

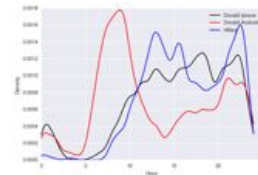
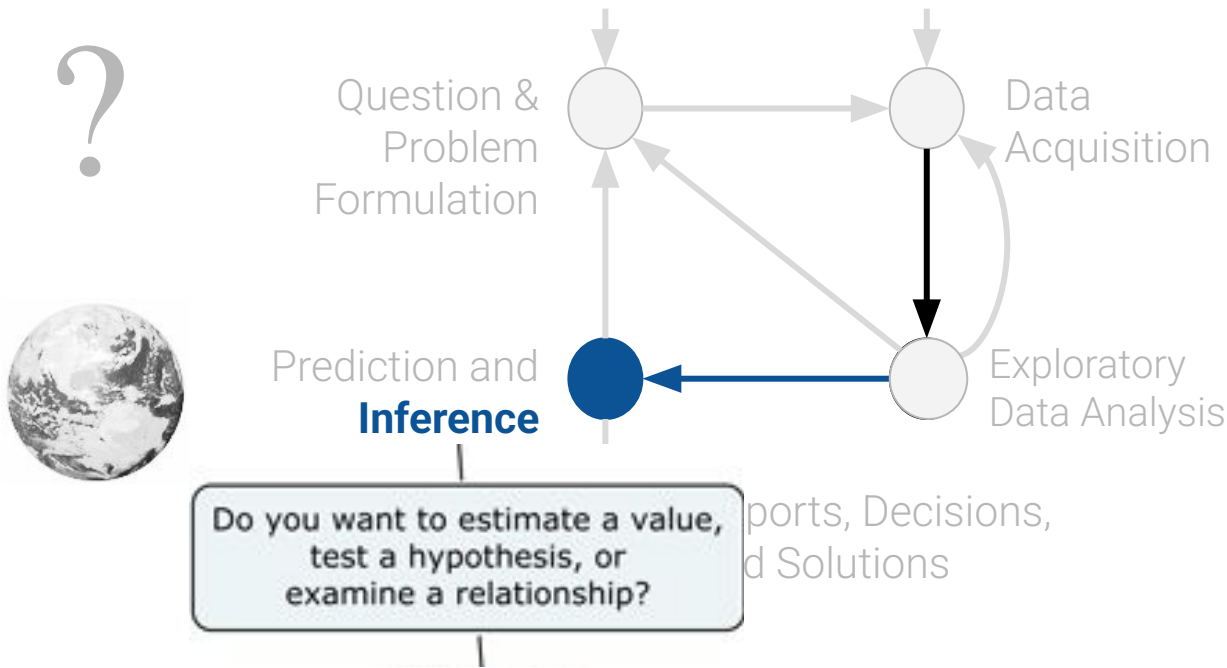
LESSON 18

Hypothesis Testing

Course Logistics: Your 8th Week At A Glance

Mon	Tues	Wed	Thurs 10/17	Fri
Attend & Participate in Class		Attend & Participate in Class		Attend & Participate in Class
	Exam 1 feedback/grades posted			HW 8 released HW 7 due 11:59pm MT

Plan for next few weeks: Statistical Inference



Roadmap

- [Finish lesson 17:](#)
 - The Central Limit Theorem**
 - Parameters, Statistics and Estimators**
 - Distributions of Statistics**
- Start lesson 18: Hypothesis Testing

"Statistics is the science of making decisions under uncertainty."

-Savage, The Foundations of Statistics, 1954.



Lesson 18

- Intro to Hypothesis Testing
 - Comparing a sample to a model
- When To Use Hypothesis Testing
- Comparing Multiple Distributions
- Steps in Hypothesis Testing
 - Choose null and alternative hypotheses
 - Choose a significance level
 - Choose a test statistic
 - Gather data and calculate observed test statistic
 - Simulate/theoretically determine distribution of test statistic under null hypothesis
 - Calculate p-value
 - Make conclusion

Introduction To Hypothesis Testing

- Intro to Hypothesis Testing
 - Supreme Court Case Example

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?

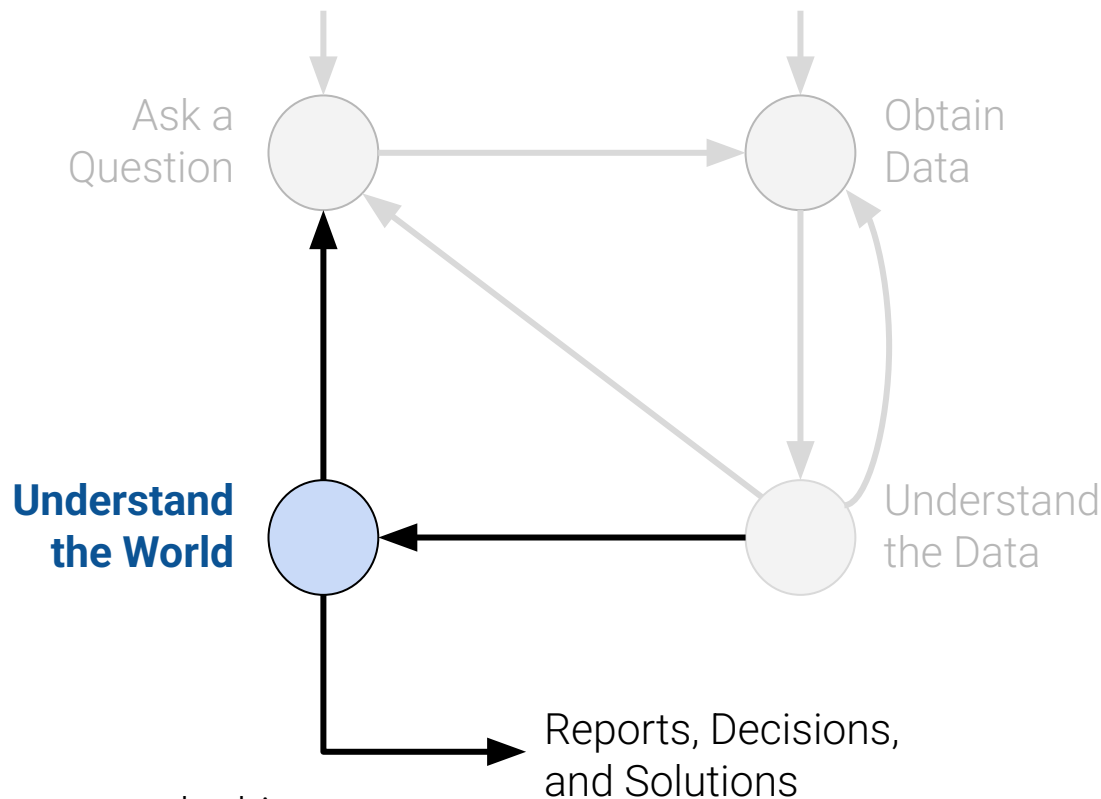
Some goals of Data Science:

- 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.



