# MODULE 6

Testing Hypotheses & Comparing Two Samples

# ASSESSING MODELS

# MODELS

- A model is a set of assumptions about the data

- In data science, many models involve assumptions about processes that involve randomness
  - "Chance models"

- **Key question:** does the model fit the data?

# APPROACH TO ASSESSMENT

## I.E., HOW CAN WE TELL IF A MODEL IS GOOD?

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts.

- We can then compare the predictions to the data that were observed.

- If the data and the model's predictions are not consistent, that is evidence against the model.

# EX. 1: JURY SELECTION

# SWAIN VS. ALABAMA, 1965

- Talladega County, Alabama
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Only men 21 years or older were allowed to serve
- 26% of this population were black
- Swain's jury panel consisted of 100 men
- 8 men on the panel were black

# SUPREME COURT RULING [IN ENGLISH]

- About disparities between the percentages in the *eligible population* and *the jury panel*, the Supreme Court wrote:

"… the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes"

- The Supreme Court denied Robert Swain's appeal

# SUPREME COURT RULING [IN DATA]

- **Paraphrase:** 8/100 is less than 26%, but not different enough to show Black men were systematically  excluded

- **Question:** is 8/100 a realistic outcome if the jury panel selection process were truly unbiased?

  - i.e., if the jury panel was selected at random

# SIMULATING JURY SELECTION 1

- First, we pick a statistic – one that can help us decide between the model and alternative views about the data.

  - What's the model? That the jury panel was picked at random (and it's just by chance that the number of black men was 8% rather than closer to 26%)

  - What's the alternative? It's what the Swain's appeal team used as their main point: the panel was not drawn at random because it contained too few black men.

- A natural statistic then is: number of black men in a simulated sample of 100 men representing the panel

# SIMULATING JURY SELECTION 2

- Next, we simulate sampling of the statistic

- We can use sample_proportions() to simulate one value of the statistic.
  - The sample size is 100, the size of the panel.
  - The distribution from which we will sample is the distribution in the population of eligible jurors.
    - Since 26% of them were Black, we will sample from the distribution specified by the proportions [0.26, 0.74].

# SAMPLING FROM A DISTRIBUTION

- Sample at random from a categorical distribution

**sample_proportions(sample_size, pop_distribution)**

- Samples at random from the population
  - Returns an array containing the distribution of the categories in the sample

(Demo – Notebook 6.1, Swain vs. Alabama)

# SIMULATING JURY SELECTION 3

- To get a sense of the variability we generate a large number of simulated values of the statistic (e.g., 10,000).

- Next, we interpret the results.
  - First, we visualize the simulated values using a histogram
  - Then, we compare the prediction (our simulation results) with the data
    - Let's plot the data value on the histogram

(Demo – Notebook 6.1, Swain vs. Alabama)

# CONCLUSION FROM THE SIMULATION

- The simulation shows that if we select a panel of 100 jurors at random from the eligible population, we are very unlikely to get counts of black men as low as the eight that were in Swain's jury panel.
  - This is evidence that the model of random selection of the jurors in the panel is not consistent with the data from the panel.

- When the data and a model are inconsistent, the model is hard to justify.
  - Why? the data is real, while the model is just a set of assumptions.

# ASSESSMENT OF THE MODEL - RESULTS

- While it is *possible* that a panel like Robert Swain's could have been generated by chance, our simulation demonstrates that it is **very unlikely**.

- Thus, our assessment is that **the model of random draws is not supported by the evidence**.
  - i.e., Swain's jury panel does not look like the result of random sampling from the population of eligible jurors.

