Hypothesis Testing

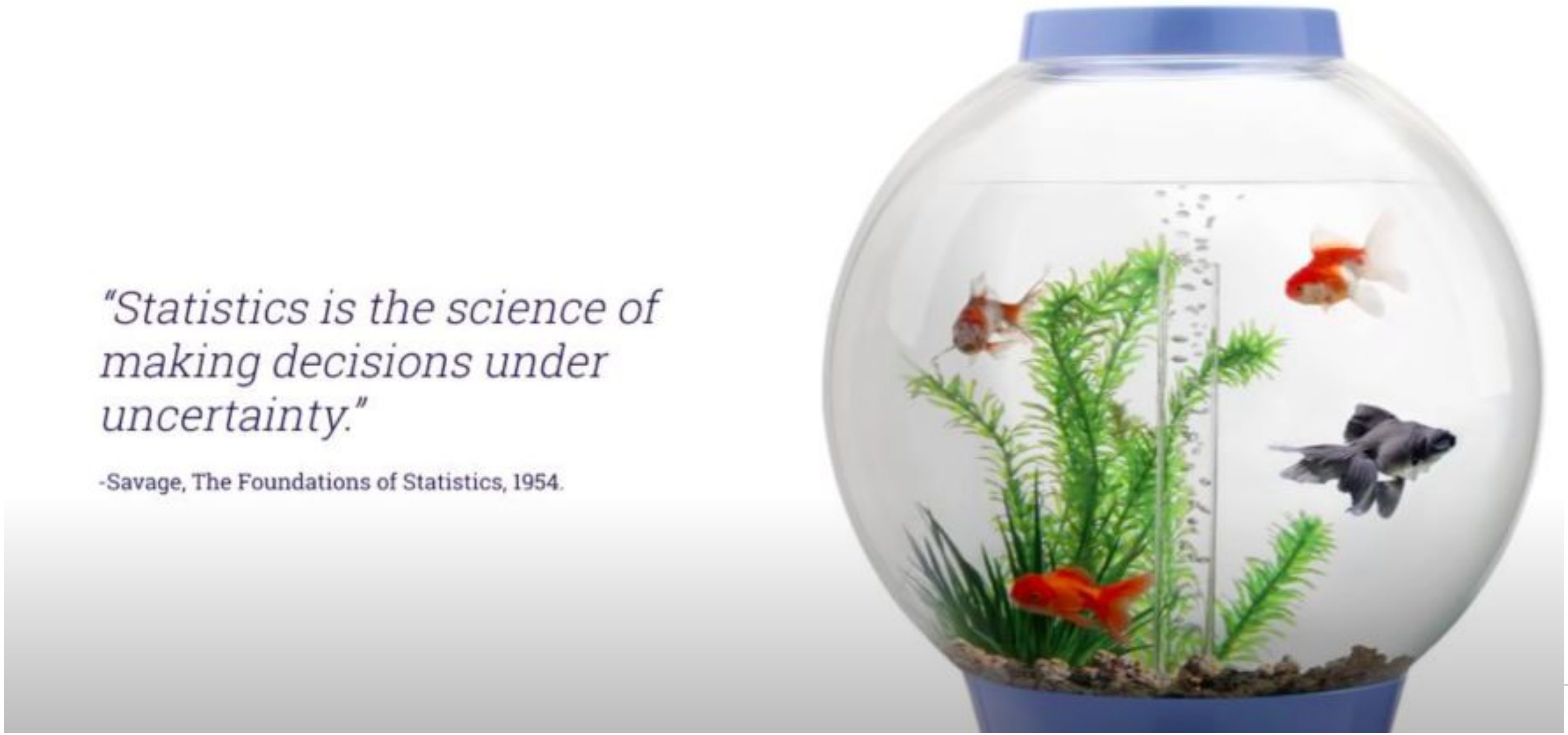
LECTURE 20

Announcements

HW 7 due tonight

HW 8 Released Tonight, due next Thursday

Today’s Gameplan



Statistical Testing

To begin, you need:

Default action
(Frequentist)

