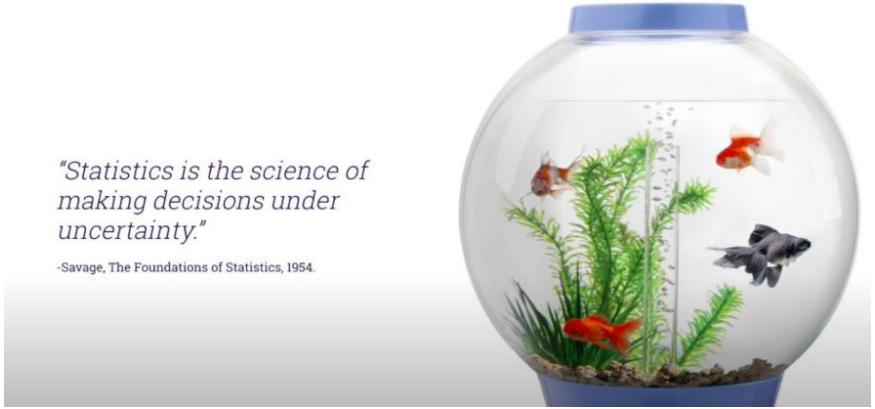


Hypothesis Testing

LECTURE 18 & 19

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 7th Week At A Glance

Mon 2/26	Tues 2/27	Wed 2/28	Thurs 2/29	Fri 3/1
Attend & Participate in Class	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class	HW 6 Due 11:59pm	Attend & Participate in Class NO QUIZ! HW 7 Released

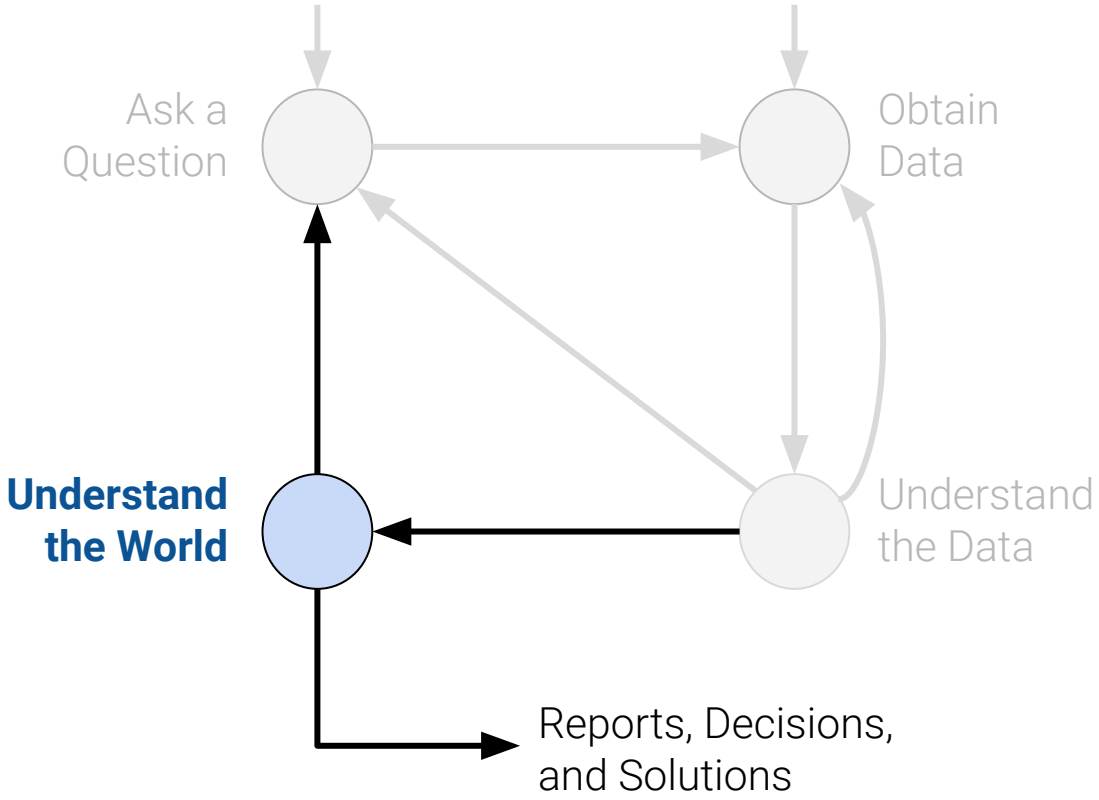
Today's Roadmap

CSCI 3022

- Day 1:
- Finish Lesson 17:
  - Central Limit Theorem
- Intro to Hypothesis Testing
  - Supreme Court Case Example
- P-Values
  - Definition/Visualization
  - Empirical vs Theoretical
- Day 2:
  - Comparing Multiple Distributions
    - Total Variation Distance
  - Hypothesis Testing with 1 Category
  - Hypothesis Testing with Sample of Numerical Data

The Data Science Lifecycle: Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



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

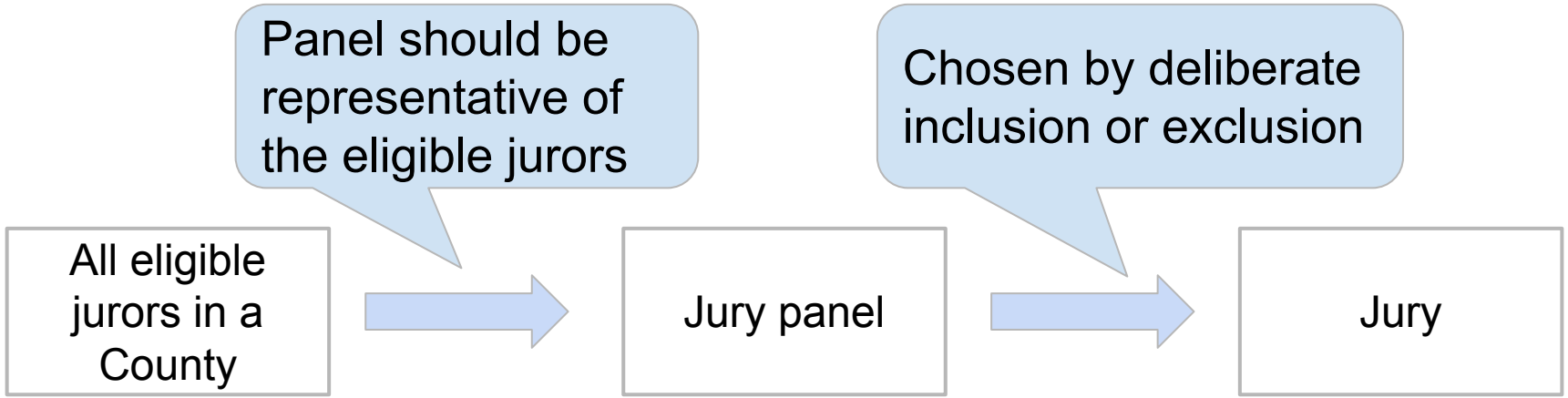
We will study a U.S. Supreme Court case in the 1960s