# EX. 2: A GENETIC MODEL

# GREGOR MENDEL, 1822-1884
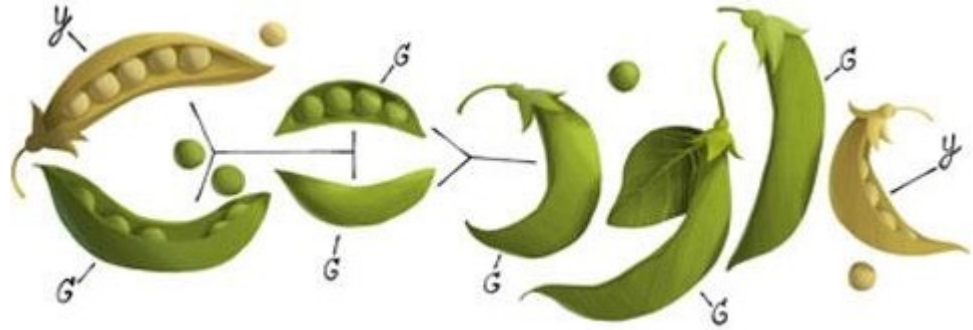
# A MODEL

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

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

- Question:
  - Is the model good, or not?

# 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:

  - | sample percent of purple-flowering plants - 75 |

- If the statistic is large, that is evidence against the model

# SIMULATING PURPLE FLOWERING PLANTS

- We have decided on a statistic

- Next, we simulate one outcome (i.e., one value of the statistic)
  - Sample 929 (total number of plants of this type that Mendel grew) times at random from the distribution specified by the model and find the sample proportion in the purple-flowering category.
  - Multiply the proportion by 100 to get a percent.
  - Subtract 75 and take the absolute value of the difference.

- Finally, run the simulation a large number of times and visualize the results in a histogram

(Demo – notebook 6.1, Mendel and Pea Flowers)

# ASSESSMENT OF THE MODEL - RESULTS

- To provide a final assessment we need to compare the prediction with the data

- The visualization of the prediction showed that the values of our statistic as small, which is desirable

- We have to plot the observed value of our statistic.
  - Of the 929 plants Mendel grew, 705 were purple flowering
  - Therefore, the observed value of the statistic = abs (100 * (705 / 929) - 75) = 0.888
  - We add that to our histogram (demo – notebook 6.1, Mendel and Pea Flowers)

- The observed value of the statistic is in the heart of the distribution predicted by Mendel's model.

- Therefore, our model is supported by Mendel's data (i.e., evidence)

# TWO VIEWPOINTS

## WHEN ASSESSING A MODEL

# MODEL AND ALTERNATIVE

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

- **Genetics:**
  - **Model:** Each plant has a 75% chance of having purple flowers
  - **Alternative viewpoint:** No, it doesn't

# STEPS IN ASSESSING A MODEL

- **Choose a statistic** to measure discrepancy between model and data

- **Simulate the statistic** under the model's assumptions

- **Compare** the data to the model's predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample

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

# 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 not."

(b)

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

# "FAIR"

For both (a) and (b),

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

- A number of heads around 200 suggests "fair"

# ANSWERS

(a) Very large or very small values of the number of heads  suggest "not fair."

- The **distance** between number of heads and 200 is  the key
- Statistic: | number of heads − 200 |
- Large values of the statistic suggest "not fair"

(b) Large values of the number of heads suggest "biased  towards heads"

- Statistic: number of heads

# ASSESSING MODELS WITH MULTIPLE CATEGORIES
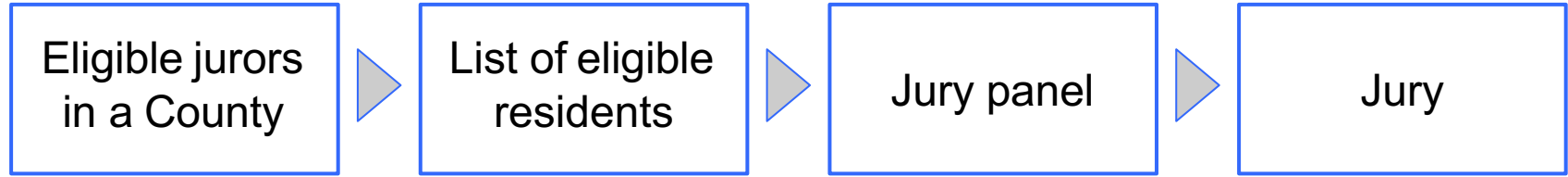
# Jury Selection in Alameda County

**Racial and Ethnic Disparities**
**in**
**Alameda County Jury Pools**

A Report by the ACLU of Northern California                      October 2010

# JURY PANELS

| Eligible jurors in a County | → | List of eligible residents | → | Jury panel | → | Jury |
|---|---|---|---|---|---|---|

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."