OR

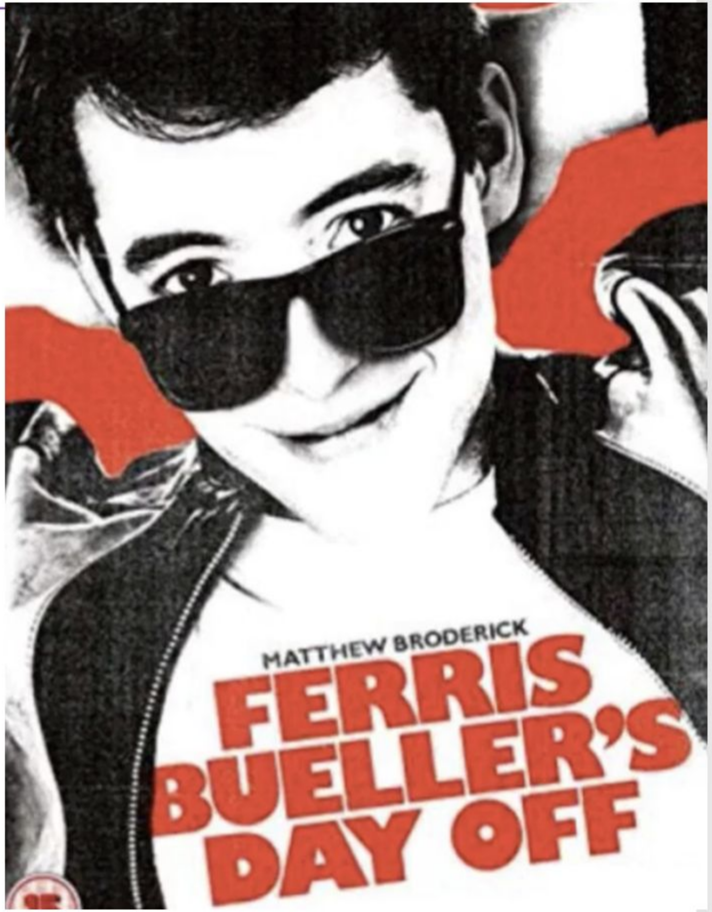
Prior opinion
(Bayesian)

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 were generated

- The views are called **hypotheses**

Null and Alternative

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:

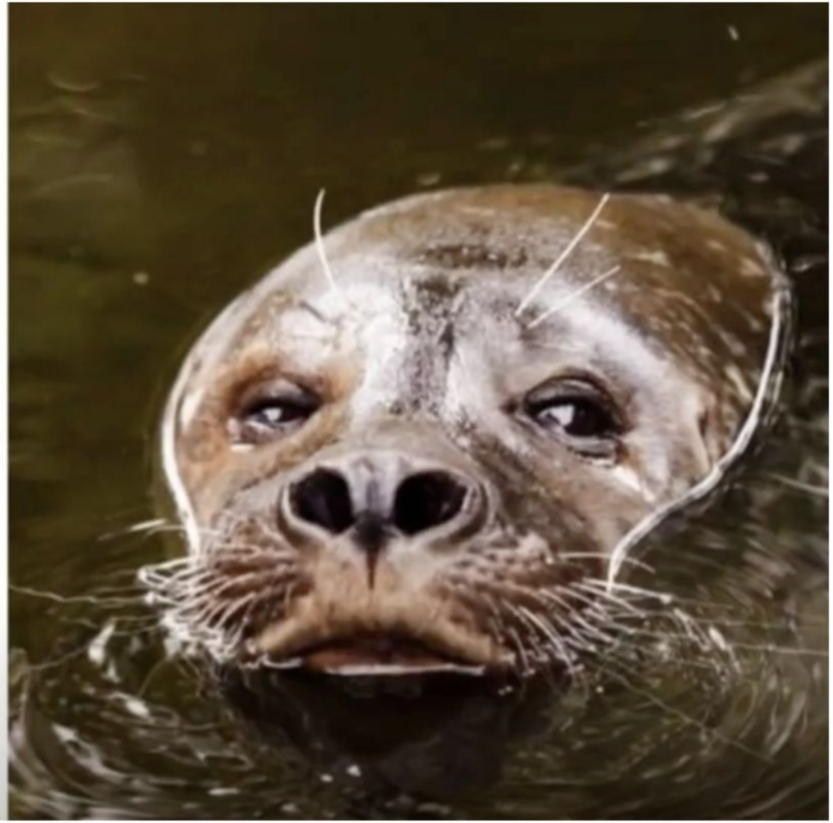
You haven't got any data
OR
You know very little
OR
Null Hypothesis is true for sure

Statistical Testing

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

You need to be convinced (with data!) that the

null hypothesis looks ridiculous




I have no idea how
it happened...