Ex 1: Supreme Court Case

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

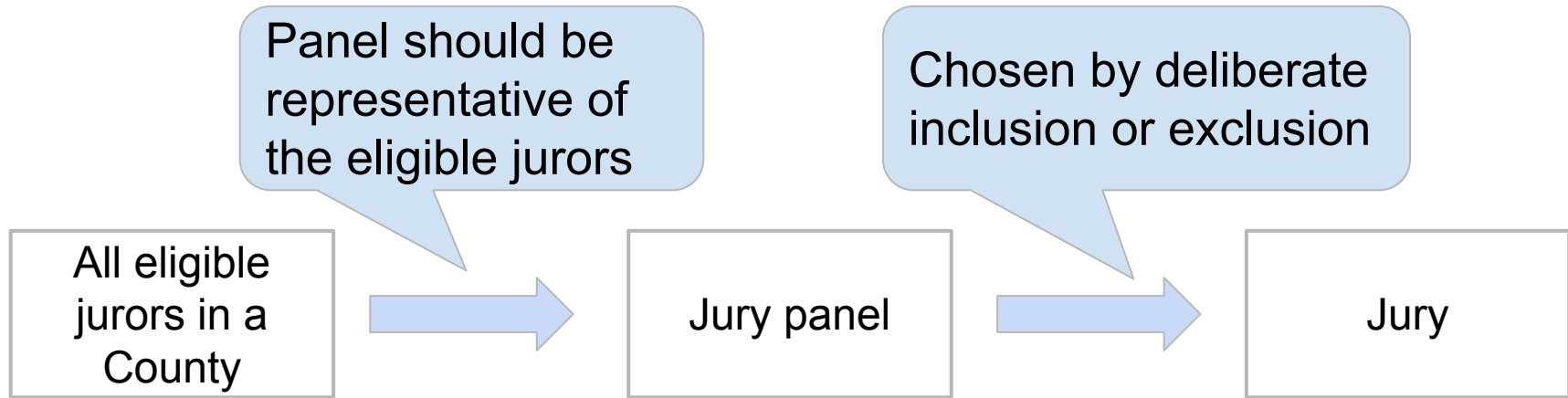
We will study a U.S. Supreme Court case in the 1960s

- 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.

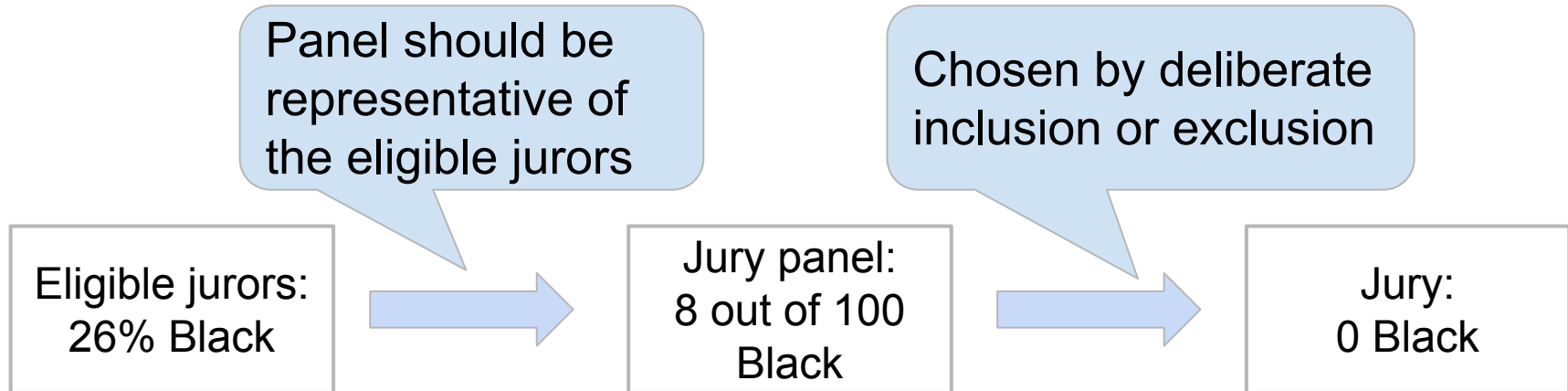
US Constitution:

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



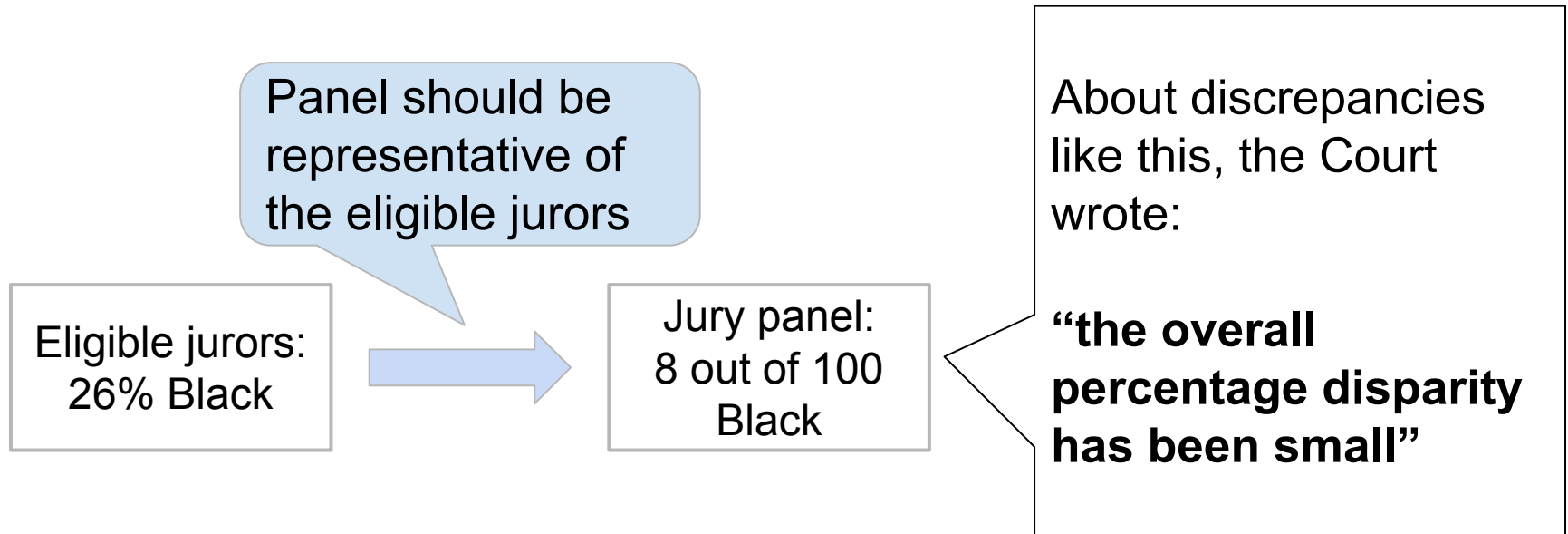
Ex 1 Background: Robert Swain's Case

- 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

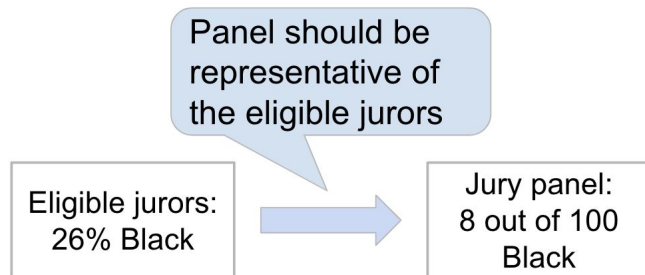


Ex 1 Background:: Supreme Court Ruling, 1965

- The Court denied Robert Swain's appeal.



Ex 1: 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?