# ALAMEDA COUNTY JURY PANELS

- The ACLU of Northern California in 2010 presented a report on jury selection in Alameda county, CA.

- The focus of the study was ethnic composition of the panels (based on data collected in 2009 and 2010)

- The data is available in the table called jury, but it shows the following to be the ethnic composition of the jury panel

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

# COMPARISON WITH PANELS DRAWN AT RANDOM

- What if we select a random sample of 1,453 people from the population of eligible jurors?

- Will the distribution of their ethnicities look like the distribution of the panels above?

<div align="center">

Demo(Notebook 6.1, Alameda County Jury Panels)

</div>

- **Technical note.**
  - Random samples of prospective jurors would be selected without replacement.
  - However, when the size of a sample is small relative to the size of the population, sampling without replacement resembles sampling with replacement; the proportions in the population don't change much between draws.

# A NEW STATISTIC

# DISTANCE BETWEEN DISTRIBUTIONS

- People on the panels are of multiple ethnicities
- Distribution of ethnicities is categorical

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

# TOTAL VARIATION DISTANCE

Every 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 – notebook 6.1, Distance Between Distributions)

# SIMULATING ETHNIC COMPOSITION IN THE JURY

- Once we've obtained a statistic, here, the distance between 2 distributions
  - For the observed statistic that would be distance between eligible and observed panel
  - For the simulated statistic that would be distance between eligible and simulated panel

- We can simulate a single statistic
  - Generate an ethnic distribution (proportions) of the jury from a sample size of 1453
  - Then compute the TVD between that distribution and eligible distribution

- Then, iterate several times (to understand the variability of the simulated statistic)
  - Append the obtained values in an array

- Then visualize the results (on a histogram)
  - And, compare the predictions (empirical TVD of the simulated statistic) against the data (actual TVD from the data).

(Demo – notebook 6.1, Total Variation Distance)

# 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

# TESTING HYPOTHESES

# TESTING HYPOTHESES

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

- The views are called **hypotheses**

- The test picks the hypothesis that is better supported by the observed data

# NULL AND ALTERNATIVE

The method only works if we can simulate data 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

# 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".

# 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

# 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 more consistent with the alternative")

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- If not, there are conventions about "consistency"

# DECISIONS AND UNCERTAINTY

# INCOMPLETE INFORMATION

- We are trying to choose between two views of the world, based on data in a sample.

- It is not always clear whether the data are consistent with one view or the other.

- Random samples can turn out quite extreme. It is unlikely, but possible.

# ANOTHER EXAMPLE

# THE PROBLEM

- Large(-ish) Statistics class divided into 12 discussion sections
- Graduate Student Instructors (GSIs) lead the sections