- 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”



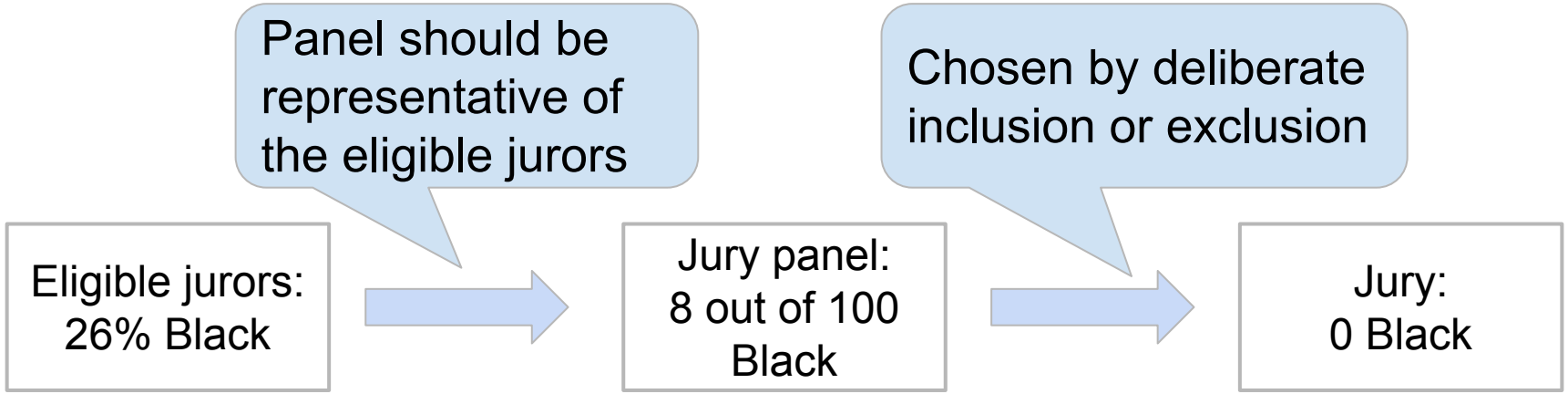
Supreme Court Case

CSCI 3022

- Day 1:
- Finish Lesson 17:
  - Central Limit Theorem
- Intro to Hypothesis Testing
  - **Supreme Court Case Example**
- P-Values
  - Definition/Visualization
  - Empirical vs Theoretical
- Day 2:
  - Comparing Multiple Distributions
    - Total Variation Distance
  - Hypothesis Testing with 1 Category
  - Hypothesis Testing with Sample of Numerical Data

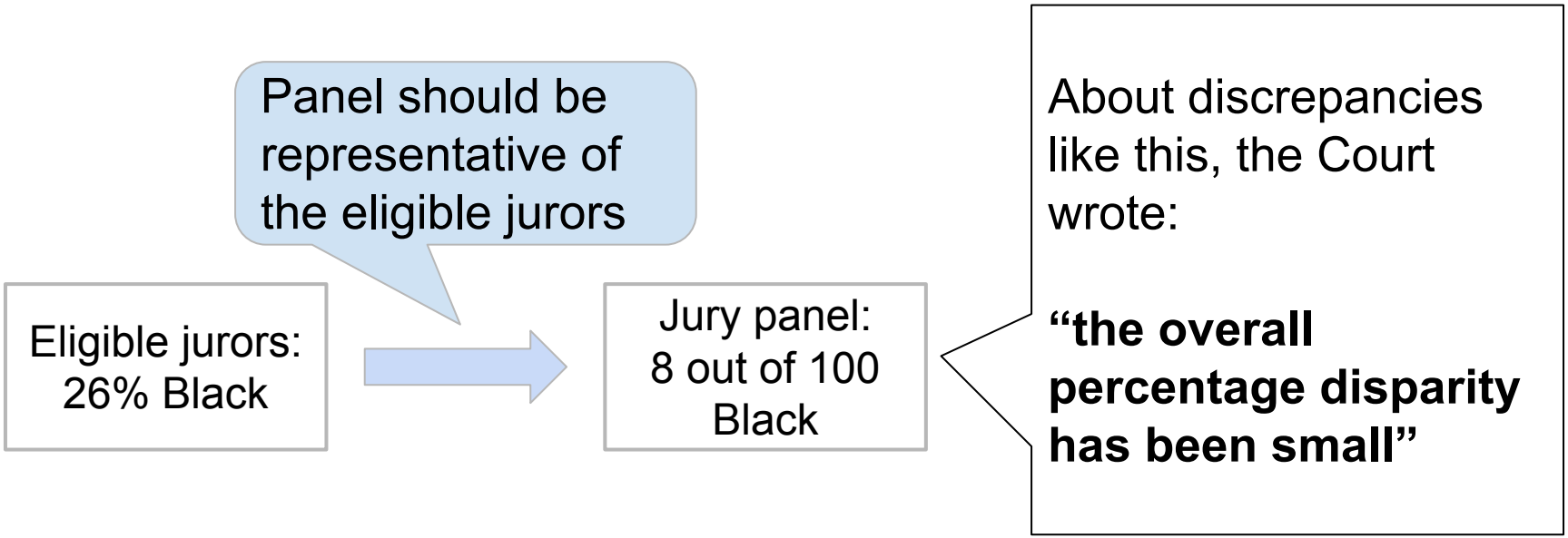
Robert Swain's Case

- 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

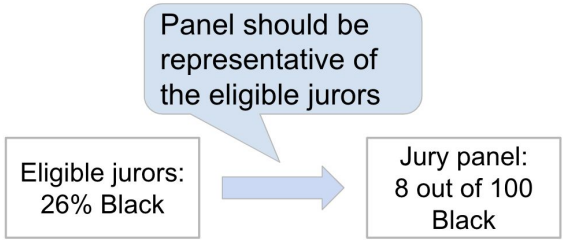


Supreme Court Ruling, 1965

- The Court denied Robert Swain's appeal.