Default action:
Don't shout at Fido.

Null hypothesis:
Fido is innocent.

Hypothesis testing:

Ridiculous? [Y/n]

Default action:
Don't shout at Fido.

Null hypothesis:
Fido is innocent.

We use the math to make a model of a world...

Hypothesis testing:

`Ridiculous? [Y/n]`

...so we can ask it how weird our evidence is.

Testing Hypotheses

- A test chooses between two views of how data were generated

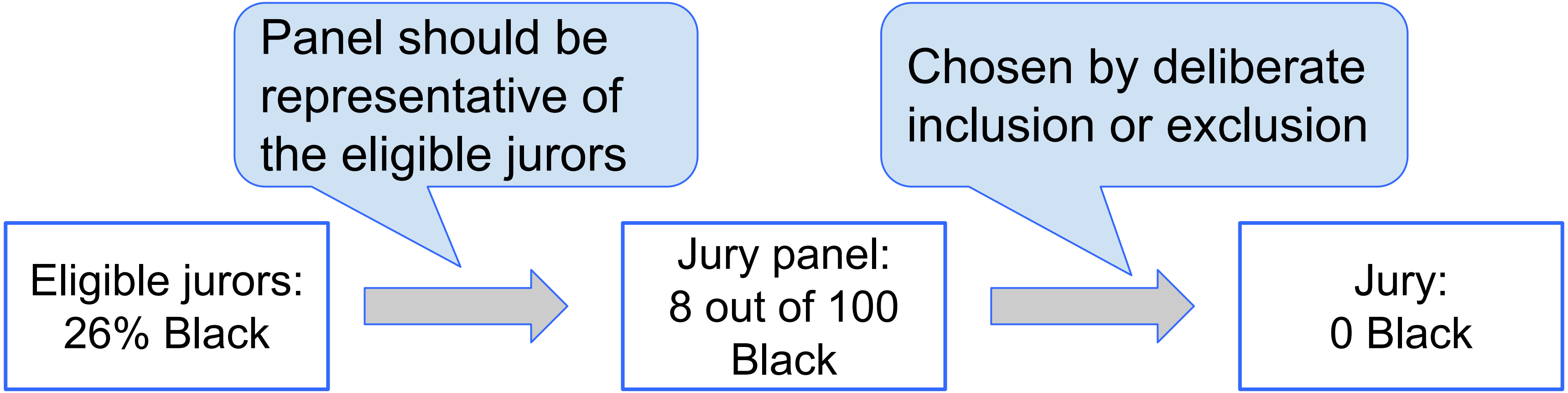
- The views are called **hypotheses**

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

  - **"Null" Hypothesis:** The people on the jury panels were selected at random from the eligible population
  - **"Alternative" Hypothesis:** No, they were biased against black people