- A statistical hypothesis 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 was 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

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

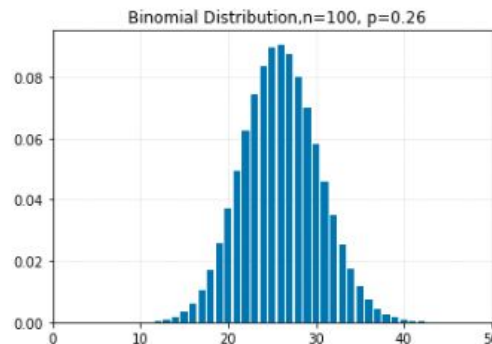
Ex: Robert Swain Jury selection Example:

- **“Null” Hypothesis:** The people on the jury panels were selected at random from the eligible population which consisted of 26% Black people. 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

Ex 1: Hypothesis Test

- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis
- Model the distribution of the test statistic under the null hypothesis (either theoretically or empirically via simulation).

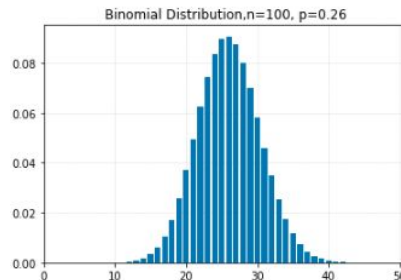
$$X \sim \text{Bin}(100, 0.26)$$



Ex 1: Hypothesis Test

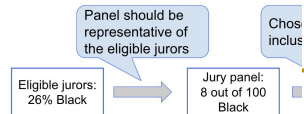
- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis
- Model the distribution of the test statistic under the null hypothesis (either theoretically or empirically via simulation).

$$X \sim \text{Bin}(100, 0.26)$$



- This displays the **distribution of the test 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**)

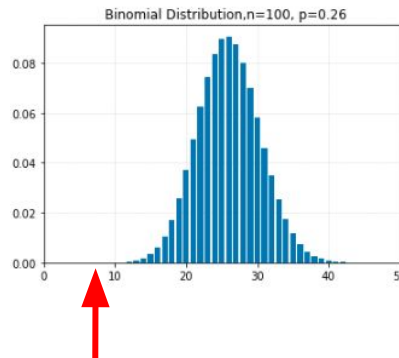
Ex 1: Hypothesis Test



- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis

- Distribution of the test statistic under the null hypothesis :

$$X \sim \text{Bin}(100, 0.26)$$

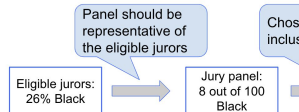


- **Observed Test Statistic:** 8

Observed Test Statistic

- If the **observed test statistic** is in the tail* of the null distribution, we reject the null hypothesis (i.e. the data is more consistent with the alternative hypothesis)

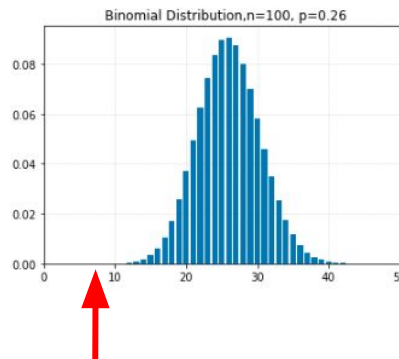
Ex 1: Hypothesis Test



- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis

- Distribution of the test statistic under the null hypothesis :

$$X \sim \text{Bin}(100, 0.26)$$



Observed Test Statistic

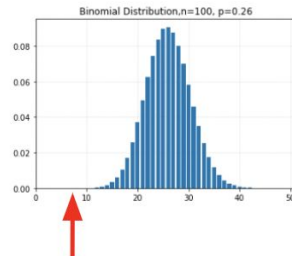
- **Observed Test Statistic:** 8
 - If the **observed test statistic** is in the tail* of the null distribution, we reject the null hypothesis (i.e. the data is more consistent with the alternative hypothesis)
- **P(observed test statistic or more extreme | null hypothesis):**
$$P(X \leq 8) = \sum_{k=0}^8 \binom{100}{k} (0.26)^k (1 - 0.26)^{100-k}$$
$$= \text{stats.binom.cdf}(8, 100, 0.26) = .00000473$$

Ex 1: Hypothesis Test

- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis

- Distribution of the test statistic under the null hypothesis :

$$X \sim \text{Bin}(100, 0.26)$$



Observed Test Statistic

- **Observed Test Statistic:** 8

- If the **observed test statistic** is in the tail* of the null distribution, we reject the null hypothesis (i.e. the data is more consistent with the alternative hypothesis)

- **P(observed test statistic or more extreme | null hypothesis):**

$$\begin{aligned} P(X \leq 8) &= \sum_{k=0}^8 \binom{100}{k} (0.26)^k (1 - 0.26)^{100-k} \\ &= \text{stats.binom.cdf}(8, 100, 0.26) = .00000473 \end{aligned}$$

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

Conventions About Inconsistency:

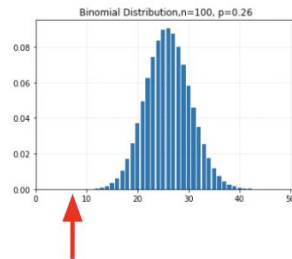
- **“In the tail,” first convention:**
 - The area in the tail is less than 5%
 - The result is “statistically significant”
- **“In the tail,” second convention:**
 - The area in the tail is less than 1%
 - The result is “highly statistically significant”

Ex 1: Hypothesis Test

- **Null Hypothesis:** The people on the jury panels were selected at random from the eligible population where 26% are Black people
- **Alternative Hypothesis:** No, the selection of the jury panels was biased against Black people
- **Test Statistic:** Number of Black people chosen out of 100 assuming null hypothesis

- Distribution of the test statistic under the null hypothesis :

$$X \sim \text{Bin}(100, 0.26)$$



Observed Test Statistic

- **Observed Test Statistic:** 8
 - If the **observed test statistic** is in the tail* of the null distribution, we reject the null hypothesis (i.e. the data is more consistent with the alternative hypothesis)

- **P(observed test statistic or more extreme | null hypothesis):**

$$\begin{aligned} P(X \leq 8) &= \sum_{k=0}^8 \binom{100}{k} (0.26)^k (1 - 0.26)^{100-k} \\ &= \text{stats.binom.cdf}(8, 100, 0.26) = .00000473 \end{aligned}$$

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

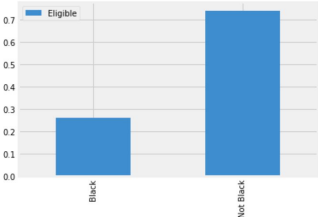
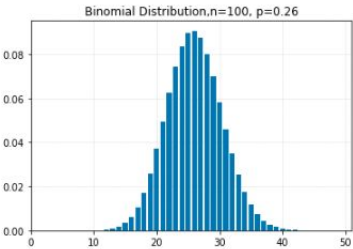
Conclusion of Test: Reject the null. Data is consistent with the alternative hypothesis that the jury selection was biased against Black people and the result is highly statistically significant.