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?

Statistical Testing

To begin, you need:

Default action
(Frequentist)

OR

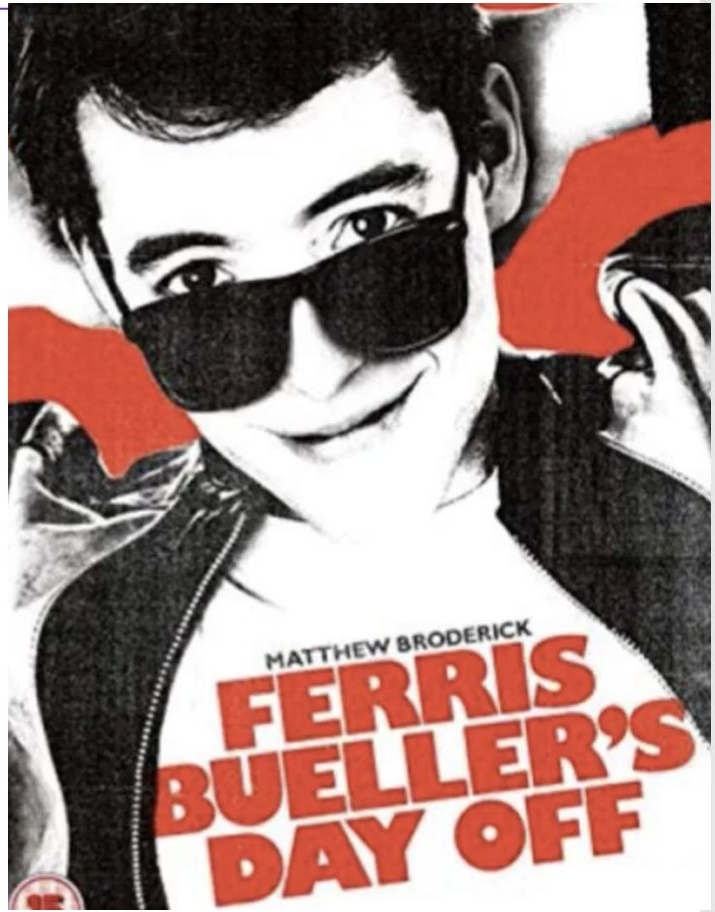
Prior opinion
(Bayesian)

Statistical Testing

Skip ALL of these statistical tests if:

1). You can answer with certainty

2). You have no prior opinion or default action.



Testing Hypotheses

- A test chooses between two views of how data was generated
- The views are called **hypotheses**

The method only works if we can simulate data (or calculate probabilities theoretically) under one of the hypotheses.

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model – “under the null hypothesis”
- **Alternative hypothesis**
  - A different view about the origin of the data

Statistical Testing

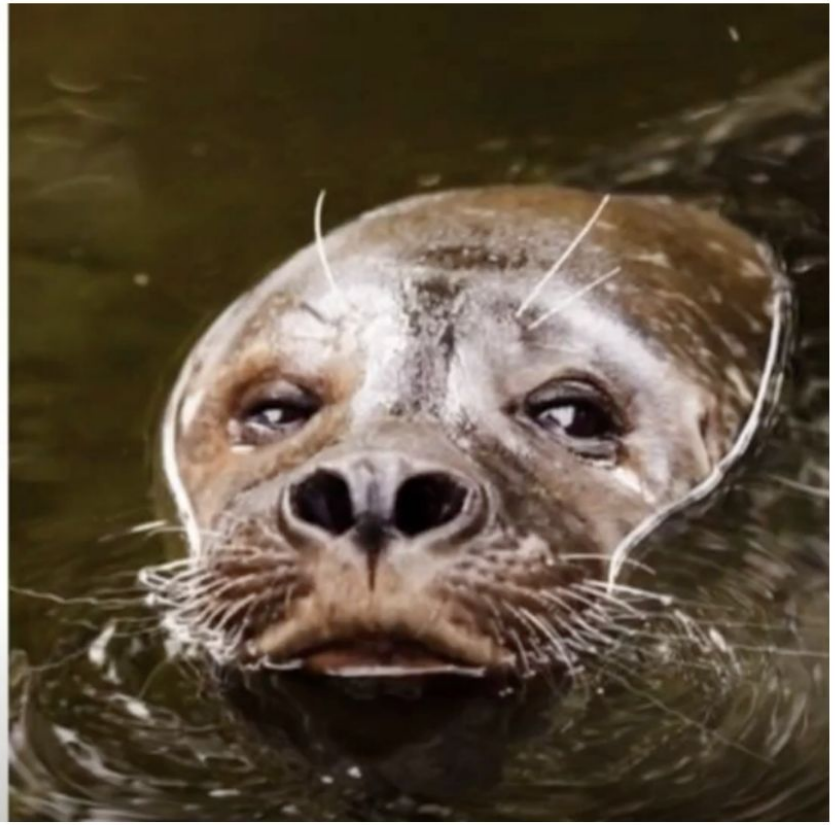
You should be happy to follow the default course of action as long as:

You haven't got any data
OR
You know very little
OR
Null Hypothesis is true for sure

Statistical Testing

In order to want to change your action from the default:

You need to be convinced (with data!) that the Alternative Hypothesis is true.



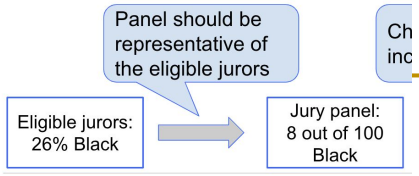
Testing Hypotheses

- A test chooses between two views of how data was generated
- The views are called **hypotheses**

  Ex: Robert Swain **Jury selection Example:**

  - **"Null" Hypothesis:** The people on the jury panels were selected at random from the eligible population. Any difference we see between the population demographics and the jury panel is due to chance
  - **"Alternative" Hypothesis:** No, they were biased against black people

