A/B Testing, Randomized Control Tests and Causality

LECTURE 21

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Announcements

- Homework 8 due tomorrow night

- Quiz 7 Friday

  Scope: HW 7;

  L17: Joint Distributions; Covariance/Correlation & Independence

  L18: Sampling

  ○

Today's Roadmap

CSCI 3022

- A/B testing
- Permutation Tests
- Randomized Control Tests
- Causality
- Errors in Hypothesis Testing

Hypothesis Testing Review

- 1 Sample: Two Categories *(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)`
- 1 Sample: 3 or More Categories *(e.g. ethnicity distribution of jury panel)*
  - Test Statistic: `tvd(observed_distribution, null_distribution)`
  - How to Simulate: `np.random.multinomial(N, null_hyp)`
- 1 Sample: Numerical Data *(e.g. scores in a lab section)*
  - Test Statistic: `observed_mean`
  - How to Simulate: `population_df.sample(n, replace=False)`
- Today: Numerical Data from Comparing 2 Samples

Today: 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**.

A/B Testing

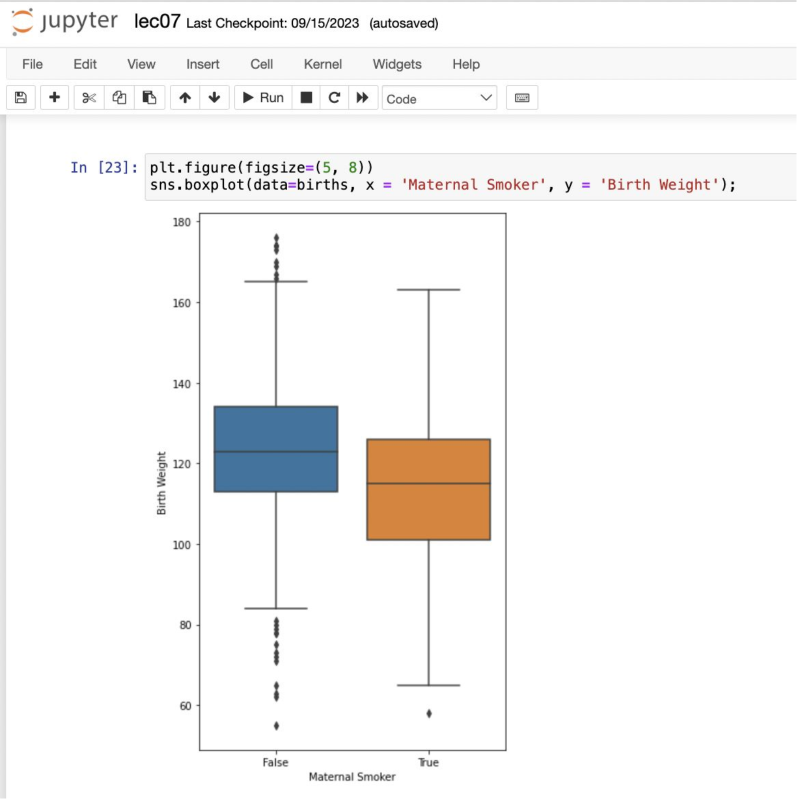
Steps in A/B Hypothesis Testing

- State Null Hypothesis:
- State Alternative Hypothesis:
- Choose Significance Level:
- Choose Test Statistic
- ***Use random permutations*** to simulate under the null hypothesis:
- Plot distribution of simulated null distribution
- Calculate Empirical p-value
- Make a concluding decision about test

The Groups and the Question

- Recall our example from Lecture 7: Random sample of mothers of newborns.

- Compare:
  - (A) Birth weights of babies of mothers who didn't smoke during pregnancy
  - (B) Birth weights of babies of mothers who did smoke



```
In [23]: plt.figure(figsize=(5, 8))
         sns.boxplot(data=births, x = 'Maternal Smoker', y = 'Birth Weight');
```

Question: Could the difference be due to chance alone?