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

# THE GSI'S DEFENSE

**GSI'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 – Notebook 6.2)

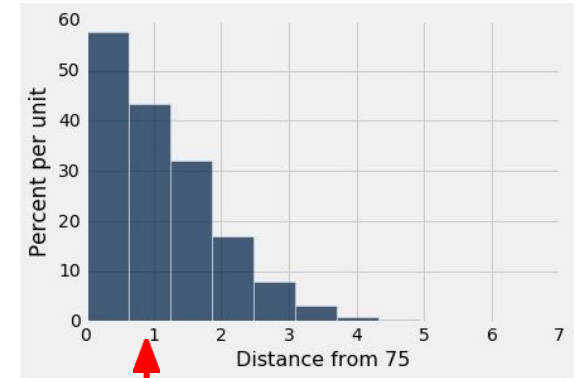# STATISTICAL SIGNIFICANCE

# TAIL AREAS
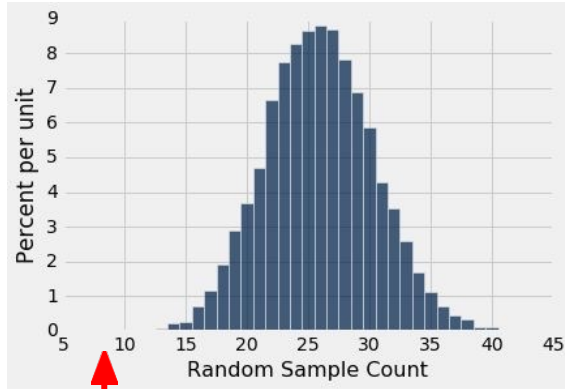


Alabama Jury — Observed Number (8)
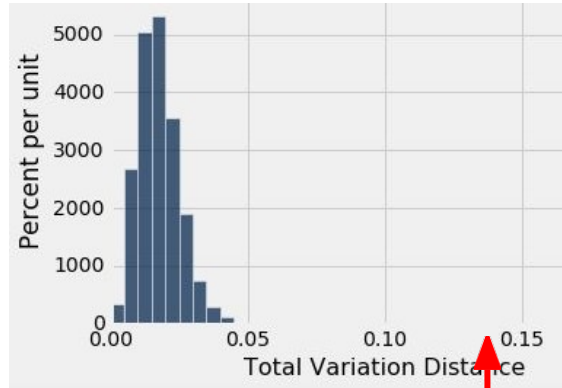
Alameda Jury — Observed TVD (0.14)

Pea Plants — Observed Distance (0.88)

# CONVENTIONS ABOUT INCONSISTENCY

- **"Inconsistent with the null":**
  - The 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:**

  (Demo - Notebook 6.2, Statistical Significance)
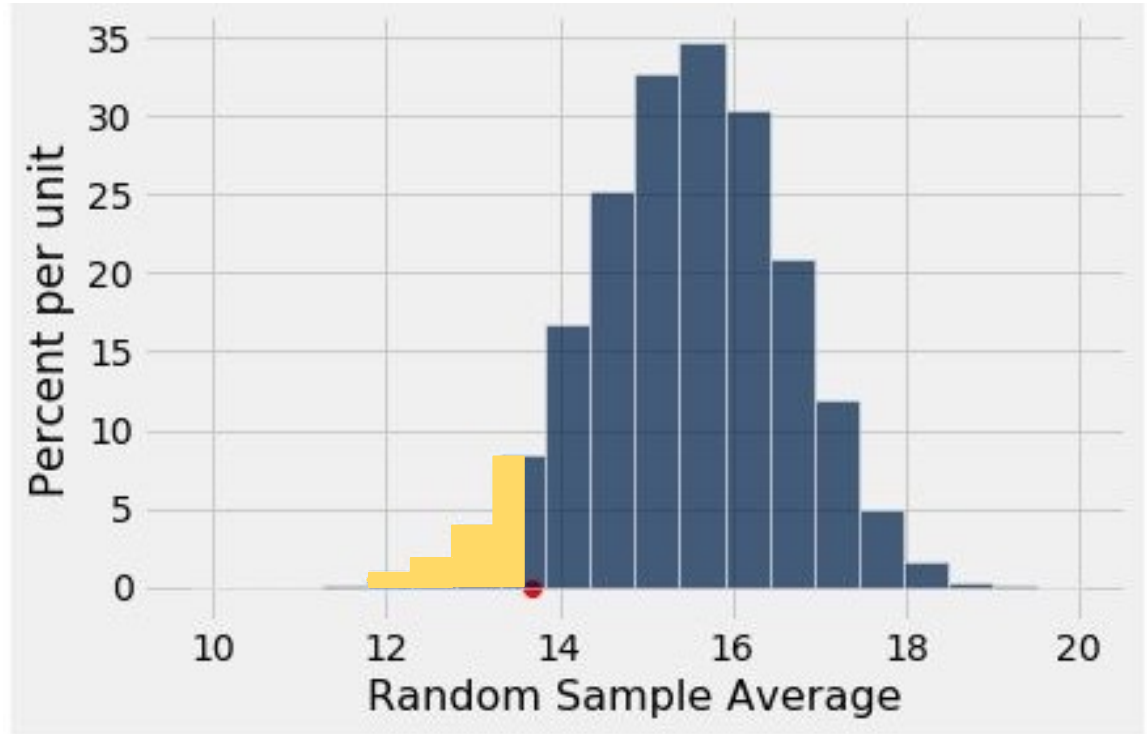
  - The area in the tail is less than 1%

  - The result is "highly statistically significant"

# THE P-VALUE AS AN AREA

Empirical distribution of the test statistic under the null hypothesis

The red dot is the observed statistic.