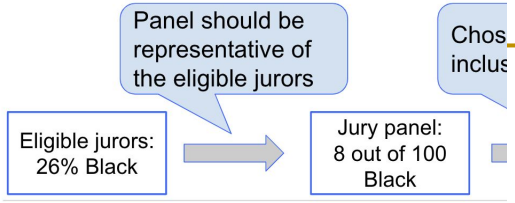
Example: Robert Swain's Case

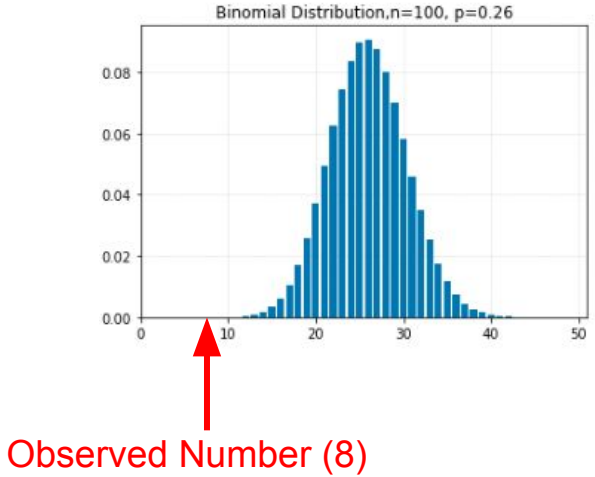


- **State the Null and Alternative Hypotheses**
  - Null Hypothesis: The people on the jury panels were selected at random from the eligible population where 26% of eligible jurors are Black. Any difference we see between the population racial demographics and the actual jury panel racial demographics is due to chance.
  - Mathematical Model of Null Hypothesis: Let X be the number of Black people out of 100 on the jury panel assuming they were selected at random from the population where 26% of eligible jurors are Black. Then X~
  - Alternative Hypothesis: The jury panel selection was biased against Black people
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
  - Test Statistic: _____
- **Simulate the test statistic distribution (or calculate directly when possible)** under the null assumption. Draw the distribution (or density histogram of simulated values).
- **Collect data and compare** the data to the null hypothesis predictions:
  - Compute the observed statistic from the real sample: _____
- If the observed statistic is in the tail* of the null distribution, we reject the null hypothesis.

Prediction Under the Null Hypothesis

- Simulate data (or when possible calculate probabilities theoretically) under the null hypothesis, draw the probability distribution (or empirical distribution if simulating values).





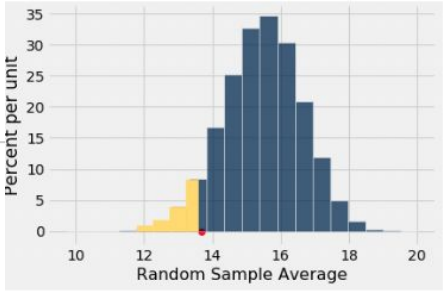
- This displays the **distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are **(if the null hypothesis is true)**

p-values

CSCI 3022

- Day 1:
- Finish Lesson 17:
  - Central Limit Theorem
- Intro to Hypothesis Testing
  - Supreme Court Case Example
- **P-Values**
  - Definition/Visualization
  - Empirical vs Theoretical
- Day 2:
  - Comparing Multiple Distributions
    - Total Variation Distance
  - Hypothesis Testing with 1 Category
  - Hypothesis Testing with Sample of Numerical Data

Definition of the *P*-value



Distribution of the test statistic under the null hypothesis

The red dot is the observed statistic.

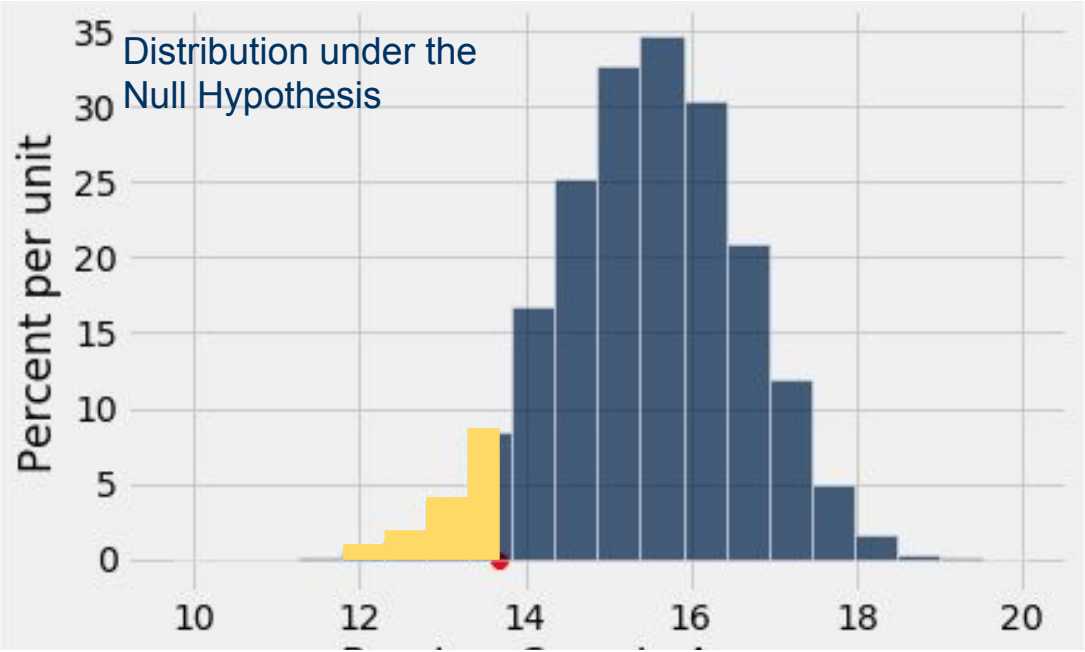
The *P*-value is the chance,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

Formal name: **observed significance level**

The p-Value as an Area

- Empirical distribution of the test statistic **under the null hypothesis**.
- Red dot denotes the observed statistic.
- Yellow area denotes the tail probability (p-value).