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

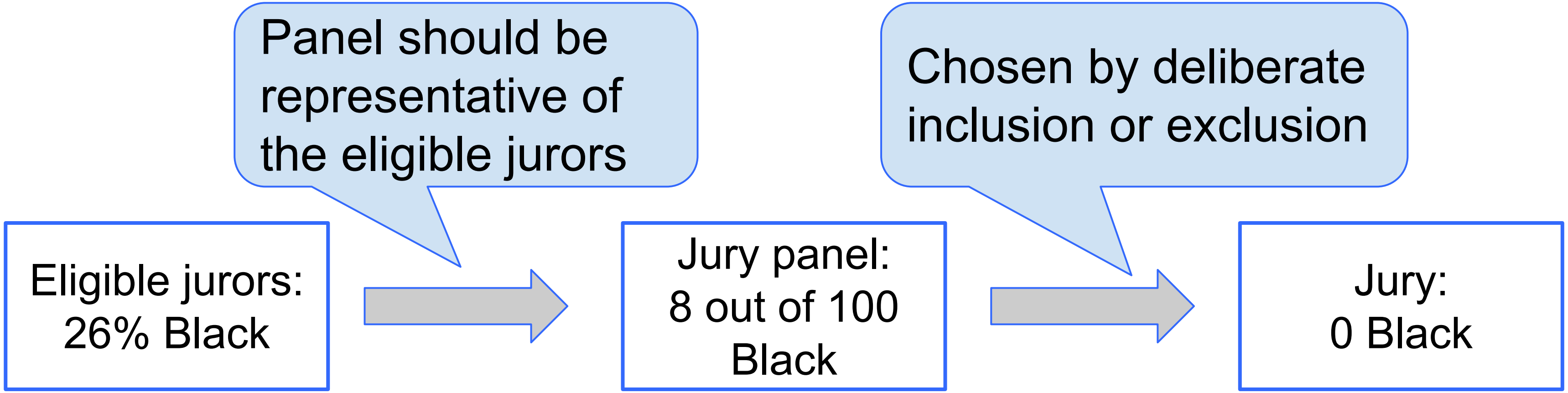


Hypothesis Testing

- **Choose a statistic** to measure “discrepancy” between null hypothesis and data
- **Simulate the statistic (or calculate directly when possible)** under the null assumptions
- **Compare** the data to the null hypothesis predictions:
  - Draw a histogram of (simulated) values of the statistic
  - Compute the observed statistic from the real sample
- If the observed statistic is in the tail* of the empirical distribution, we reject the null hypothesis.
-

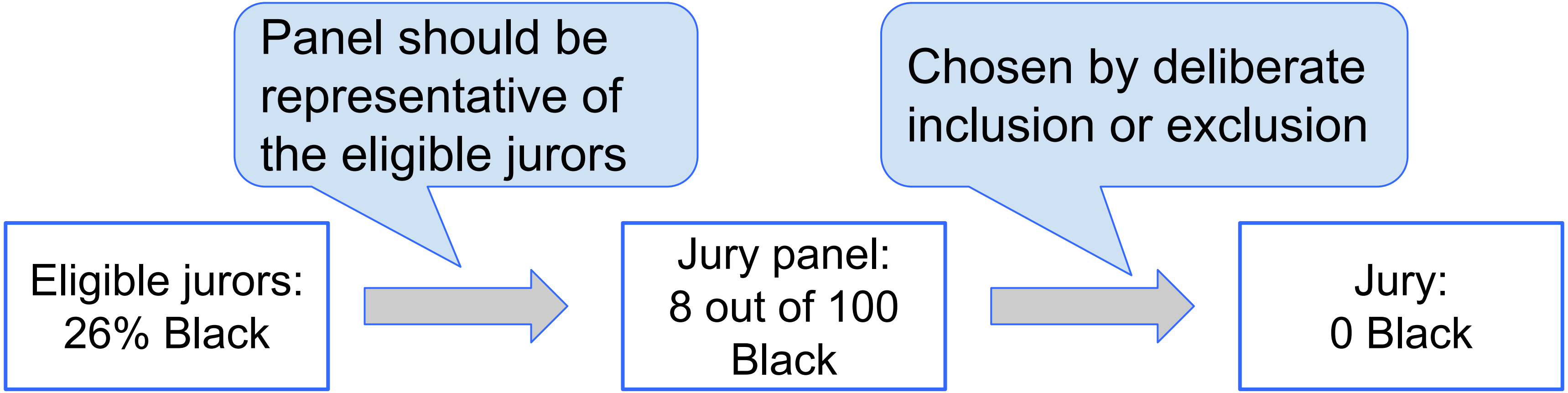
Robert Swain’s Case

- Null Hypothesis: The people on the jury panels were selected at random from the eligible population where 26% of people are black (i.e. Binomial distribution, with $p=0.26$)
- Alternative Hypothesis:No, they were biased against black men
- Test Statistic: ??
- Observed test statistic: ??