# 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.

# A/B TESTING

# HOW WE'VE TESTED THUS FAR

# HYPOTHESIS TESTING REVIEW

- **One Category** *(ex: percent of flowers that are purple)*
  - Test Statistic (1): empirical_percentage
  - Test Statistic (1): abs(empirical_percentage - null_percentage)
  - How to Simulate: sample_proportions(n, null_dist)

- **Multiple Categories** *(ex: ethnicity distribution of jury panel)*
  - Test Statistic: tvd(empirical_dist, null_dist)
  - How to Simulate: sample_proportions(n, null_dist)

- **Numerical Data** *(ex: scores in a lab section)*
  - Test Statistic: empirical_mean
  - How to Simulate: population_data.sample(n, with_replacement=False)

# COMPARING TWO SAMPLES

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.

- Question: Do the two sets of values come from the same underlying distribution?

- Answering this question by performing a statistical test is called **A/B testing**.

# THE GROUPS AND THE QUESTION

- Random sample of mothers of newborns. Compare:
  - (A) Birth weights of babies of mothers who smoked  during pregnancy
  - (B) Birth weights of babies of mothers who didn't  smoke

(Demo – Notebook 6.3, Comparing Two Samples)

# THE GROUPS AND THE QUESTION

- Random sample of mothers of newborns. Compare:
  - (A) Birth weights of babies of mothers who smoked  during pregnancy
  - (B) Birth weights of babies of mothers who didn't  smoke

- Question: Could the difference be due to chance alone?

# HYPOTHESES

- Null:
  - In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance)
- Alternative:
  - In the population, the babies of the mothers who smoked weigh less, on average, than the babies of the non-smokers.

# TEST STATISTIC

- Group A: non-smokers
- Group B: smokers

- Statistic: Difference between average weights
  - i.e., Group B average - Group A average

- Negative values of this statistic favor the alternative

(Demo – Notebook 6.3, Test statistic)

# THE DATA



| Non-smoker | Non-smoker | Smoker | Smoker | ... | Non-smoker |
|:---:|:---:|:---:|:---:|:---:|:---:|
| 120 oz | 113 oz | 128 oz | 108 oz | ... | 117 oz |

# SHUFFLING LABELS UNDER THE NULL



| Smoker | Non-smoker | Non-smoker | Smoker | ... | Smoker |
| 120 oz | 113 oz | 128 oz | 108 oz | ... | 117 oz |

# SHUFFLING ROWS

# RANDOM PERMUTATION

- **tbl.sample(n)**
  - Table of n rows picked randomly with replacement
- **tbl.sample()**
  - Table with same number of rows as original **tbl**, picked randomly with replacement
- **tbl.sample(n, with_replacement = False)**
  - Table of n rows picked randomly without replacement
- **tbl.sample(with_replacement = False)**
  - All rows of tbl, in random order

(Demo – Notebook 6.3, Random Permutation (Shuffling)

# SIMULATING UNDER THE NULL

- If the null is true, all rearrangements of labels are equally likely
- Plan:
  - Shuffle all group labels
  - Assign each shuffled label to a birth weight
  - Find the difference between the averages of the two shuffled groups
  - Repeat (i.e, iterate to get a sense of variability of the simulated value of the test statistic)

(Demo, Notebook 6.3,

Simulation under null hypothesis **and** permutation test)

# HOW WE'VE TESTED THUS FAR

# HYPOTHESIS TESTING REVIEW

- **1 Sample: One Category** *(e.g. percent of flowers that are purple)*
  - Test Statistic: empirical_percent, abs(empirical_percent -null_percent)
  - How to Simulate: sample_proportions(n, null_dist)