Analogy for p-values

Demo

https://www.youtube.com/watch?v=9jW9G8MO4PQ

Theoretical p-value vs Empirical p-value

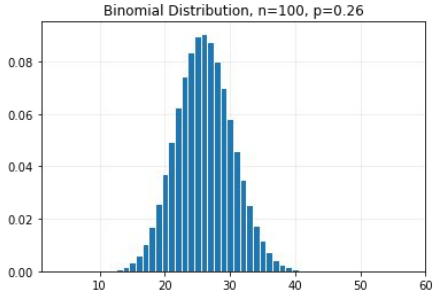
Recall: any random variable has a distribution:

<u>Probability (aka Population or Theoretical) Distribution</u> These are the distributions (pmf or pdf) of random variables or the distribution of some feature of some population.

```
k = np.arange(101)
p = special.comb(100, k)*(0.26**k)*(0.74**(100-k))


fig, ax = plt.subplots()

ax.bar(k, p, width=1, ec='white');
ax.set_axisbelow(True)
ax.grid(alpha=0.25)
plt.xlim(1,60)
plt.title("Binomial Distribution, n=100, p=0.26");
```



- <u>Empirical (aka Simulated or Sample) Distribution:</u> based on random samples (or simulations)
- Observations can be from **repetitions of an experiment or random samples from a population**
  - All observed values
  - The proportion of times each value appears

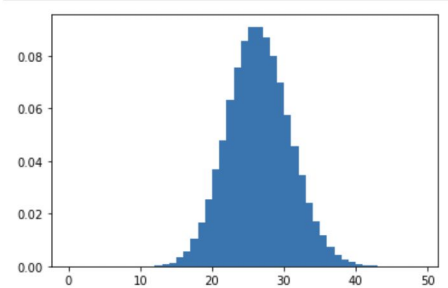
```
#Simulate one experiment
def heads_in_n_tosses(n=100):
    return sum(np.random.choice(["H","T"],size=n,p=[.26, .74]) == 'H')
```

```
# Repeat the experiment m times:
num_simulations = 50000;

outcomes=[]

for i in np.arange(num_simulations):
    outcomes = np.append(outcomes, heads_in_n_tosses())

plt.hist(outcomes,bins=np.arange(0,50),   density=True);
```



The p-value calculated using the **theoretical distribution** under the null hypothesis is called the **theoretical p-value (or just the p-value)**.

As your number of simulations increases, the empirical p-value will converge to the theoretical p-value

The p-value calculated using the **empirical distribution** under the null hypothesis is called the **empirical p-value**.

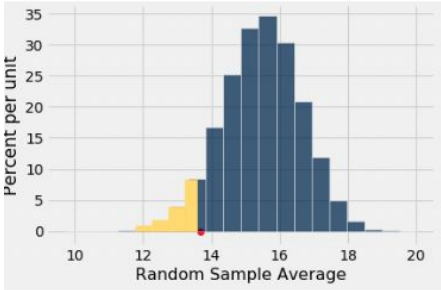
Conclusion of the Test

Determine whether observed test statistic is consistent null hypothesis:

- If p-value is less than your chosen significance level:
  - Reject the null hypothesis in favor of the alternative
  - Else: Fail to reject the null hypothesis

Conventions About Inconsistency

- **"Inconsistent with the null":** The observed test statistic is in the tail of the empirical distribution under the null hypothesis



- **"In the tail," first convention:**
  - The area in the tail is less than 5%
  - The result is "statistically significant"
- **"In the tail," second convention:**
  - The area in the tail is less than 1%
  - The result is "highly statistically significant"

Day 2: Roadmap

CSCI 3022

Hypothesis Testing:

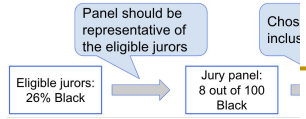
- Recap
- Comparing Multiple Distributions
  - Total Variation Distance
- Hypothesis Testing with 1 Category
- Hypothesis Testing with Simple Sample of Numerical Data

`

Course Logistics: 8th Week At A Glance

Mon 3/4	Tues 3/5	Wed 3/6	Thurs 3/7	Fri 3/8
Attend & Participate in Class	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class	HW 7 Due 11:59pm	Attend & Participate in Class QUIZ 5: Scope: L13-L16, nb 7, HW 6 HW 8 Released

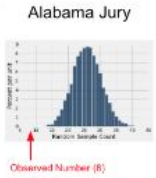
Recap: Robert Swain's Case



- Null Hypothesis: The people on the jury panels were selected at random from the eligible population where 26% are Black people (i.e. Binomial distribution, with n=100, p=0.26)
- Alternative Hypothesis: No, they were biased against Black people
- Test Statistic: Number of Black people chosen out of 100 assuming null hypothesis
- Observed test statistic: 8

Definition of the *P*-value

Formal name: **observed significance level**



The *P*-value is the chance,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

If the area in the tail is less than 1%
The result is "highly statistically significant"

$$X \sim Bin(100, 0.26)$$

$$P(X \le 8) = \sum_{k=0}^{8} \binom{100}{k} (0.26)^k (1-0.26)^{100-k}$$

$$= stats.binom.cdf(8, 100, 0.26) = .00000473$$

Conclusion: Our p-value = _____ which is less than _____ ,
so we _____null
and _____ the alternative and our result is "highly statistically significant"

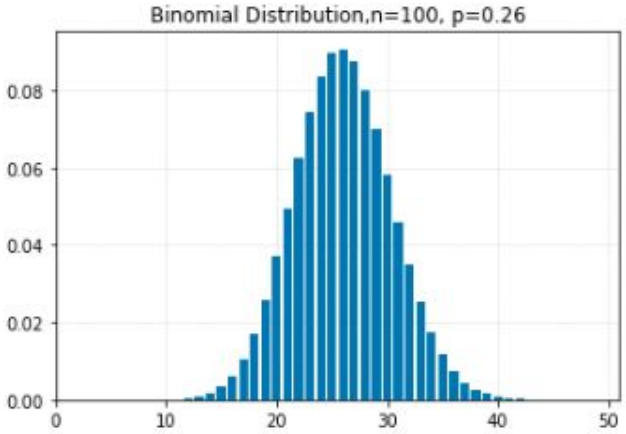
Statistical Bias

- *Bias*: when errors are systematically in one direction

- 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 lower than expected under random sampling over multiple years



Sampling Biases

Selection Bias

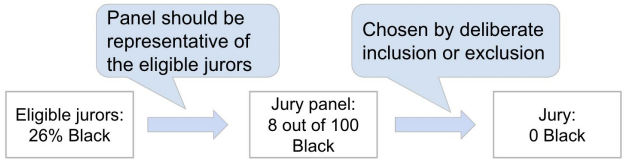
- Systematically excluding (or favoring) particular groups.

Robert Swain's case:

The Supreme Court judgment says that Talladega County jury panels were selected from a jury roll of names that the jury commissioners acquired from:

"city directories, registration lists, club and church lists, conversations with other persons in the community, both white and [not white], and personal and business acquaintances."

This process was clearly biased against Black people and in favor of people in the commissioners' social and professional circles. Such systematic exclusion of Black people from the jury rolls meant that very few Black people were selected for the jury panels.