Conventions About Inconsistency

- **“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”

Summary of Hypothesis Test

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>The 100 people on the jury panel were selected at random from the eligible population (where 26% are Black as shown in distribution above).</p> <p>Any difference we see between the population demographics and the jury panel is due to chance</p>	<p>The selection of the jury panels was biased against Black people</p>	<p>1%</p>	<p>X = Number of Black people chosen out of 100 assuming null hypothesis</p>	<p><code>x = np.random.binomial(100, 0.26)</code></p>  <p>Observed Test Statistic (8)</p>	<p>8 out of 100 on the jury panel were Black.</p>	<p>P-value = $P(X \leq 8) = 0.00000473$</p> <p>Conclusion: Since $0.000473 < 0.01$ We reject the null.</p> <p>The data is consistent with the alternative and the result is highly statistically significant.</p>

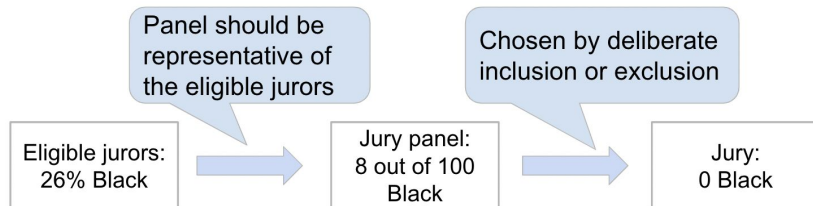
Ex 1: Supreme Court Case

The analysis above provides quantitative evidence of unfairness in Robert Swain's trial.

The data support his position that he was denied the impartial jury to which he was entitled by the U.S. Constitution.

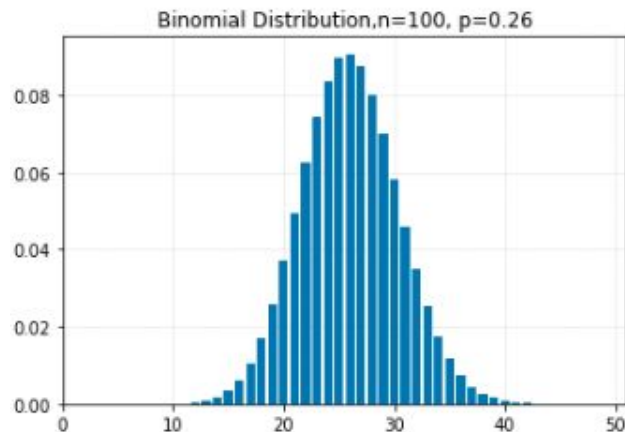
In the 1960s, the Supreme Court looked at the data but drew the wrong conclusion that "the overall percentage disparity has been small."

A hypothesis test does not establish the reason *why* the difference is not due to chance. Establishing causality is usually more complex than running a test of hypotheses.



Ex 1: Supreme Court Case

- 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



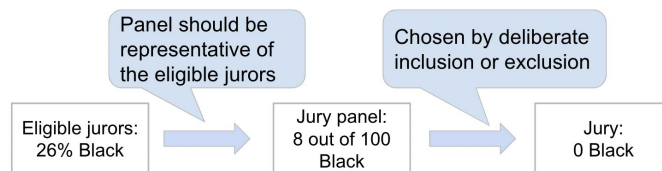
For more details: https://inferentialthinking.com/chapters/11/1/Assessing_a_Model.html#racial-bias

Ex 1: Additional Context

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 information indicates the sampling frame (jury roll) didn't reflect the target population, instead it was 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.



Selection Bias

- Systematically excluding (or favoring) particular groups.

Hypothesis Testing Framework

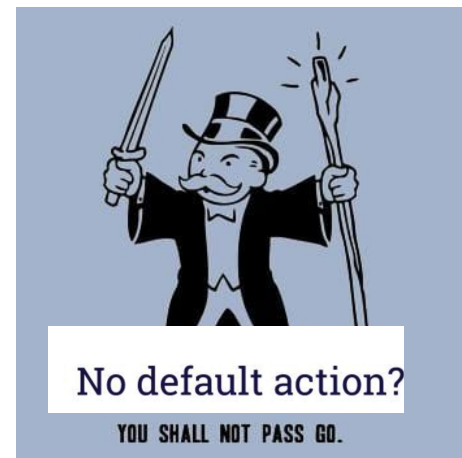
- When to use hypothesis testing

Hypothesis Testing Framework (Frequentist)

To begin, you need:

A Default course of action:

- 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 the null hypothesis is true for sure*



Choosing a default action requires domain knowledge

Examples where society considers the default to be fairly obvious:

- innocent-until-proven-guilty (default = don't convict if there's no evidence)
- testing new medications (default = don't approve if there's no evidence)
- scientific publication (default = don't publish if there's no evidence)

Skip hypothesis testing if:

1. You can answer with certainty
2. You have no default action.

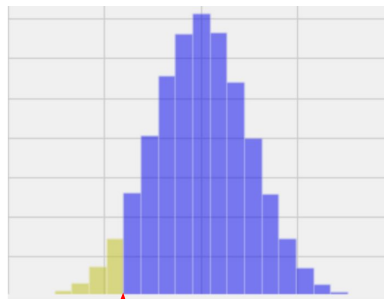
"Statistics is the science of making decisions under uncertainty."

-Savage, The Foundations of Statistics, 1954.



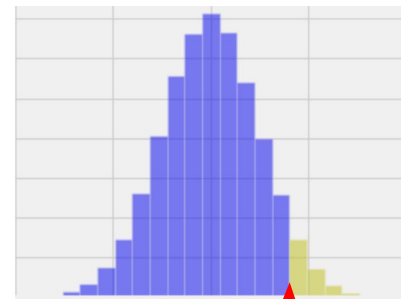
Steps in (Frequentist) Hypothesis Testing

- **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 distribution of the test statistic (or calculate directly when possible)** under the null hypothesis assumptions
- **Gather observed data and compare** to the null hypothesis predictions:
 - Compute the **observed statistic and the p-value** from the real sample
- If the **p-value is less than** the significance level: Reject Null. Otherwise Fail to Reject Null. We can't PROVE either hypothesis is true in this framework.



Simulated values when **LOW** test statistics support the alternative hypothesis

Observed Test Statistic



Simulated values when **HIGH** test statistics support the alternative hypothesis

Observed Test Statistic

- Yellow area denotes the p-value

Comparing Multiple Distributions

Hypothesis Testing:

- Comparing Multiple Distributions
 - Total Variation Distance

Example: 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.

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California

October 2010

Example: 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."

The ACLU compiled data on the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010.

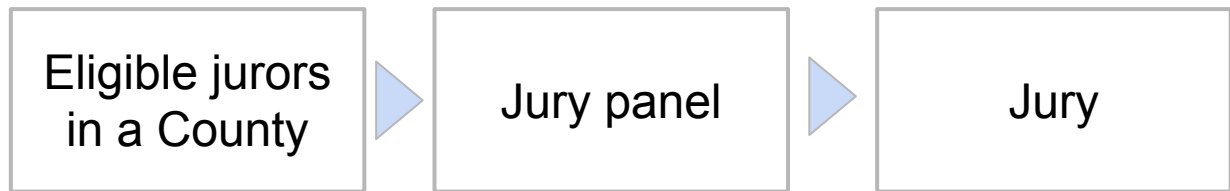
In those panels, the total number of people who reported for jury service was 1453.

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/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04

The first numerical value is the proportion of all eligible juror candidates in that category. The second value is the proportion of people in that category among those who appeared for the process of selection into the jury.

Example: Jury Selection in Alameda County



- **Question:** Would the composition of the jury panels observed in the ACLU report be a realistic outcome if the jury panel selection process were truly unbiased?