- **1 Sample: Multiple Categories** *(e.g. ethnicity distribution of jury panel)*
  - Test Statistic: tvd(empirical_dist, null_dist)
  - How to Simulate: sample_proportions(n, null_dist)

- **1 Sample: Numerical Data** *(e.g. scores in a lab section)*
  - Test Statistic: empirical_mean, abs(empirical_mean - null_mean)
  - How to Simulate: population_data.sample(n, with_replacement=False)

- **2 Samples: Numerical Data** *(e.g. birth weights of smokers vs. non-smokers)*
  - Test Statistic: group_a_mean - group_b_mean,
    group_b_mean - group_a_mean, abs(group_a_mean - group_b_mean)
  - How to Simulate: empirical_data.sample(with_replacement=False)

# IMPORTANCE OF RANDOM ASSIGNMENT



Apple users more willing to pay for apps
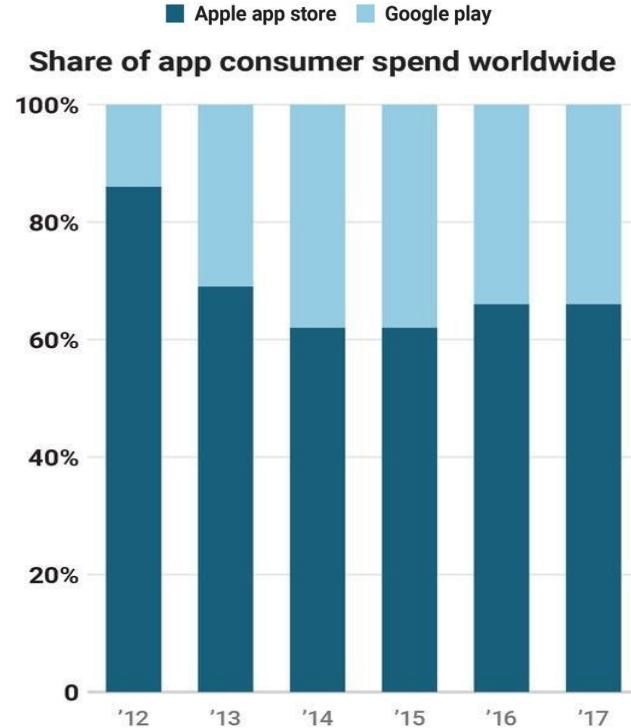
# IMPORTANCE OF RANDOM ASSIGNMENT

- iOS users spend 2x as much as Android users on 3rd party apps
  - *Is **higher spending** <u>caused by</u> users owning **iPhone**?*
  - Can't Tell:
    - Users aren't randomly assigned a phone
    - Other factors contribute to their phone purchasing decisions (e.g. income, geography)
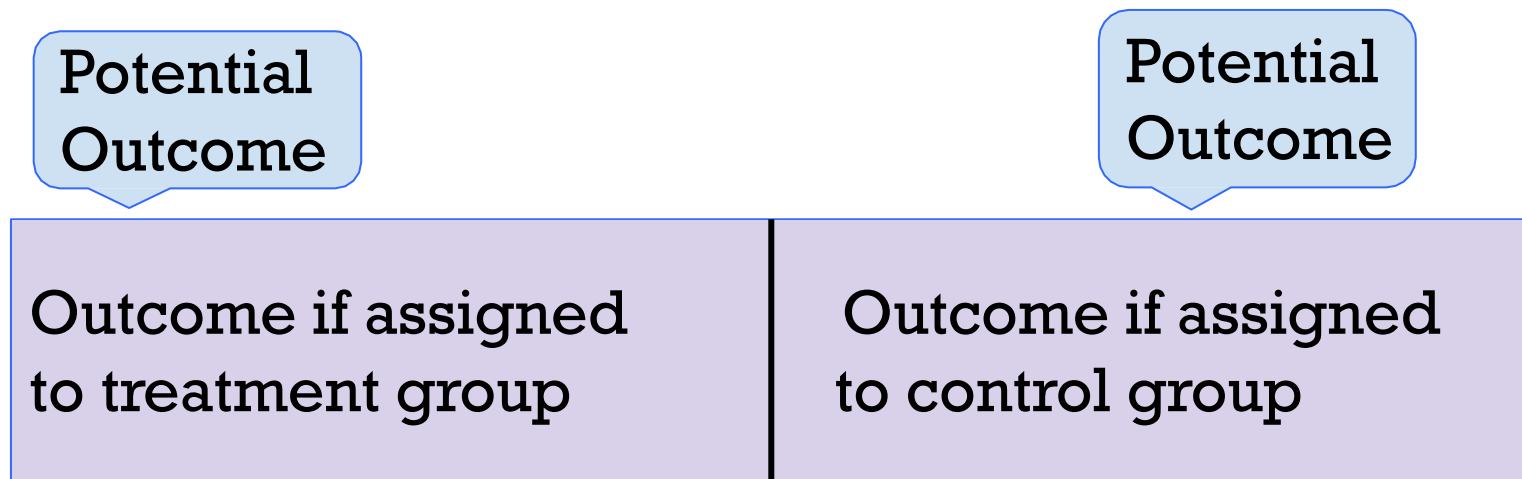
# CAUSALITY