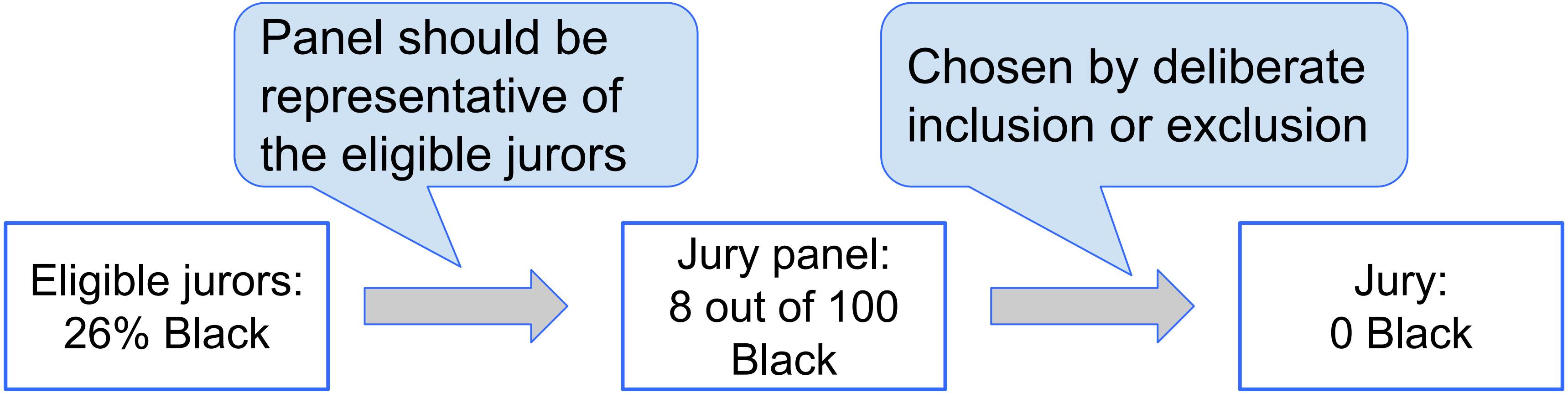
Robert Swain's Case

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



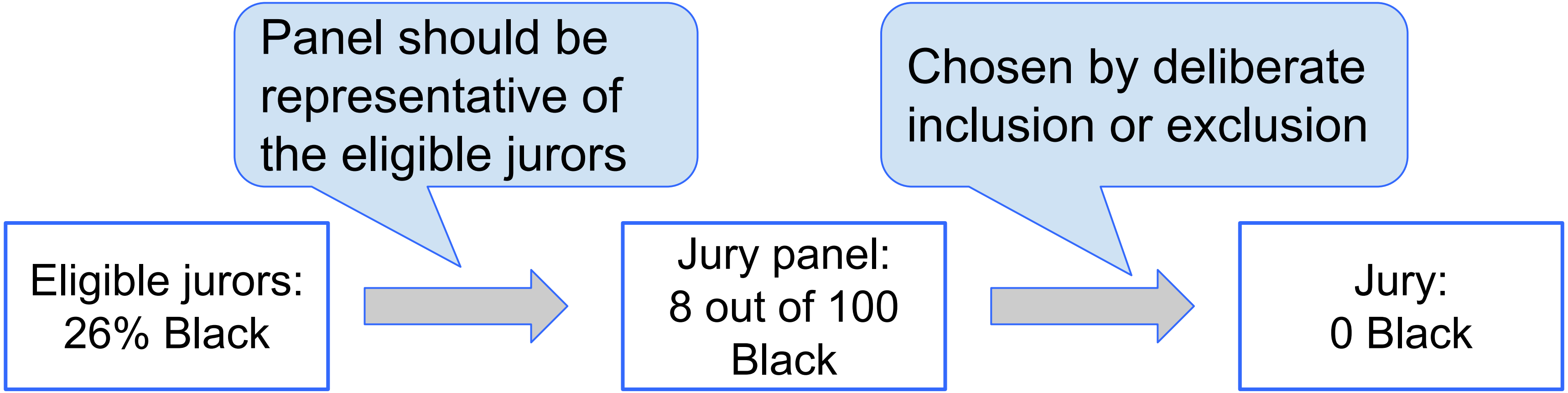
Robert Swain's Case

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



Robert Swain’s Case

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

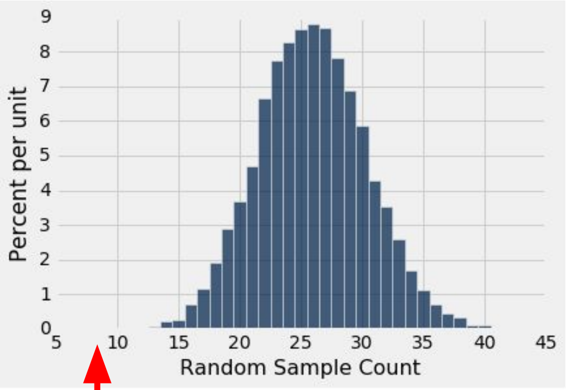


Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical 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)**
- The probabilities are approximate, because we can't generate all the possible random samples

Demo: Tail Areas

Alabama Jury



Observed Number (8)

Recap: Robert Swain's Case

- Null Hypothesis: The people on the jury panels were selected at random from the eligible population where 26% of people are black (i.e. Binomial distribution, with p=0.26)
- Alternative Hypothesis:No, they were biased against black men
- 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.