Further reading about this case and bias in jury selection:
https://inferentialthinking.com/chapters/11/1/Assessing_a_Model.html#further-reading

Recap: Steps in Hypothesis Testing

- **Define the null hypothesis and the alternative hypothesis**
- **Choose a significance level** (cutoff tail probability after which you will decide the null hypothesis is inconsistent with the observed data)
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
- **Simulate the test statistic (or calculate directly when possible)** under the null assumptions
- **Gather observed data and compare** to the null hypothesis predictions:
  - Draw a histogram of (simulated) values of the statistic
  - Compute the observed statistic and the p-value from the real sample
- If the p-value is less than the significance level: Reject Null and Accept Alternative. Otherwise Fail to Reject Null.

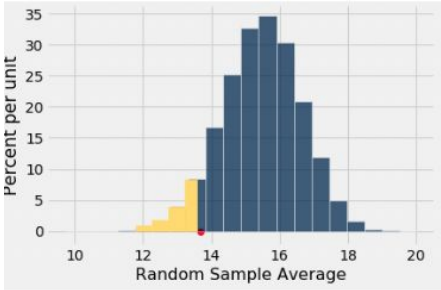
Conclusion of the Test

Determine whether observed test statistic is consistent null hypothesis:

- If p-value is less than your chosen significance level:
  - Reject the null hypothesis in favor of the alternative
  - Else: Fail to reject the null hypothesis

Conventions About Inconsistency

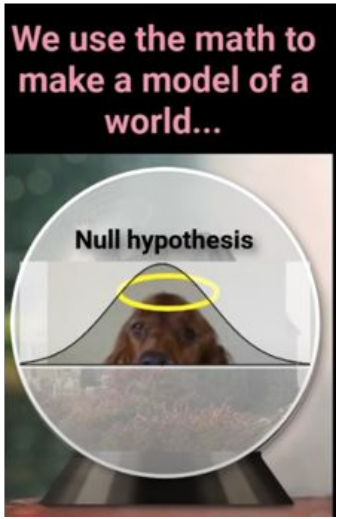
- **"Inconsistent with the null":** The observed test statistic is in the tail of the empirical distribution under the null hypothesis



- **"In the tail," first convention:**
  - The area in the tail is less than 5%
  - The result is "statistically significant"
- **"In the tail," second convention:**
  - The area in the tail is less than 1%
  - The result is "highly statistically significant"

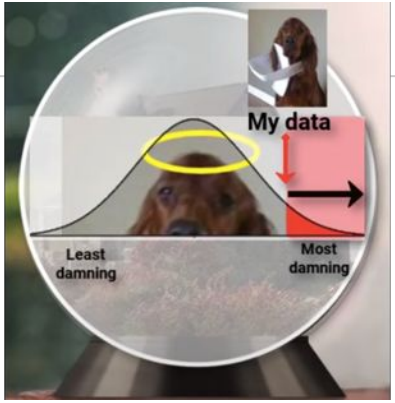
Default action:
Don't shout at Fido.

Null hypothesis:
Fido is innocent.



Hypothesis testing:

`Ridiculous? [Y/n]`



The lower the p-value, the more surprising the evidence is, the more ridiculous our null hypothesis looks

> A p-value doesn't *prove* anything. It's simply a way to use surprise as a basis for making a reasonable decision.
>
> — Cassie Kozyrkov

Hypothesis Testing Caveats

Whether you use a conventional cutoff or your own judgment, it is important to keep the following points in mind:

- Always provide the observed value of the test statistic and the p-value, so that readers can decide whether or not they think the p-value is small.
- Don’t look to defy convention only when the conventionally derived result is not to your liking.
- Even if a test concludes that the data don’t support the chance model in the null hypothesis, it typically doesn’t explain *why* the model doesn’t work. Don’t make causal conclusions without further analysis, unless you are running a randomized controlled trial. We will analyze those in a later section.

Guidelines for Choosing Test Statistics

Test Statistic: the statistic that we choose to simulate, to decide between the two hypotheses.

Questions before choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?

- What values will make us lean towards the alternative?
  - Preferably, the answer should be just “high”.
  - In this class, try to **avoid “both high and low”**.

Choosing Test Statistics

In each of (a) and (b), choose a statistic that will help you decide between the two viewpoints.

Data: the results of 400 tosses of a coin

a)
- Null: “This coin is fair.”
- Alternative: “No, it’s biased towards heads.”

Large values of the percent of heads suggest “biased towards heads”

b)
- Null: “This coin is fair.”
- Alternative: “No, it’s not”

Very **large** or very **small** values of the percent of heads suggest “not fair.”

- The **distance** between percent of heads and 50% is the key

Possible Test Statistic:

- percent of heads
- or number of heads

- abs(% heads - 50%)
- or abs(#heads - 200)

Comparing Multiple Distributions

CSCI 3022

Hypothesis Testing:

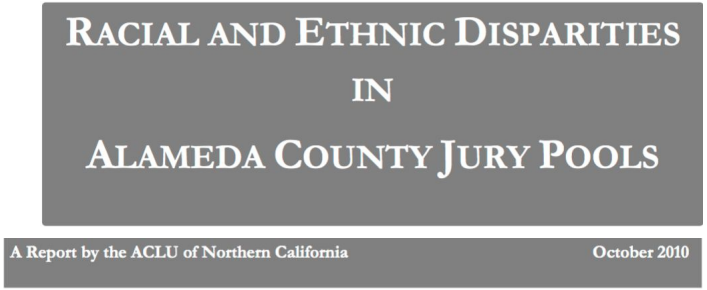
- Recap
- **Comparing Multiple Distributions**
  - Total Variation Distance
- Hypothesis Testing with 2 Categories
- Hypothesis Testing with Simple Sample of Numerical Data

`

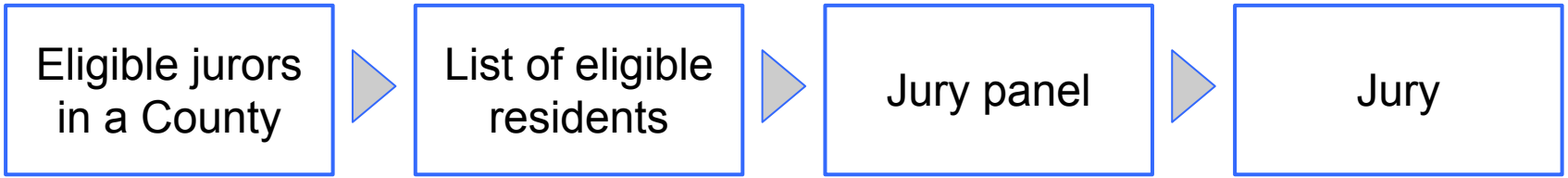
Jury Selection in Alameda County

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 racial and 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.