Discuss: Null, Alternative and Test Statistic:

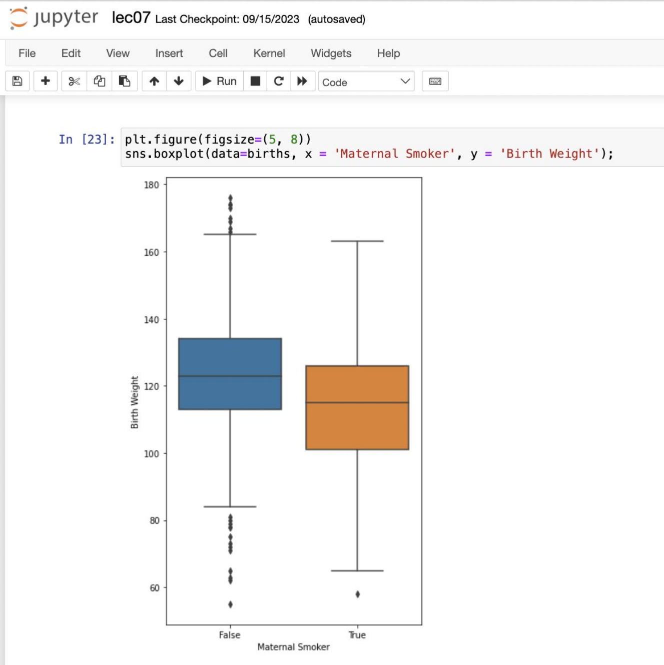
Poll: What is your test statistic?

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.

Choose Significance Level:

Alpha = 0.01



```
plt.figure(figsize=(5, 8))
sns.boxplot(data=births, x = 'Maternal Smoker', y = 'Birth Weight');
```

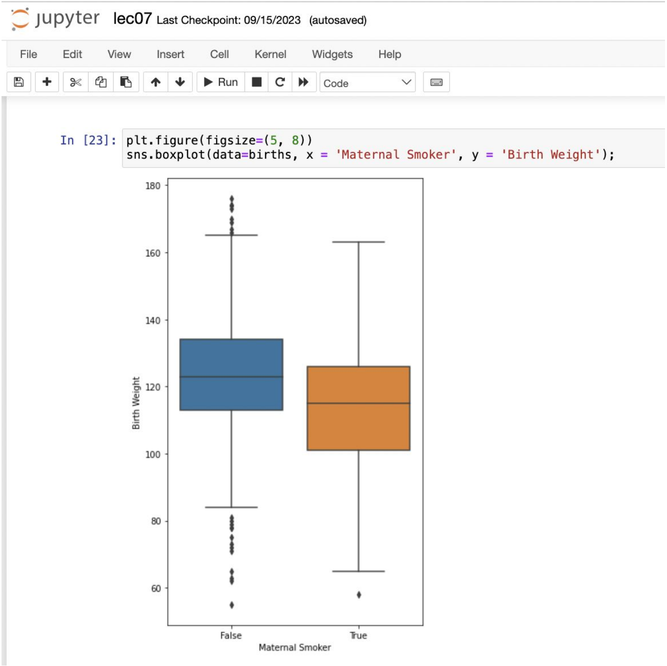
Test Statistic

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

- Statistic: Difference between average weights

  Group B average - Group A average

- Negative values of this statistic favor the alternative



```
In [23]: plt.figure(figsize=(5, 8))
         sns.boxplot(data=births, x = 'Maternal Smoker', y = 'Birth Weight');
```

(Demo)

The Data

Non-smoker

120 oz

Non-smoker

113 oz

Smoker

128 oz

Smoker

108 oz

...

...

Non-smoker

117 oz

Permutations: Shuffling Labels Under the Null

Smoker

120 oz

Non-smoker

113 oz

Non-smoker

128 oz

Smoker

108 oz

...

...

Smoker

117 oz

Shuffling Rows

Random Permutation

- **`df.sample(n)`**
  - Dataframe of n rows picked randomly (default is WITHOUT replacement)
- **`df.sample(frac=1)`**
  - All rows of df, in random order (default is WITHOUT replacement)

(Demo)

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

(Demo)

Recap: Steps in A/B Hypothesis Testing for Birthweight Data

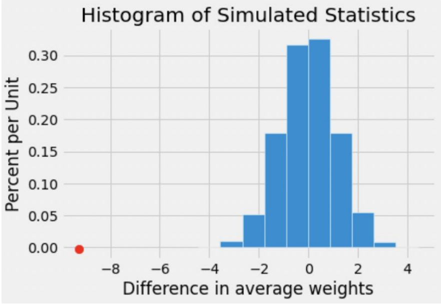
- State Null Hypothesis:
  - ***In the population***, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)
- State Alternative Hypothesis:
  - ***In the population***, the babies of the mothers who smoked weigh less, on average, than the babies of the non-smokers.
- Choose Significance Level: 1%
- Choose Test Statistic:
  - Statistic: Difference between average weights:

    Average weight of babies whose moms smoked - Ave weight of babies whose moms didn’t smoke
- Use random permutations to simulate under the null hypothesis
  - See Python Code
- Plot distribution of simulated null distribution
- Calculate Empirical p-value and make conclusion:

  p<0.01 (highly statistically significant):

  We reject the null hypothesis and

  accept the alternative hypothesis.



Importance of Random Assignment

We’ve concluded that in the population, birth weights of babies whose mothers smoke weigh less than those whose mothers do not

- *Is **lower birth weight** caused by maternal **smoking**?*
- Can’t Tell:
    - Moms aren’t randomly assigned whether to smoke
    - Other factors contribute to their decision to smoke (e.g. income, geography, diet)

Hypothesis Testing Review

- **1 Sample: Two Categories** *(e.g. percent of flowers that are purple)*

  Test Statistic: `observed_proportion` OR `abs(observed_proportion - null_proportion)` `(depending on alternative)`
  How to Simulate: `np.random.binomial(N, null_hyp)` `(this is with replacement)`

- **1 Sample: 3 or more Categories** *(e.g. ethnicity distribution of jury panel)*

  Test Statistic: `tvd(observed_dist, null_dist)` `(this is with replacement)`
  How to Simulate: `np.random.multinomial(N, null_hyp)`

- **1 Sample: Numerical Data** *(e.g. scores in a lab section)*

  Test Statistic: `observed_mean`, `abs(observed_mean - null_mean)`

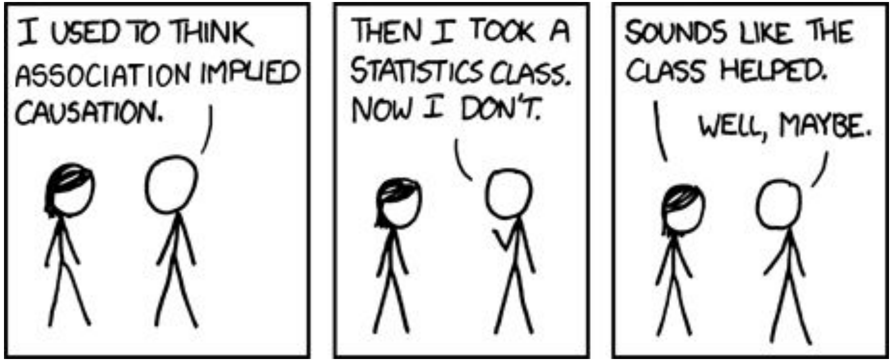
  How to Simulate: How to Simulate: `population_df.sample(n, replace=False)`

- **2 Samples: Numerical Data** *(e.g. birth weights of smokers vs. non-smokers)*

  Test Statistic: `group_a_mean - group_b_mean` OR `group_b_mean - group_a_mean`,
  `OR abs(group_a_mean - group_b_mean)`

  How to Simulate: `observed_df.sample(frac=1, replace=False)`

Causality


I USED TO THINK ASSOCIATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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

