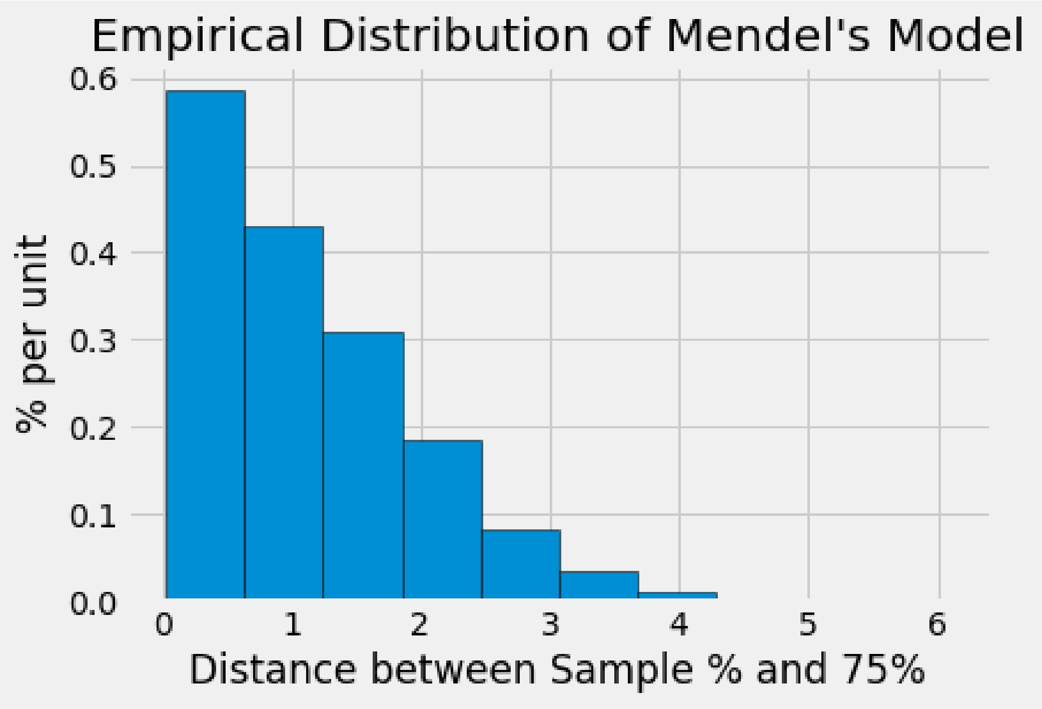
- We have to figure out how many times to sample. To do this, remember that we are going to compare our simulation with Mendel's plants. So we should simulate the same number of plants that he had.

- Mendel grew a lot of plants. There were 929 plants of the variety corresponding to this model. So we have to sample 929 times.
  - Sample 929 times at random from the distribution specified by the model and find the sample proportion in the purple-flowering category.
  - Multiply the proportion by 100 to get a percent.
  - Subtract 75 and take the absolute value of the difference.

- That's the statistic: the distance between the sample percent and 75.