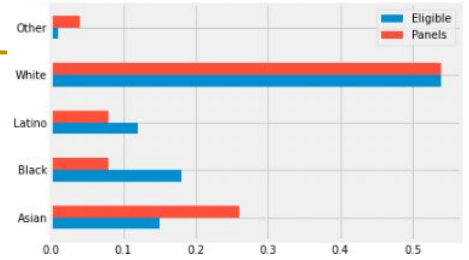


Jury Panels



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."

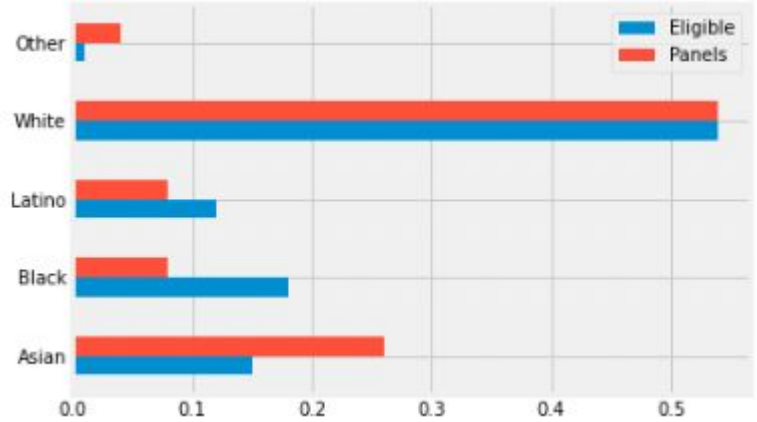
Example: Jury Selection in Alameda County



- **State the Null and Alternative Hypotheses**
  - Null: the people on the jury panels were selected at random from the eligible population.. Any difference we see between the population demographics and the actual jury panel demographics is due to chance.
    - Mathematical Model of Null Hypothesis:
- **Alternative Hypothesis**: The jury panel selection was biased
- **Choose a Significance Level:** _____
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
  - Test Statistic: _____
- **Simulate the test statistic distribution** under the null assumption. Draw the distribution (or density histogram of simulated values).
- **Collect data and compare** the data to the null hypothesis predictions:
  - Compute the observed statistic from the real sample
- If the p-value is less than the chosen significance level, we reject the null and accept the alternative. Otherwise we FAIL to reject the null.

Distance Between Distributions

- 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

TVD

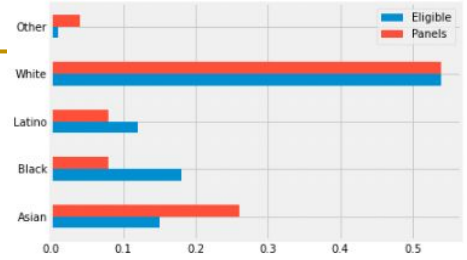
CSCI 3022

Hypothesis Testing:

- Recap
- Comparing Multiple Distributions
  - **Total Variation Distance**
- Hypothesis Testing with 1 Category
- Hypothesis Testing with Simple Sample of Numerical Data

`

Test Statistic: Total Variation Distance (TVD)

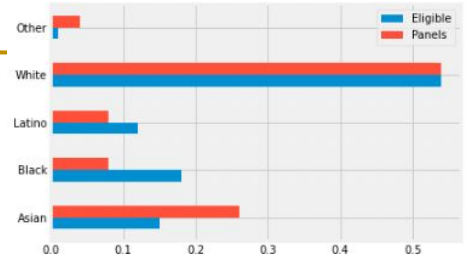


Total Variation Distance (TVD):

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

	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

Test Statistic: Total Variation Distance (TVD)



Total Variation Distance (TVD):

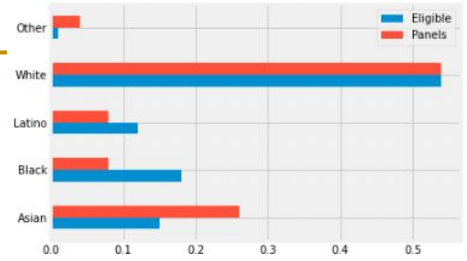
- For each category, compute the difference in proportions between two distributions

Ethnicity	Eligible	Panels	Difference
Asian/PI	0.15	0.26	0.11
Black/AA	0.18	0.08	-0.1
Caucasian	0.54	0.54	0
Hispanic	0.12	0.08	-0.04
Other	0.01	0.04	0.03

- Take the absolute value of each difference

- Sum, and then divide the sum by 2

Test Statistic: Total Variation Distance (TVD)



Total Variation Distance (TVD):

- For each category, compute the difference in proportions between two distributions

Ethnicity	Eligible	Panels	Difference
Asian/PI	0.15	0.26	0.11
Black/AA	0.18	0.08	-0.1
Caucasian	0.54	0.54	0
Hispanic	0.12	0.08	-0.04
Other	0.01	0.04	0.03

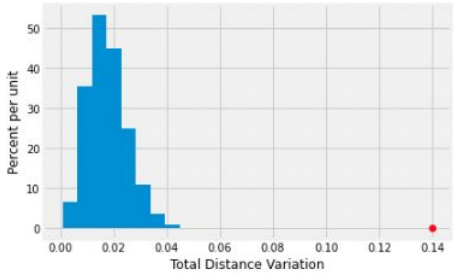
- Take the absolute value of each difference

Ethnicity	Eligible	Panels	Difference	Absolute Difference
Asian/PI	0.15	0.26	0.11	0.11
Black/AA	0.18	0.08	-0.1	0.1
Caucasian	0.54	0.54	0	0
Hispanic	0.12	0.08	-0.04	0.04
Other	0.01	0.04	0.03	0.03

- Sum, and then divide the sum by 2

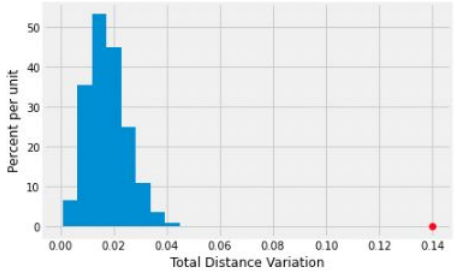
Example: Jury Selection in Alameda County

- **State the Null and Alternative Hypotheses**
  - Null: the people on the jury panels were selected at random from the eligible population.. Any difference we see between the population demographics and the actual jury panel demographics is due to chance.
    - Mathematical Model of Null Hypothesis: Multinomial (N=1423, p=[0.15, 0.18, 0.12, 0.54, 0.01]
  - **Alternative:**: The jury panel selection was biased
- **Choose a Significance Level:** 1 %
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
  - Test Statistic: TVD
- **Simulate the test statistic distribution** under the null assumption. Dr the distribution (or density histogram of simulated values).
- **Collect data and compare** the data to the null hypothesis predictions
  - Compute the observed statistic from the real sample
  - Compute empirical p-value
- If the empirical p-value is less than the chosen significance level, we reject the null and accept the alternative. Otherwise we FAIL to reject the null.