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

Conclusion: Our p-value = ______ which is less than _______ , so we __________null and say result is "highly statistically significant"

Conventions About Inconsistency

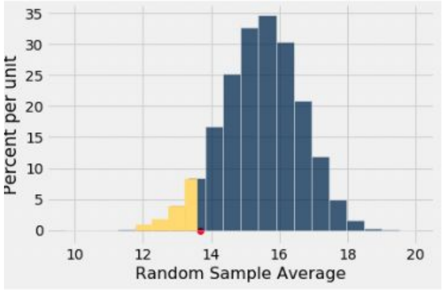
- **“Inconsistent with the null”:** The observed test statistic is in the tail of the empirical distribution under the null hypothesis

- **“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”

Definition of the *P*-value

Formal name: **observed significance level**

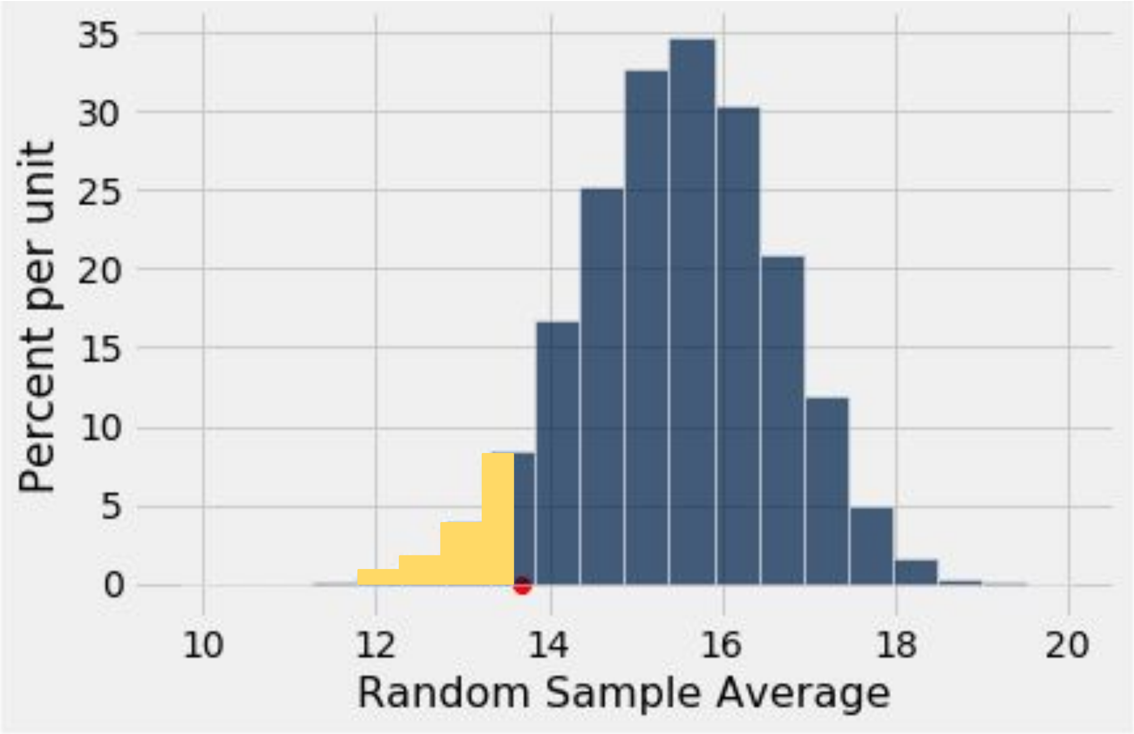


The *P*-value is the chance,

- 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.

The P-Value as an Area

Empirical distribution of the test statistic under the null hypothesis



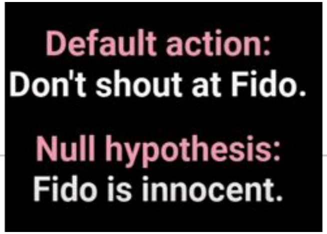
The red dot is the observed statistic.