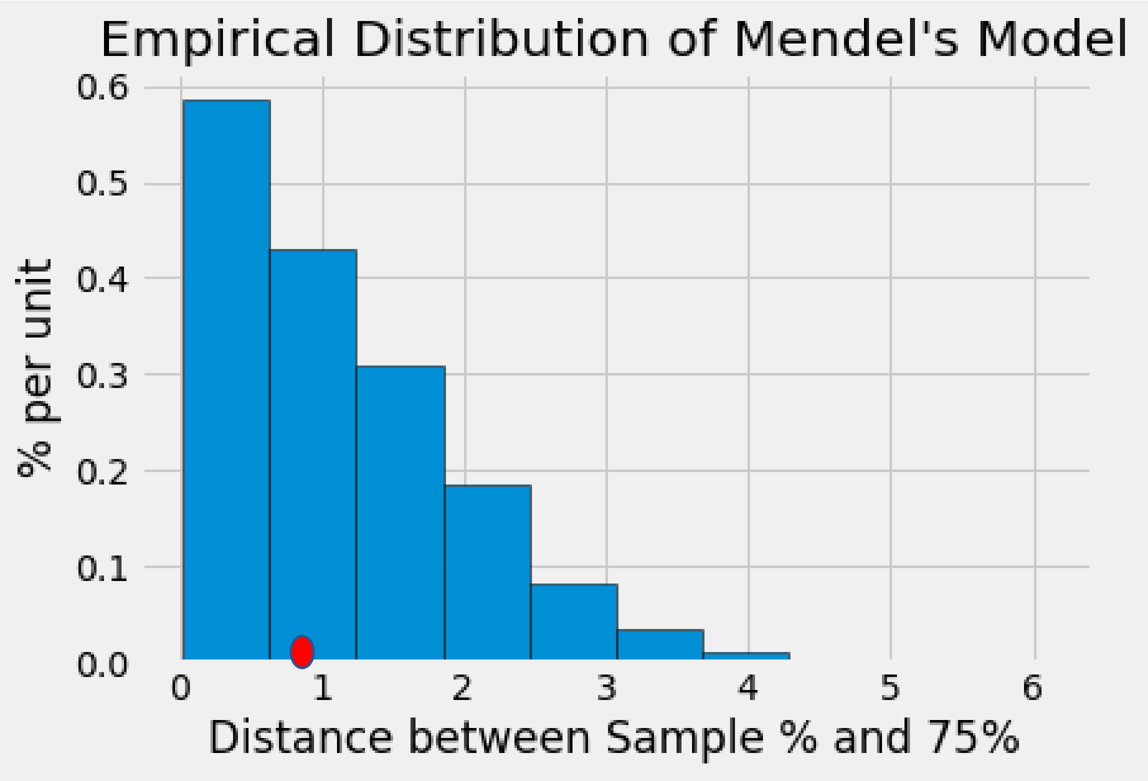
Running the Simulation and making the prediction

- We will generate 10,000 values of the distance.
- The empirical histogram of the simulated values shows the distribution of the distance as predicted by Mendel's model.
- Observing the horizontal axis, we see the typical values of the distance, as predicted by the model. They are rather small. For example, a high proportion of the distances are in the range 0 to 1, meaning that for a high proportion of the samples, the percent of purple-flowering plants is within 1% of 75%, that is, the sample percent is in the range 74% to 76%.



Comparing the Prediction and the Data

- To assess the model, we have to compare this prediction with the data. Mendel recorded the number of purple and white flowering plants.
- Among the 929 plants that he grew, 705 were purple flowering. That's just about 75.89%.
- So the observed value of our statistic – the distance between Mendel's sample percent and 75 is about 0.89:
- The observed statistic is like a typical distance predicted by the model. By this measure, the data are consistent with the histogram that we generated under the assumptions of Mendel's model.
- **This is evidence in favor of the model.**



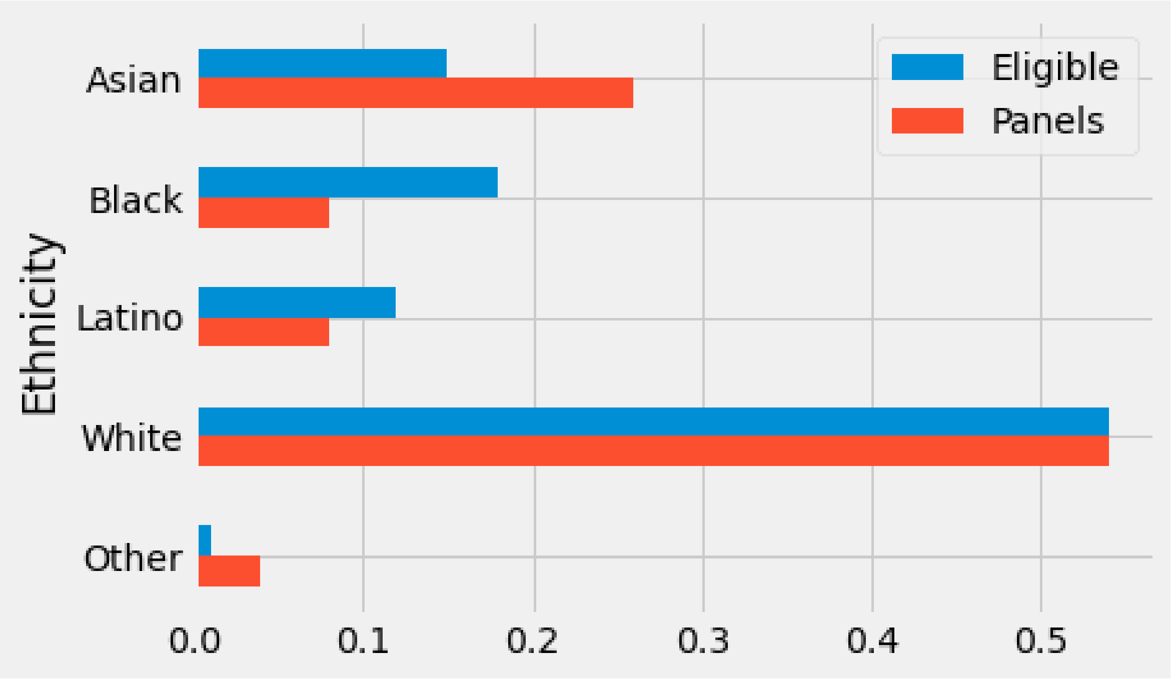
Multiple Categories - Jury Selection in Alameda County