Ethnicity	Eligible	Panels
Asian/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04

Example: Jury Selection in Alameda County: Hypothesis Test

- **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 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 Hypothesis:** No, the selection of the jury panels was biased.
- **Significance Level:** _____
- **Test Statistic:** _____
 - Model the distribution of the test statistic under the null hypothesis (either theoretically or empirically via simulation).

Ethnicity	Eligible	Panels
Asian/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04

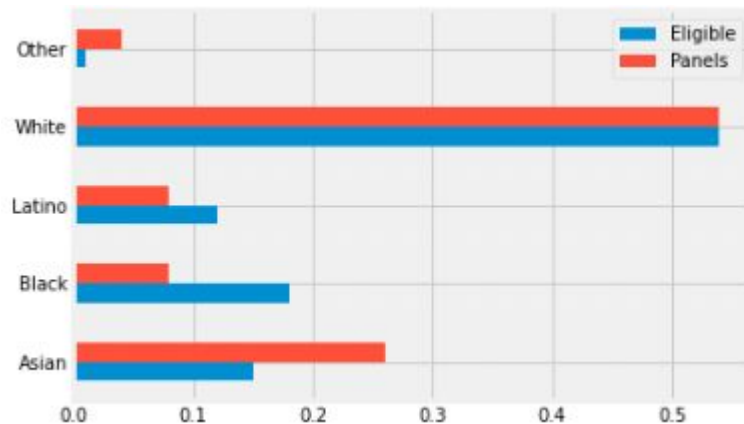
*[See Lec 16](#)

*Technical note. Random samples of prospective jurors would be selected without replacement. 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 use a multinomial distribution (i.e. sample with replacement) to model this.

Example: Choosing a Test Statistic

Need a Statistic To Measure the “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

Total Variation Distance

Hypothesis Testing:

- Comparing Multiple Distributions
 - **Total Variation Distance**

Test Statistic: Total Variation Distance (TVD)

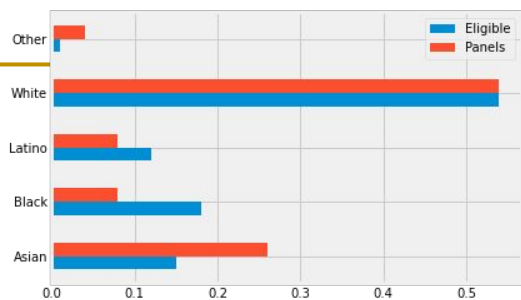
Total Variation Distance (TVD):

- 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

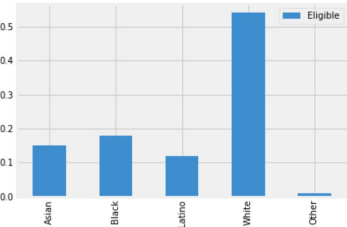
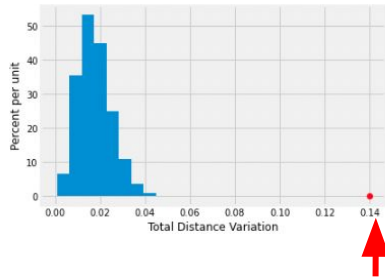
- Sum (and then divide the sum by 2)



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

Summary of This Hypothesis Test

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>The 1423 people on the jury panels were selected at random from the eligible population (with ethnicity distribution above).</p> <p>Any difference we see between the population ethnicities and the actual jury panel ethnicities is due to chance.</p>	<p>The selection of the jury panels was biased.</p>	<p>1%</p>	<p>$X = \text{Total Variation Distance (TVD)}$</p>	<pre>Null_dist = [0.15, 0.18, 0.54, 0.12, 0.01] Sim_dist = np.random.multinomial(N=1423, p=Null_dist)/1423 X=np.sum(np.abs(Null_dist-Sim_dist))/2</pre>  <p>Observed Test Statistic (0.14)</p>	<p>Observed TVD: 0.14</p>	<p>Empirical P-value = $P(X \geq 0.14) = 0/1000000$</p> <p>Conclusion: Since $0 < 0.01$ We reject the null. The data is consistent with the alternative hypothesis.</p> <p>The result is highly statistically significant.</p>

Example: Jury Selection in Alameda County

Based on the data available, the results of our hypothesis test supports the ACLU's conclusion that the jury panels were not representative of the distribution provided for the eligible jurors.

However, hypothesis tests don't tell us WHY the result was inconsistent with the null model.

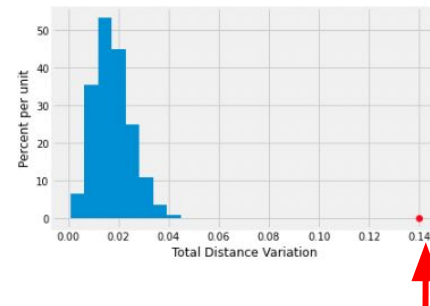
Some potential explanations based on the ACLU report:

- 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 - may not be feasible
 - 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.

Example: Jury Selection in Alameda County: Hypothesis Test

- **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 actual jury panel demographics is due to chance.
 - Mathematical Model of Null Hypothesis
 - Multinomial Distribution ($N=1423$, $p=[0.15, 0.18, 0.12, 0.54, 0.01]$)
- **Alternative Hypothesis:** No, the selection of the jury panels was biased.
- **Significance Level:** 0.01
- **Test Statistic:** Total Variation Distance (TVD)
 - Model the distribution of the test statistic under the null hypothesis (empirically via simulation).
- **Collect Data and Calculate the Observed Test Statistic:** 0.14
- **Calculate the (empirical) p-value:**
P(observed test statistic or more extreme | null hypothesis): 0/1000000

Ethnicity	Eligible	Panels
Asian/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04



- **Conclusion of Test:** Since the p-value is _____ than the significance level 0.01, we _____ the null. Data is consistent with the _____ hypothesis and the result is _____.

ACLU estimates of demographics of eligible jurors:

- The ACLU used estimates developed by a San Diego State University professor for an Alameda County trial in 2002. Those estimates were based on the 2000 Census and also took into account the criteria required for eligibility as a juror.
 - Using estimates based on the 2000 Census for populations in 2010 might not be accurate due to the changing demographics in California.

Estimates of demographics of the jury panel:

- Unclear exactly how the 1453 panelists were classified into the different ethnic categories. The report says only that “attorneys ... cooperated in collecting jury pool data”.
- Non-response bias: Data on panelists was obtained from those who reported for service. Not all panelists do so. The reasons for not reporting are often associated with race and ethnicity, and disproportionately affect panelists from under-resourced communities.

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. Otherwise Fail to Reject Null.

Hypothesis Testing: Comparing A Sample To A Model

Testing whether a sample looks like random draws from a specified chance model.

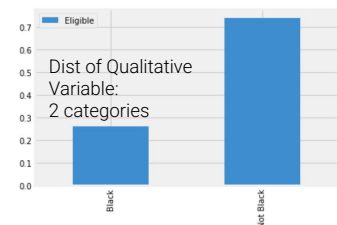
Recap: Hypothesis Testing So Far

1). Test whether a **single sample** looks like **random** draws from a **specified chance model (null)**.

- Do jury panel demographics look like a **random** sample from the **known population demographics of eligible jurors**?