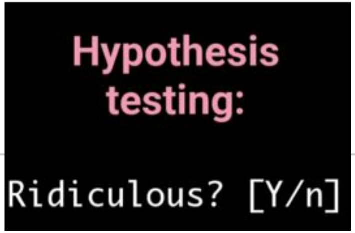
Lesson 20: Day 2 Announcements

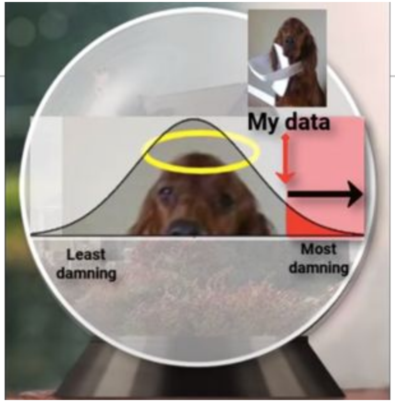
HW 8 due Thursday

Quiz 7 Friday

- Scope: HW 7;
- L17: Joint Distributions; Covariance/Correlation & Independence
- L18: Sampling







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

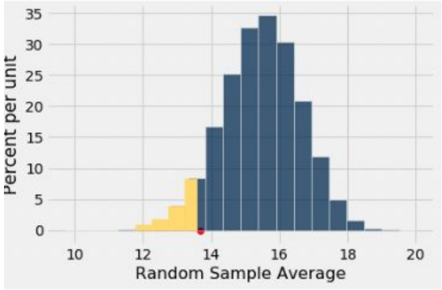
> A p-value doesn't *prove* anything. It's simply a way to use surprise as a basis for making a reasonable decision.
>
> — Cassie Kozyrkov

Hypothesis Testing

- **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 statistic** to measure “discrepancy” between null hypothesis and data
- **Simulate the 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 from the real sample
- If the observed statistic is in the tail* of the empirical distribution, we reject the null hypothesis (calculate the p-value to determine this)

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.

Conclusion of the Test

Determine whether observed test statistic is consistent null hypothesis:

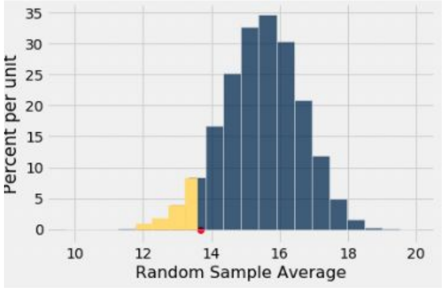
- If p-value is less than your chosen significance level:
  - Reject the null hypothesis in favor of the alternative
  - Else: Fail to reject the null hypothesis

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"



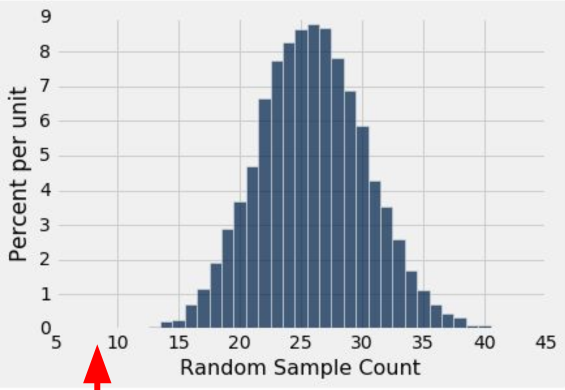
Hypothesis Testing Review

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.

Recap: Robert Swain Ex

Alabama Jury



Observed Number (8)

Ex 2: Comparing Distributions

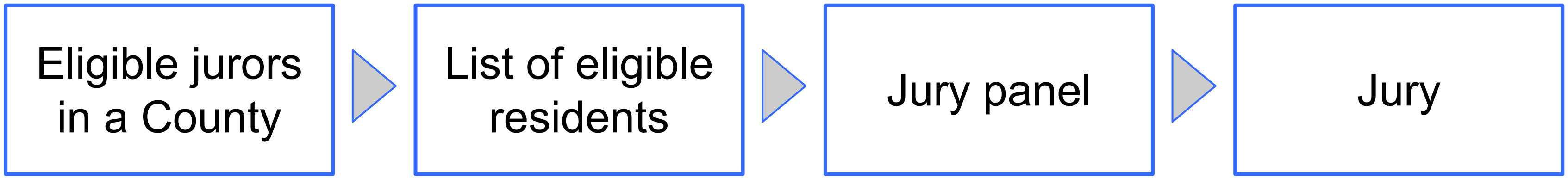
Jury Selection in Alameda County

Racial and Ethnic Disparities in Alameda County Jury Pools

A Report by the ACLU of Northern California

October 2010

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."