- The process of assessment is the same as with 2 categories, the only difference being defining a new statistic to simulate.
- A jury panel is a group of people chosen to be prospective jurors; the final trial jury is selected from among them
- But the initial panel is supposed to resemble a random sample of the population of eligible jurors.
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- The report concluded that certain ethnic groups are underrepresented among jury panelists in Alameda County, and suggested some reforms of the process by which eligible jurors are assigned to panels.
- In this section, we will perform our own analysis of the data and examine some questions that arise as a result.

Composition of Panels in Alameda County

- The focus of the study by the ACLU of Northern California was the ethnic composition of jury panels in Alameda County.
- The ACLU compiled data on the ethnic composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010.
- In those panels, the total number of people who reported for jury service was 1,453.
- The ACLU gathered demographic data on all of these prospective jurors, and compared those data with the composition of all eligible jurors in the county.

Ethnicity	Eligible	Panels
Asian	0.15	0.26
Black	0.18	0.08
Latino	0.12	0.08
White	0.54	0.54
Other	0.01	0.04



Comparison with Panels Selected at Random

- What if we select a random sample of 1,453 people from the population of eligible jurors? Will the distribution of their ethnicities look like the distribution of the panels above?
- We can answer these questions by using sample proportions and augmenting the jury table with a column of the proportions in our sample.

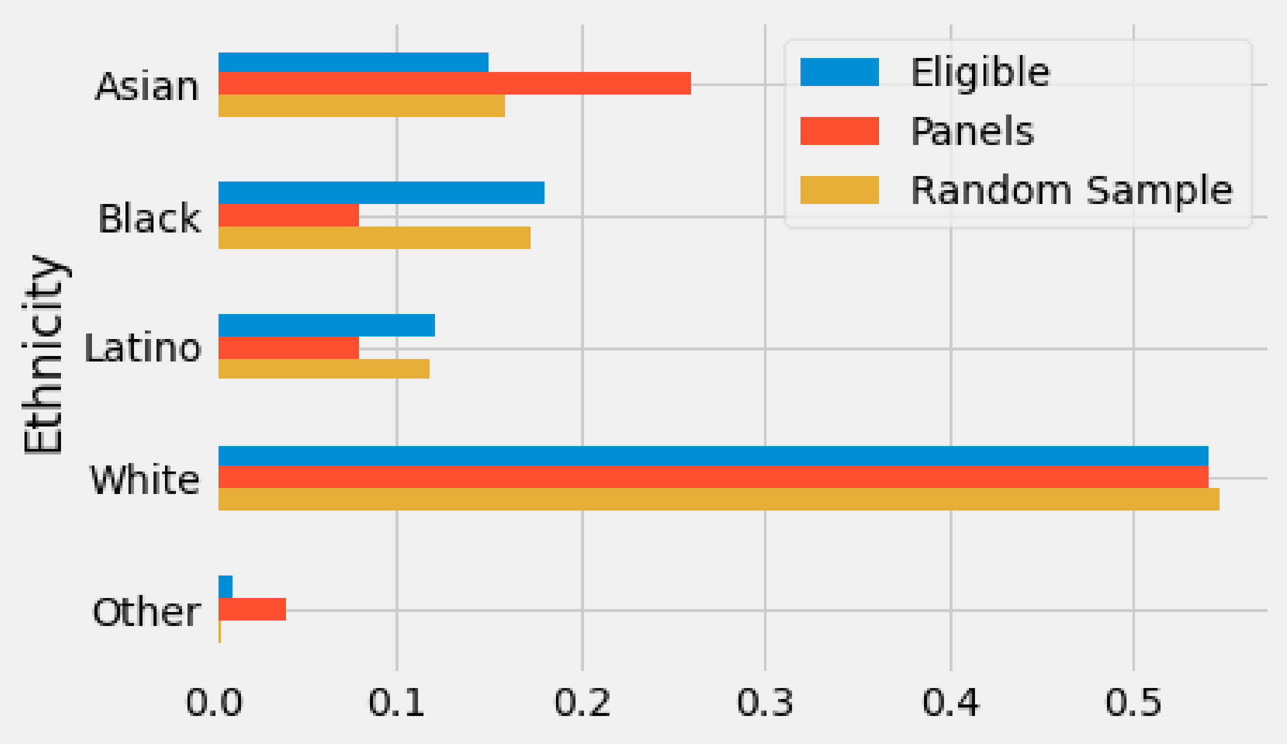
- Random samples of prospective jurors would be selected without replacement in real life.
- However, when the size of a sample is small relative to the size of the population, sampling without replacement resembles sampling with replacement; the proportions in the population don't change much between draws.
- The population of eligible jurors in Alameda County is over a million, and compared to that, a sample size of about 1500 is quite small.
- We will therefore sample with replacement