Null Hypothesis

Selection was based on repeated random draw from known distribution:



Alternative Hypothesis

Selection biased against Black people

Choose Significance Level: 1%

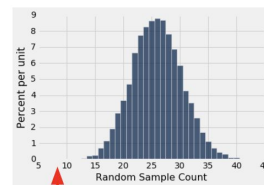
Test Statistic

Count of 1 category

Simulate Distribution of Test Statistic Under Null

```
np.random.  
binomial(100, 0.26)
```

Calculate Observed Test Statistic

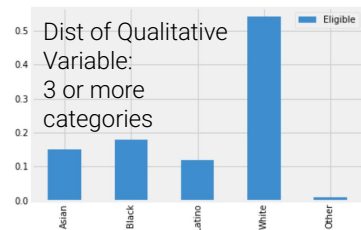


Observed Count (8)

Calculate p-value & Conclusion

Empirical p-value = $0 < 0.01$
Reject Null

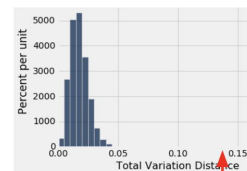
"Data is consistent with the alternative"



Selection was biased

Total Variation Distance (TVD)

```
sim_dist = np.random.  
multinomial(929,  
null_dist)/929  
  
sum(abs(sim_dist - null_dist))/2
```



Observed TVD (0.14)

Empirical p-value = $0 < 0.01$
Reject null

"Data is consistent with the alternative"

In these cases there was an observed sample and we knew the population distribution, so we were testing whether the observed sample was really **randomly** chosen from that population

Hypothesis Testing

1). Test whether a **single sample** looks like **random** draws from a **specified chance model (null)**.

What if we don't know the population distribution?

- We can make a “best guess” about the population distribution
- Take a random sample and use a hypothesis test to determine whether the data is consistent with our guess.

Example: Testing hypothesis about population distribution

Gregor Mendel (1822-1884) was an Austrian monk and founder of the modern field of genetics.

Among many experiments, he tested the hypothesis that pea plants will bear purple or white flowers at random, in the ratio 3:1.

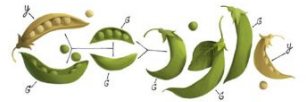
- Mendel's hypothesis:
 - Each plant is purple-flowering with chance 75%,
 - regardless of the colors of the other plants

Let's test this using a hypothesis test:

What should he use for the null hypothesis?

What should he use for the alternative hypothesis?

Gregor Mendel, 1822-1884



Ex: Testing Mendel's hypothesis

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
<div><p>For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation.</p></div>	The chance of purple flowers is not 75%.	5%				

Choosing the Test Statistic

Null: For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation

Alternative: The chance of purple flowers is **not** 75%.

Choosing a test statistic: What values will make us lean towards the alternative?

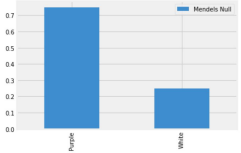
- Preferably, the answer should be just “high” or just “low”
- In this class, **try to avoid “both high and low”**.

Poll: Which of the following test statistics could we use to test these hypotheses? Select all that apply.

- A. The proportion of plants with purple flowers.
- B. The proportion of plants with white flowers.
- C. $\text{abs}(p - 75)$, where p is the percent of plants with purple flowers.
- D. The number of different colors in the plants flowers.
- E. The total variation distance between the distribution in the observed data, vs the model distribution (0.75, 0.25)

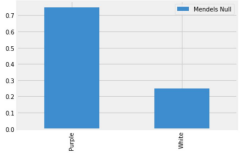
Ex: Testing Mendel's hypothesis

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation.</p>	The chance of purple flowers is not 75%.	5%	<p>$X = \text{absolute difference}$ between the simulated percent of purple flowers and 75</p> <p>Other options:</p> <p>absolute difference between the simulated number of purple flowers and 0.75(929)</p> <p>OR:</p> <p>absolute difference between the simulated proportion of purple flowers and 0.75 (This is equivalent to the TVD)</p>			

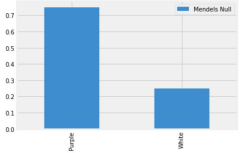
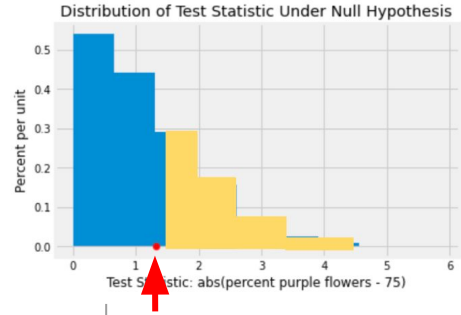
Ex: Testing Mendel's hypothesis

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation.</p>	The chance of purple flowers is not 75%.	5%	<p>$X = \text{absolute difference}$ between the simulated percent of purple flowers and 75</p> <p>Other options:</p> <p>absolute difference between the simulated number of purple flowers and 0.75(929)</p> <p>OR:</p> <p>absolute difference between the simulated proportion of purple flowers and 0.75 (This is equivalent to the TVD)</p>	<pre>sim_perc = np.random.binomial(N=929, p=0.75) X = np.abs(sim_perc-75)</pre>		

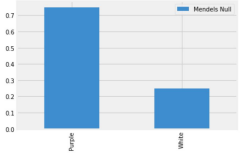
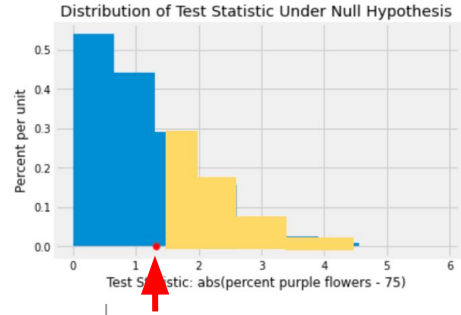
Ex: Testing Mendel's hypothesis

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation.</p>	<p>The chance of purple flowers is not 75%.</p>	<p>5%</p>	<p>$X = \text{absolute difference}$ between the simulated percent of purple flowers and 75</p> <p>Other options:</p> <p>absolute difference between the simulated number of purple flowers and 0.75(929)</p> <p>OR:</p> <p>absolute difference between the simulated proportion of purple flowers and 0.75 (This is equivalent to the TVD)</p>	<pre>sim_perc = np.random.binomial(N=929, p=0.75)</pre> <p>$X = np.abs(sim_perc - 75)$</p>	<p>Mendel grew 929 pea plants of this variety. Among these 929 plants, 705 had purple flowers.</p> <p>Observed Test Statistic:</p> <p>$abs(705/929 * 100 - 75) = 1.32$</p>  <p>Observed Test Statistic (1.32)</p>	<p>P-value =</p> <p>Conclusion:</p> <p>Yellow area is the p-value</p>