(Demo)

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”. Try to **avoid “both high and low”**.

A New Statistic

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

(Demo)

Total Variation Distance

Every distance has a computational recipe

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

(Demo)

Summary of the Method

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

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

(Demo)

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:

  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)

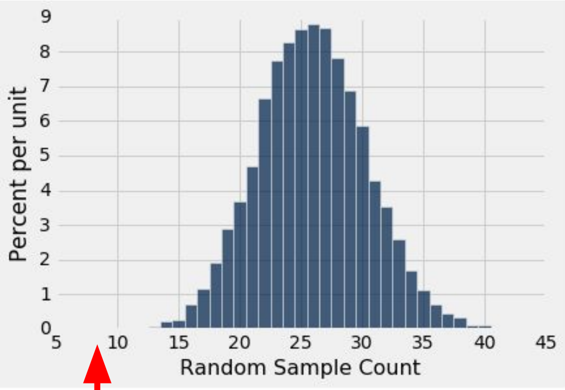
Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** 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")

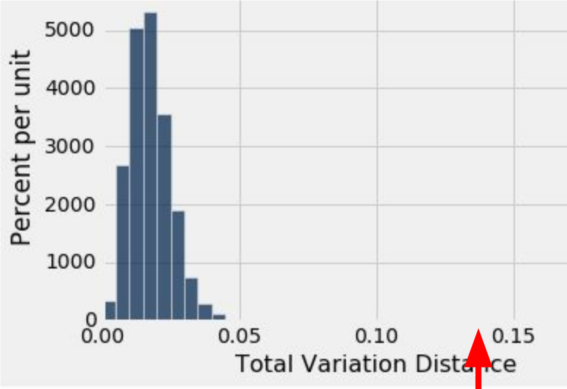
Tail Areas

Alabama Jury



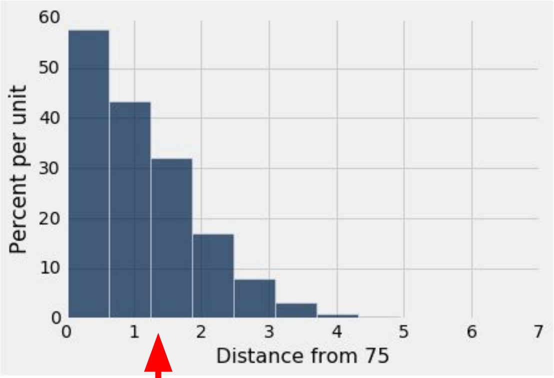
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Pea Plants



Observed Distance (1.32)

Discussion Questions

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)

- “This coin is fair.”
- “No, it’s biased towards heads.”

(b)

- “This coin is fair.”
- “No, it’s not.”

“Fair”

For both (a) and (b),

- The percent of heads in the 400 tosses is a good starting point, but might need adjustment

- A percent of heads around 50% suggests “fair”

Answers

(a) **Large** values of the percent of heads suggest “biased towards heads”

- Statistic: percent of heads

(b) 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
- Statistic: | percent of heads − 50% |
- Large values of the statistic suggest “not fair”

Ex 4: Another Example

- Large(-ish) Calculus class divided into 12 recitation sections
- TA's lead the sections

- After the midterm, students in Recitation 3 notice that the average score in their section is lower than in others

The TA's Defense

TA's position (Null Hypothesis):

- If we had picked my section at random from the whole class, we could have got an average like this one.

Alternative:

- No, the average score is too low. Randomness is not the only reason for the low scores.

(Demo)

Hypothesis Testing Review

- **One Category** *(e.g. percent of flowers that are purple)*
  - Test Statistic (1): `observed_proportion`
  - Test Statistic (2): `abs(observed_proportion - null_proportion)`
  - How to Simulate: `np.random.binomial(N, null_hyp)`
- **Multiple Categories** *(e.g. ethnicity distribution of jury panel)*
  - Test Statistic: `tvd(observed_distribution, null_distribution)`
  - How to Simulate: `np.random.multinomial(N, null_hyp)`
- **Numerical Data** *(e.g. scores in a lab section)*
  - Test Statistic: `observed_mean`
  - How to Simulate: `population_df.sample(n, with_replacement=False)`