Comparison with Panels Selected at Random

- We sample at random 1453 times from the distribution of eligible jurors, and display the distribution of the random sample along with the distributions of the eligible jurors and the panel in the data.
- The distribution of the random sample is quite close to the distribution of the eligible population, unlike the distribution of the panels.
- The bar chart shows that the distribution of the random sample resembles the eligible population but the distribution of the panels does not.

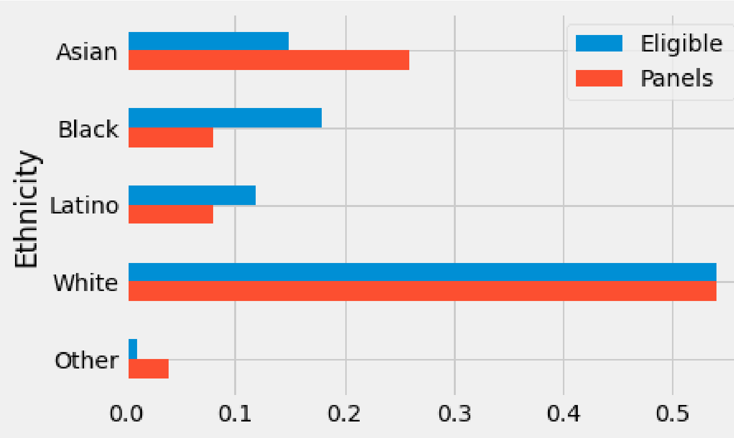
- To assess whether this observation is particular to one random sample or more general, we can simulate multiple panels under the model of random selection and see what the simulations predict.
- Cannot review thousands of bar charts!.

Ethnicity	Eligible	Panels	Random Sample
Asian	0.15	0.26	0.15898
Black	0.18	0.08	0.17275
Latino	0.12	0.08	0.11700
White	0.54	0.54	0.54646
Other	0.01	0.04	0.00482



A New Statistic: The Distance between Two Distributions

- We know how to measure how different two numbers are – if the numbers are x and y , the distance between them is |x−y| .
- Now we have to quantify the distance between two distributions.
- For example, we have to measure the distance between the blue and red distributions below
- We will compute a quantity called the *total variation distance* between two distributions.
- The calculation is as an extension of the calculation of the distance between two numbers



TVD: Total Variational Distance

- To compute the total variation distance, we first take the difference between the two proportions in each category.
- As sum of its entries is 0 (in this case), we drop the negative signs and then add all the entries and divide by 2, giving a total of 0.14
- This quantity 0.14 is the *total variation distance* (TVD) between the distribution of ethnicities in the eligible juror population and the distribution in the panels.