Ex: Testing Mendel's hypothesis

1). Test whether a **single sample** looks like **random** draws from a **specified chance model**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
 <p>For every plant there is a 75% chance that it will have purple flowers, regardless of the colors in all the other plants. Any observed deviation from the model is the result of chance variation.</p>	<p>The chance of purple flowers is not 75%.</p>	<p>5%</p>	<p>$X = \text{absolute difference}$ between the simulated percent of purple flowers and 75</p> <p>Other options:</p> <p>absolute difference between the simulated number of purple flowers and 0.75(929)</p> <p>OR:</p> <p>absolute difference between the simulated proportion of purple flowers and 0.75 (This is equivalent to the TVD)</p>	<pre>sim_perc = np.random.binomial(N=929, p=0.75)</pre> <p>$X = np.abs(sim_perc - 75)$</p>	<p>Mendel grew 929 pea plants of this variety. Among these 929 plants, 705 had purple flowers.</p> <p>Observed Test Statistic:</p> <p>$abs(705/929 * 100 - 75) = 1.32$</p>  <p>Observed Test Statistic (1.32)</p>	<p>P-value = 0.368 > 0.05</p> <p>Conclusion: Fail to reject the null.</p> <p>Data is consistent with the null</p> <p>Yellow area is the p-value</p>

Recap: Hypothesis Testing So Far

- 1). Test whether a **single sample** looks like **random** draws from a **specified chance model (null)**.
 - Do jury panel demographics look like a **random** sample from the **known population demographics of eligible jurors**
 - In these cases we knew the population distribution (the specified chance model), but we didn't know if the observed sample was actually obtained via random sampling, so we were testing whether the observed sample was really **randomly** chosen from that population
 - Did the pea plants that Mendel grew in his **random sample** have colors that were consistent with the **chances he specified in his model**?
 - In this case we didn't know the population distribution, but we did know that the observed sample was obtained via random sampling, so our test was about our **guess for the specified chance model** (not the randomness).
- 2). Up next: Test whether **two samples** looks like **random** draws from the **same underlying distribution**.

Null & Alternative Hypotheses

Steps in Hypothesis Testing

- **Choose null and alternative hypotheses**
- Choose a significance level
- Choose a test statistic
- Gather data and calculate observed test statistic
- Simulate/theoretically determine distribution of test statistic under null hypothesis
- Calculate p-value
- Make conclusion

Null vs Alternative Hypotheses:

All statistical tests attempt to choose between two views of the world:

Null hypothesis:

- Data was generated at random under clearly specified assumptions that make it possible to compute chances.
- The word "null" reinforces the idea that if the data look different from what the null hypothesis predicts, **the difference is due to *nothing* but chance.**

Null vs Alternative Hypotheses:

All statistical tests attempt to choose between two views of the world:

Null hypothesis:

- Criteria for selecting null:
 - Need to be able to compute/simulate chances under the null assumption
 - Ex: The sample of data is randomly chosen from this distribution (any difference we see is due to chance)
 - Models the state of a world where you'd be happy to take your default action without any further data.
 - Ex: The control and test group come from the same underlying distribution, any difference between the control and test group for this new medication is just due to chance.

- **Alternative hypothesis:**

- Some reason other than chance made the data differ from what was predicted by the null hypothesis.
- Informally, the alternative hypothesis says that the observed difference is "real" and not just due to chance.
- Null and alternative hypothesis can't be true at the same time. Alternative hypotheses will take 3 flavors in this class:
 - Either:
 - "Null is wrong"
 - "Null is wrong and the real value is greater than the null-proposition"
 - "Null is wrong and the real value is less than the null-proposition."
- We choose an alternative hypothesis based on the setting and situation we are trying to test.

The game of hypothesis testing is all about determining whether the evidence we have makes our null hypothesis look ridiculous.

If data convinces you live in the alternative hypothesis world,
change your action!



Significance Level

Steps in Hypothesis Testing

- Choose null and alternative hypotheses
- **Choose a significance level**
- Choose a test statistic
- Gather data and calculate observed test statistic
- Simulate/theoretically determine distribution of test statistic under null hypothesis
- Calculate p-value
- Make conclusion

- It is your **threshold** for deciding whether or not you think the p-value is small.
- The cutoff **does not depend on the data**. It is chosen before the data is collected.

"It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not." [Fisher 1925]

Guidelines for Choosing the Significance Level (i.e. p-value cutoff)

- Decide on it **before** seeing the results
 - Don't change it!
- Common values at 5% and 1%
 - follow conventions in your area

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author [Fisher] prefers to set a low standard of significance at the 5 percent point ..." [Fisher 1926]



Sir Ronald Aylmer Fisher [1890-1962]
Pioneer of Modern Statistics

Discussion Question

Suppose there are 500 students enrolled in CSPB 3022. We give each student a separate coin and have them toss it 160 times to test whether or not the coin is fair.

Null: The coin is _____

Alternative: The coin is _____

- Test Statistic: _____
- Significance level (cutoff for the P-value): 5%

Suppose in reality all the coins are fair.

About how many students will conclude that their coins are unfair using this hypothesis test?

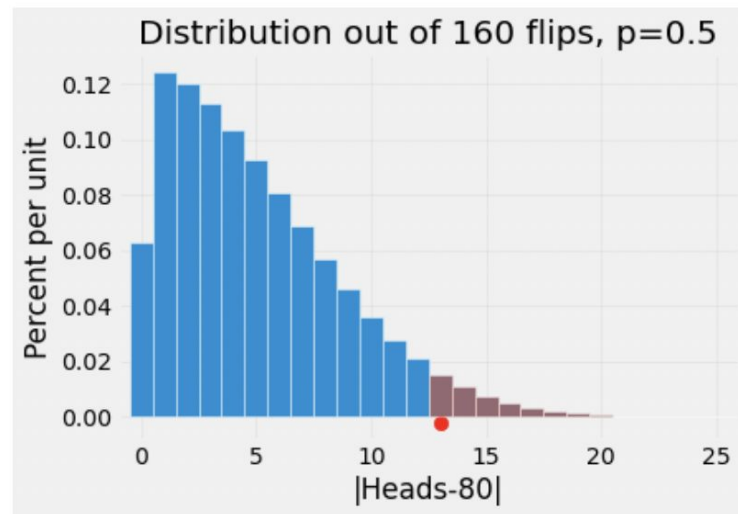
- A). 5 B). 25 C). 50 D). 120 E). 160

Recall: Significance Level as an Error Probability

- If:
 - your **significance level (i.e. p-value cutoff) is 5%**
 - and the **null hypothesis happens to be true**
- Then there is a **5% chance** that **the test will INCORRECTLY reject the null hypothesis**.

Thus, the significance level is actually a conditional probability of making one type of error:

- **Significance level = $P(\text{reject null} \mid \text{null hypothesis is true})$**

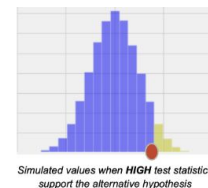
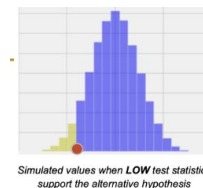


When null is true,
5% of the time you
will get an observed
test statistic in tail
shaded pink even
when the coin is fair
JUST BY CHANCE!

- It is your **threshold** for deciding whether or not you think the p-value is small.
- The cutoff **does not depend on the data**. It is chosen before the data are collected.
- It is an ***error probability***: approximately the chance that the test concludes the alternative when the null is true
 - You get to choose the cutoff.

P-value vs Significance Level (P-value cutoff)

- Significance level (i.e. P-value cutoff): You Pick It
 - Does not depend on observed data or simulation
 - **P(reject null | null hypothesis is true)**
 - “Acceptable” probability of rejecting the null hypothesis when it is true.
 - Common Conventions
 - Significance level = 5%
 - If your p-value < 5%, then reject null and result is called “statistically significant”
 - Significance level = 1%
 - If your p-value < 1%, then reject null and result is “highly statistically significant”
- P-value (You Compute It)
 - Depends on the observed data and simulation
 - **P(data you observed or more extreme | null hypothesis is true)**
 - Probability under the null hypothesis that the test statistic is the observed value or more extreme in the direction of the alternative



- Yellow area denotes the p-value
- Red dot denotes the observed statistic.