CONCLUSION:

Example: Jury Selection in Alameda County



Hypothesis Tests Don't Tell us WHY the result was biased.

Based on the ACLU report:

- Didn't use valid random sampling software
- Only sampled people who were registered with DMV or registered voters
- Can't reach people with out-of-date addresses
- Potential panelists have to be able to appear. 1st day of jury service is not compensated, and compensation for subsequent days is $15 per day. While employers are required by law to excuse employees who have jury duty, they are not required to provide compensation, and some employers don't.

CONCLUSION: There WAS bias in the jury selection

https://inferentialthinking.com/chapters/11/2/Multiple_Categories.html#reasons-for-the-bias

Testing with 2 Categories

CSCI 3022

Hypothesis Testing:

- Recap
- Comparing Multiple Distributions
- **Hypothesis Testing with 2 Categories**

`

Ex 3: Another Example


Gregor Mendel, 1822-1884

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

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

- Let's test this hypothesis

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:

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:

  abs(sample percent of purple flowering plants - 75)

- If the statistic is large, that is evidence against the model
- Notice: the statistic above is just the TVD for the binomial case

(Demo)

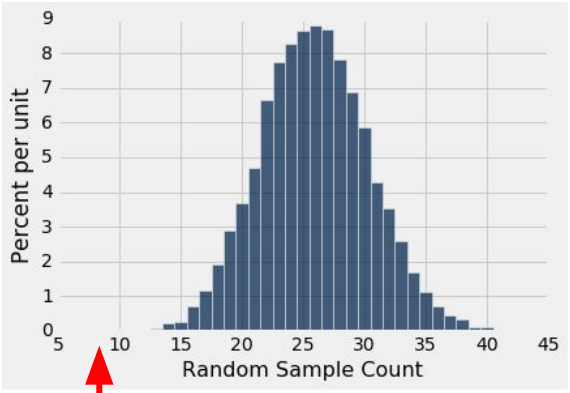
Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic (via simulation)** and its empirical distribution under the null hypothesis
- If the observed value is **not consistent** with the distribution, then the test favors the alternative (“data is consistent with the alternative”)

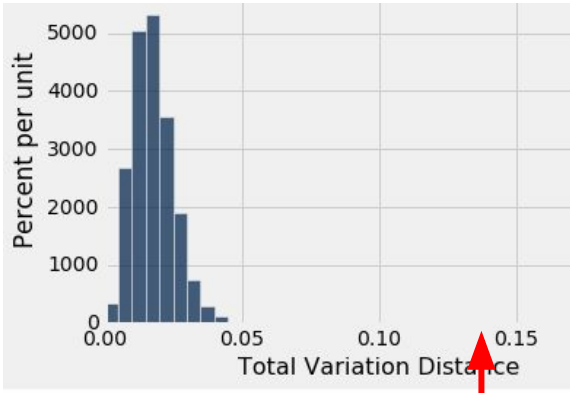
p-values as Tail Areas

Alabama Jury



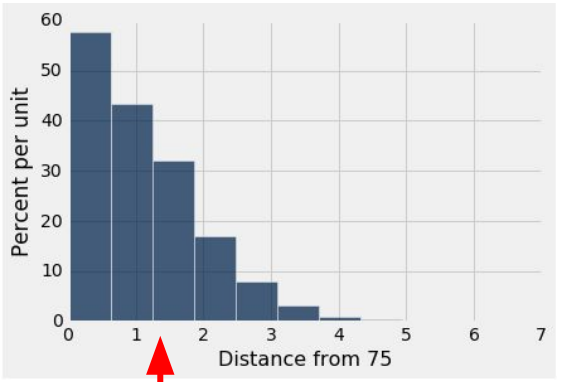
Observed Count (8)

Alameda Jury



Observed TVD (0.14)

Pea Plants



Observed Distance (1.32)

Test Scores: Unfair Disadvantage?

Example:

- Consider a Calculus class with 359 students divided into 12 recitation sections. TA's lead the sections.
- After the midterm, students in Section 3 notice that the average score in their section is lower than in others. They complain that it must be due to their particular TA.

	Midterm	
	student count	mean
Section		
1	32	15.593750
2	32	15.125000
3	27	13.666667
4	30	14.766667
5	33	17.454545
6	32	15.031250
7	24	16.625000
8	29	16.310345
9	30	14.566667
10	34	15.235294
11	26	15.807692
12	30	15.733333

Can we design a Hypothesis Test to test the students' concerns?

Null Hypothesis (TA's Defense):

If students had been picked at random to be in Section 3 we could have gotten a midterm average like the one we observed. Any difference we see in the Section 3 scores and the rest of the sections' scores is due to chance.

Alternative Hypothesis:

- No, the average score for Section 3 is too low to be explained solely by random chance.

Hypothesis Testing Example: Test Scores

- **State the Null and Alternative Hypotheses**
  - **Null Hypothesis:** If students had been picked at random to be in Section 3 we could have gotten a midterm average like the one we observed. Any difference we see in the Section 3 scores and the rest of the sections' scores is due to chance.
    - Mathematical Model: Randomly select the midterm scores of 27 students from the class (_____ replacement) and calculate the test statistic
  - **Alternative Hypothesis**: No, the average score for Section 3 is too low to be explained solely by random chance.
- **Choose a Significance Level:** _____
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
  - Test Statistic: _____
- **Simulate (or calculate theoretically) the test statistic distribution** under the null assumption.
- **Collect data and compare** the data to the null hypothesis predictions:
  - Compute the observed statistic from the real sample
  - Compute (empirical) p-value
- If the (empirical) p-value is less than the chosen significance level, we reject the null and accept the alternative. Otherwise we FAIL to reject the null.