# RANDOMIZED CONTROLLED EXPERIMENT

- Sample A: **control group**
- Sample B: **treatment group**

- **If the treatment and control groups are selected at random, then you can make causal conclusions.**

- Any difference in outcomes between the two groups could be due to:
  - chance
  - the treatment

(Demo – Notebook 6.4,
Randomized Control Experiment)

# BEFORE THE RANDOMIZATION

- In the population there is one imaginary ticket for each of the 31 participants in the experiment.
- Each participant's ticket looks like this:

Potential Outcome

Potential Outcome

| Outcome if assigned to treatment group | Outcome if assigned to control group |

# THE DATA

16 randomly picked tickets show:

| | Outcome if assigned to control group |
|---|---|

The remaining 15 tickets show:

| Outcome if assigned to treatment group | |
|---|---|

# THE HYPOTHESES

- **Null:**
  - In the population, the distribution of **all potential control scores** is the same as the distribution of **all potential treatment scores**.
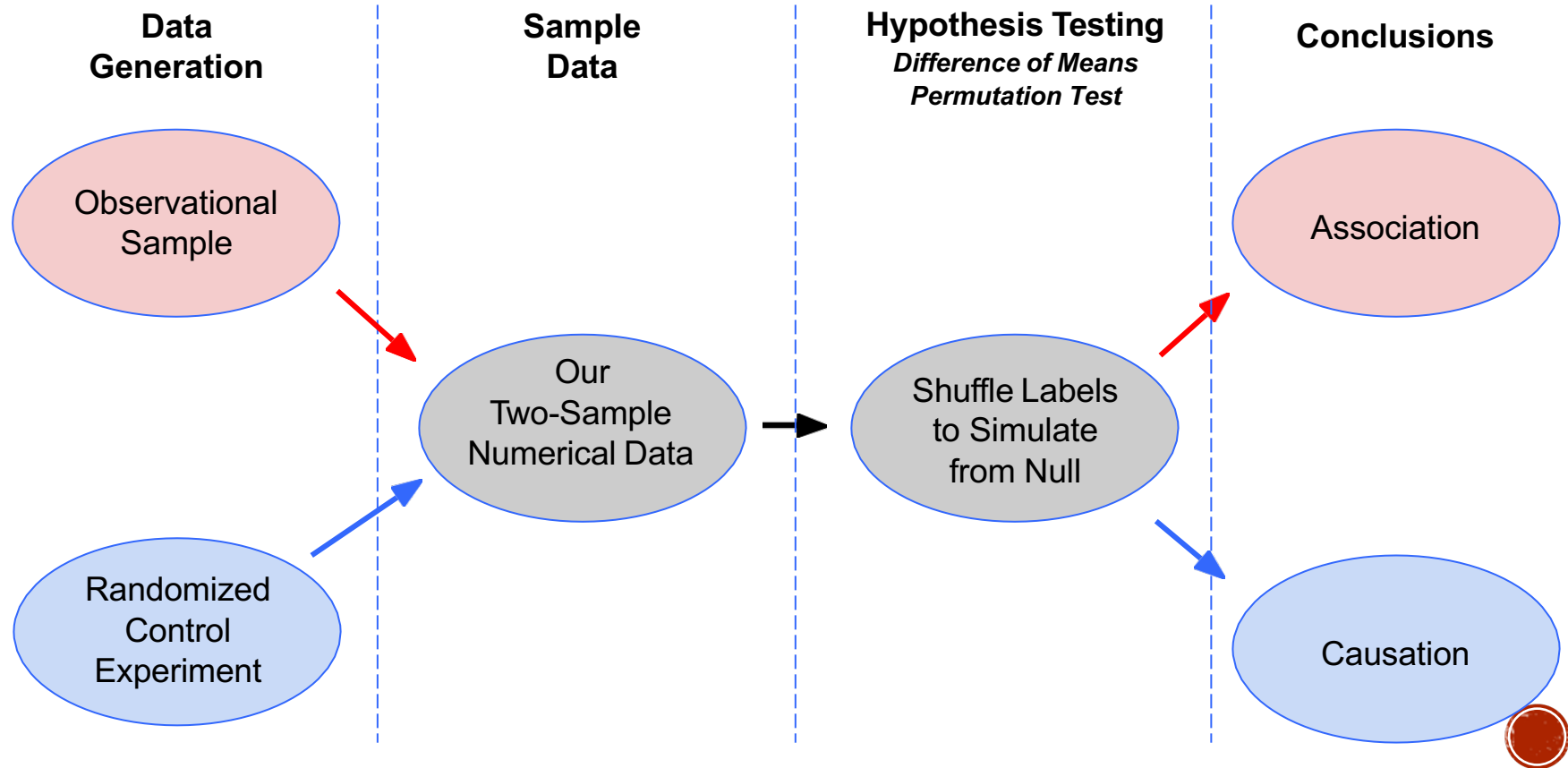  - tl;dr the treatment has no effect

- **Alternative:**

  - In the population, **more of the potential treatment scores are 1** (pain improves) than the **potential control scores**.

# RANDOM ASSIGNMENT & SHUFFLING

**Data Generation**

**Sample Data**

**Hypothesis Testing**
*Difference of Means Permutation Test*

**Conclusions**

Observational Sample

Randomized Control Experiment

Our Two-Sample Numerical Data