Test Statistics

Steps in Hypothesis Testing

- Choose null and alternative hypotheses
- Choose a significance level
- **Choose a test statistic**
- Gather data and calculate observed test statistic
- Simulate/theoretically determine distribution of test statistic under null hypothesis
- Calculate p-value
- Make conclusion

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” or just “low”
 - In this class, try to **avoid “both high and low”**.

Example: 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 500 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
- number of heads
- (% heads - 50%)
- (# heads - 250)
- abs(% heads - 50%)
- or abs(#heads - 250)

To choose a **test statistic**, look at the alternative hypothesis.

- If the alternative is “**the null is wrong**” then
 - If comparing counts or means:
 - Use **Absolute Value** of Difference
 - If comparing proportions:
 - Use **Total Variation Distance (TVD)**
- If the alternative specifies a **direction** (e.g. “fewer than”)
 - Use a count, or proportion or average, or difference from null

p-values

Steps in Hypothesis Testing

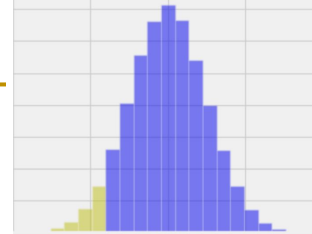
- Choose null and alternative hypotheses
- Choose a significance level
- Choose a test statistic
- Gather data and calculate observed test statistic
- Simulate/theoretically determine distribution of test statistic under null hypothesis
- **Calculate p-value**
- Make conclusion



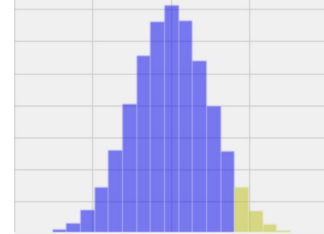
Video

<https://www.youtube.com/watch?v=9jW9G8M04PQ>

Definition of the p -value



Simulated values when **LOW** test statistics support the alternative hypothesis



Simulated values when **HIGH** test statistics support the alternative hypothesis

The p -value is the chance (probability),

- under the null hypothesis (i.e. given the null)
 - 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.
- Yellow area denotes the p -value
 - Red dot denotes the observed statistic.

Formal name: **observed significance level**

Notice: The p -value is actually a Conditional Probability!

$p\text{-value} = P(\text{observed data or more extreme} \mid \text{null hypothesis})$

p -value is large \rightarrow evidence of consistency with the null (fail to reject null)

p -value is small \rightarrow more evidence for the alternative (reject null)

P-value is based on tail area

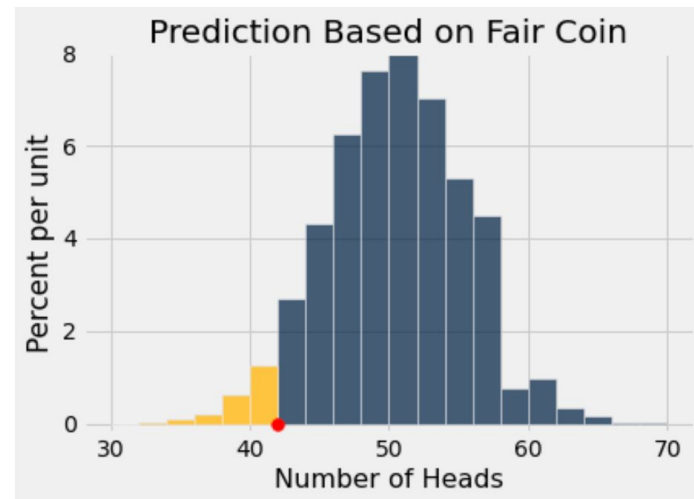
- Start at the observed value of the test statistic
- **Look in the direction that favors the alternative hypothesis**
 - If that tail is small, the data are not consistent with the null
 - Otherwise, the data are consistent with the null (fail to reject null)

Ex: Suppose we want to test whether or not a coin is biased toward tails

- **Null:** The coin is fair
- **Alternative:** The coin is biased towards tails
- **Statistic:** Number of heads (or heads - tails or tails or ...)

Fill in the blanks:

- _____ values of the number of heads favor the alternative
- So start at the observed number of heads and look to the _____



Theoretical p-value vs Empirical p-value

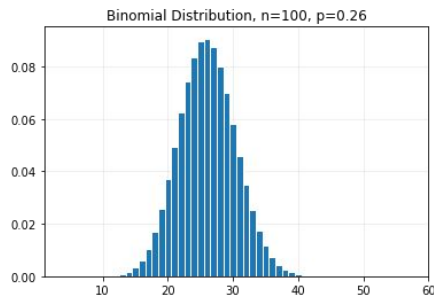
Recall: any random variable has a distribution:

Probability (aka Population or Theoretical)

Distribution 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");
```



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

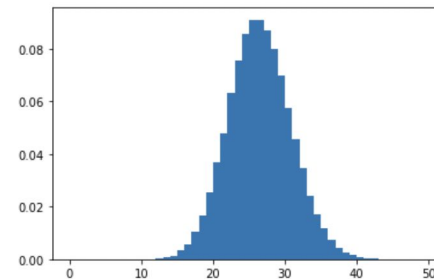
- **Empirical (aka Simulated or Sample) Distribution:** based on random samples (or simulations)
- Observations can be from **repetitions of an experiment or random samples from a population**
 - All observed values
 - The proportion of times each value appears

```
#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);
```



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**.

Which of the following does the p-value depend on?

Select all that apply

- Null hypothesis
- Alternative hypothesis
- The choice of test statistic
- The data in the sample
- The significance level (e.g. 5%)

Which of the following does the p-value depend on?

Select all that apply

- Null hypothesis
- Alternative hypothesis
- The choice of test statistic
- The data in the sample
- The significance level (e.g. 5%)

Answer: All except the significance level

True or False:

- A p-value is the probability that your null hypothesis is true
- A p-value is the probability that your alternative hypothesis is false
- A p-value is the probability that the observed effects was produced by random chance alone.
- A small p-value means you have a large effect size.
- A small p-value means you have a large effect size.

Conclusion of test

Steps in Hypothesis Testing

- Choose null and alternative hypotheses
- Choose a significance level
- Choose a test statistic
- Gather data and calculate observed test statistic
- Simulate/theoretically determine distribution of test statistic under null hypothesis
- Calculate p-value
- **Make conclusion**

Hypothesis Test Conclusions

Recall: Before we started the hypothesis test we needed: **A Default course of action**

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 the null hypothesis is true for sure*

Conclusion of Hypothesis Test:

- If $p\text{-value} \leq$ your predetermined significance level
 - Reject null (data is consistent with the alternative hypothesis)
- Else
 - Fail to reject null hypothesis (data is consistent with the null hypothesis)

If data convinces you live in the alternative hypothesis world, ***change your action!***

No reason to change your mind? Proceed with the default action as planned. Is it the right action?
~_(\ツ)/~ Welcome to uncertainty.

Hypothesis Test Conclusions: Caveats

- Hypothesis tests cannot PROVE whether or not a hypothesis is true.

Whether you use a conventional cutoff for the significance level 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.