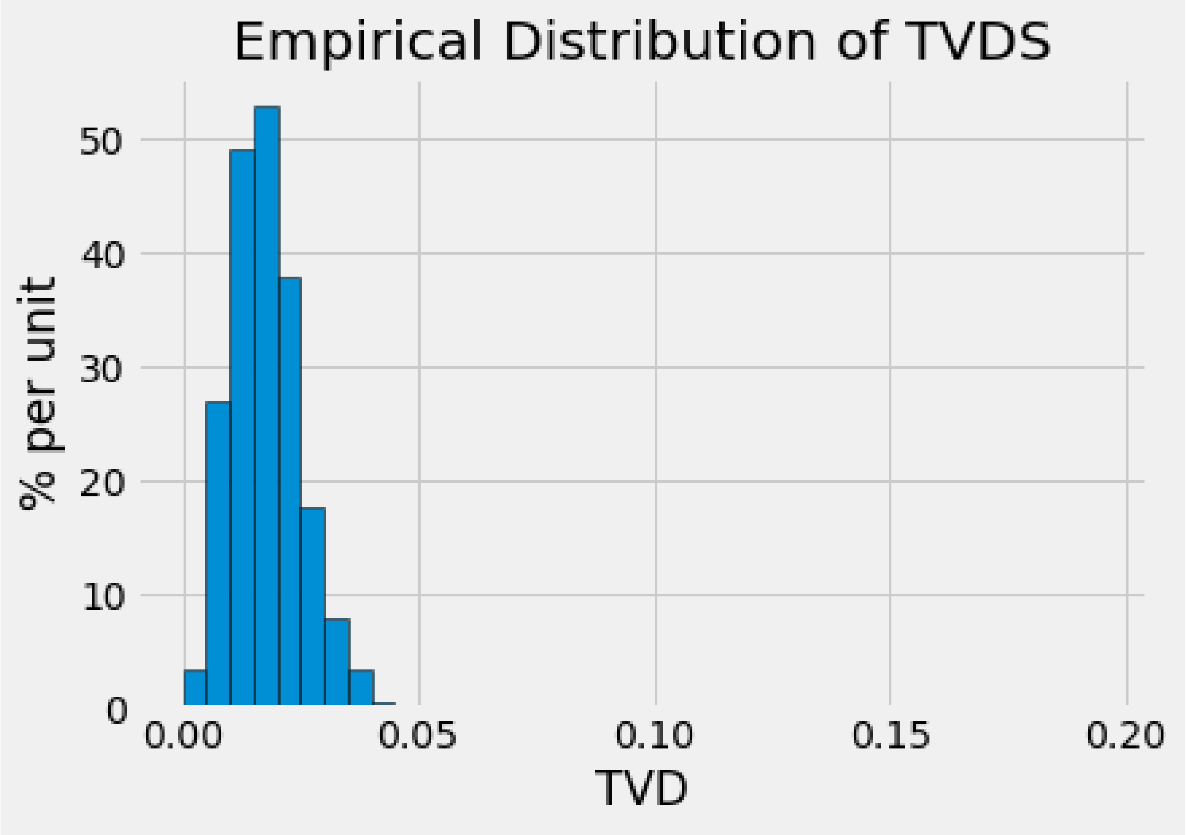
Ethnicity	Eligible	Panels	Difference
Asian	0.15	0.26	0.11
Black	0.18	0.08	-0.10
Latino	0.12	0.08	-0.04
White	0.54	0.54	0
Other	0.01	0.04	0.03

Simulating One Value of the Statistic

- We will use the total variation distance between distributions as the statistic to simulate.
- It will help us decide whether the model of random selection is good, because large values of the distance will be evidence against the model.
- **The observed value of our statistic is 0.14**
- Since we are going to be computing total variation distance repeatedly, we will write a function to compute it.
- This function will help us calculate our statistic in each repetition of the simulation.
- For a single run, that the distance is quite a bit smaller than 0.14, the distance between the distribution of the panels and the eligible jurors.

Predicting the Statistic Under the Model of Random Selection

- The total variation distance between the distributions of the random sample and the eligible jurors is the statistic that we are using to measure the distance between the two distributions.
- By repeating the process of sampling, we can see how much the statistic varies across different random samples



- The panels in the study, however, were not quite so similar to the eligible population.
- The total variation distance between the panels and the population was 0.14, which is far out in the tail of the histogram above.

Assessing the Model of Random Selection

- The data in the panels is not consistent with the predicted values of the statistic based on the model of random selection.
- So this analysis supports the ACLU's calculation that the panels were not representative of the distribution provided for the eligible jurors.
- We have developed a powerful technique that helps decide whether one distribution looks like a random sample from another.
- But data science is about more than techniques. In particular, data science always involves a thoughtful examination of how the data were gathered.
- While the panels do not look like a random sample from the distribution provided for eligible jurors – questions about the nature of the data that was used to prepare the sample prevent us from concluding anything broader in this specific case.

Summary

- Assessing if a sample was drawn randomly from a known population:
    - Decide on a statistic that measures the distance between distributions
    - Compute the statistic from the sample; that is, the distance between distributions of sample and known population
    - Sample at random and from the population and compute the statistic from the random sample; repeat numerous times
    - Compare