Shuffle Labels to Simulate from Null

Association

Causation

# AN ERROR PROBABILITY

# CAN THE CONCLUSION BE WRONG?

**Yes.**

|  | Null is true | Alternative is true |
|---|---|---|
| **Test favors the null** | ✅ | ❌ |
| **Test favors the alternative** | ❌ | ✅ |

# AN ERROR PROBABILITY

- The cutoff for the *P*-value is an error probability.

- If:
    - your **cutoff is 5%**
    - and the **null hypothesis happens to be true**

- then there is about a **5% chance** that **your test will reject the null hypothesis**.

# P-VALUE CUTOFF VS P-VALUE

- P-value cutoff
  - Does not depend on observed data or simulation
  - Decide on it before seeing the results
  - Conventional values at 5% and 1%
  - Probability of hypothesis testing making an error
- P-value
  - Depends on the observed data and simulation
  - Probability under the null hypothesis that the test statistic is the observed value or further towards the alternative

# TESTING HYPOTHESES - CONCLUSION

# HOW TO DO A HYPOTHESIS TEST

- **Before computing anything:** figure out the viewpoint the question wants to test, and formulate:
    - **Null hypothesis**: Completely specified chance model under which you can simulate data
    - **Alternative hypothesis**: Viewpoint from the question
    - **Test statistic**: to help you choose one viewpoint
- Compute the value of the test statistic in your data
- Simulate the test statistic under the null many times
- Compare the results

# QUESTIONS?