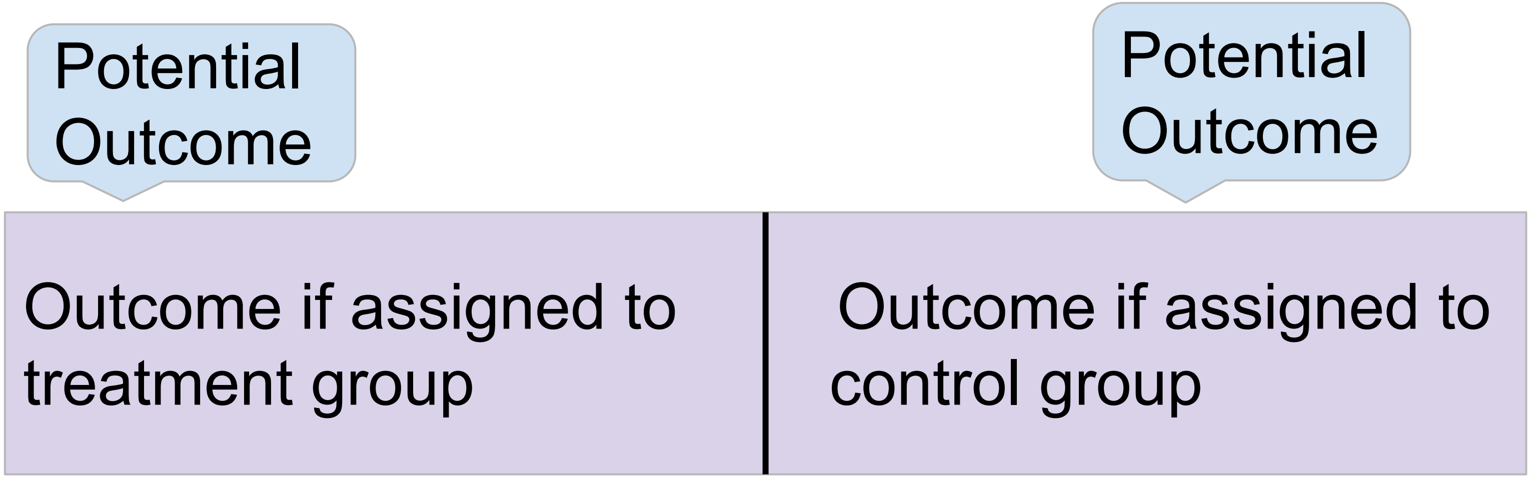
(Python Demo)

Example 2: Python Demo

(Demo)

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

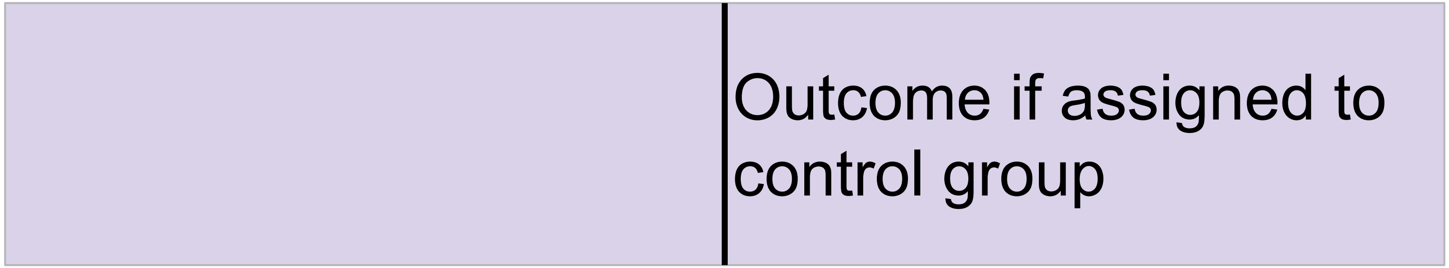
Potential Outcome

Outcome if assigned to treatment group

Outcome if assigned to control group

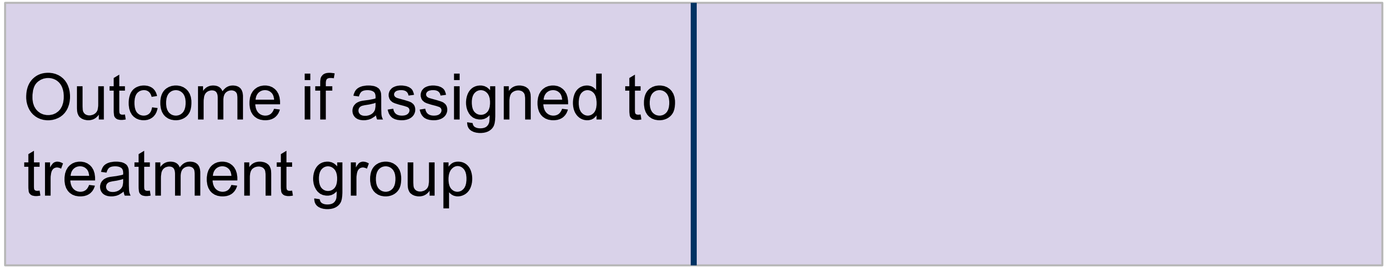
The Data

16 randomly picked tickets show:



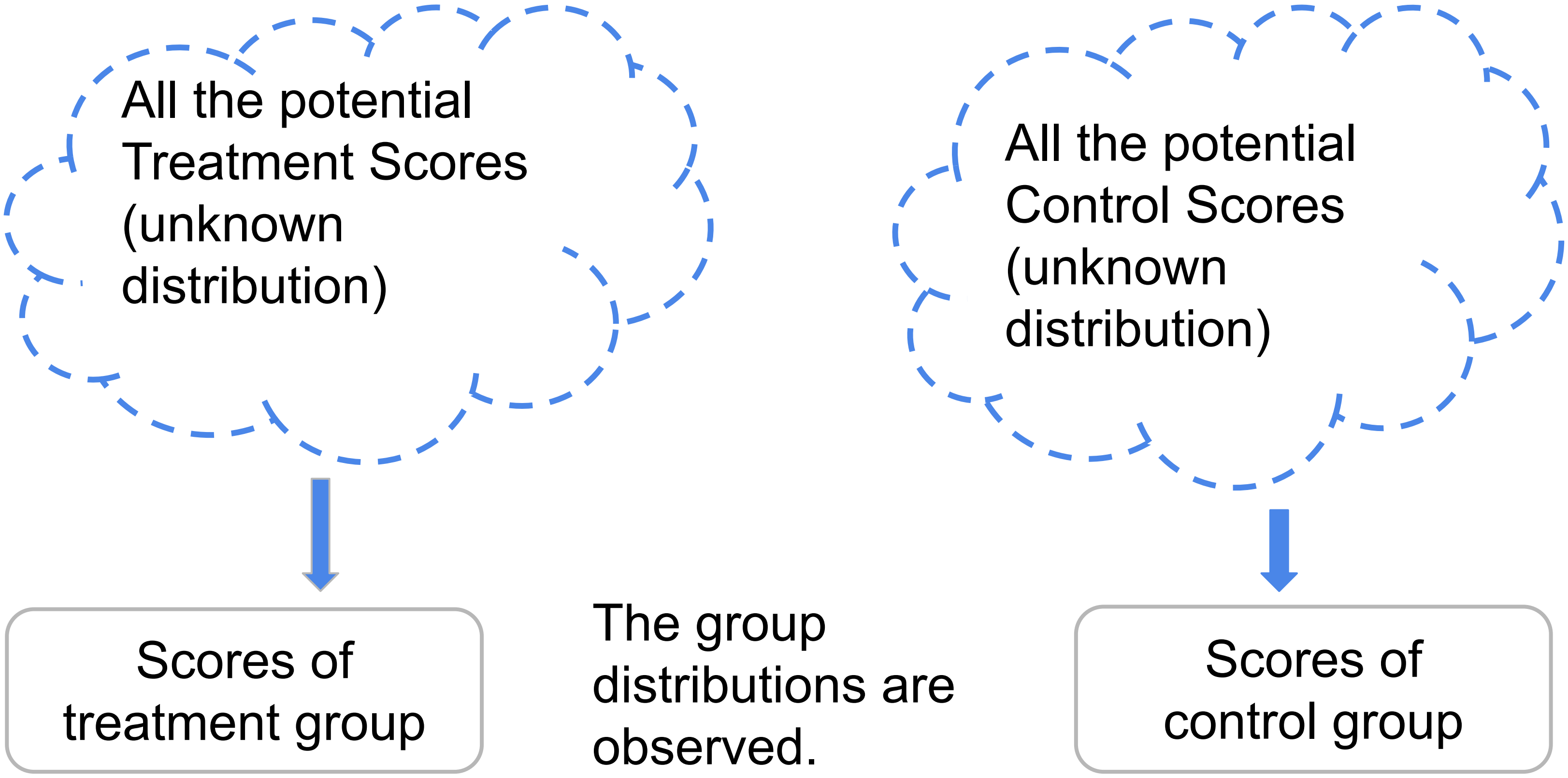
Outcome if assigned to control group

The remaining 15 tickets show:



Outcome if assigned to treatment group

The Question in the Randomized Control Trial (RCT)



All the potential Treatment Scores (unknown distribution)

All the potential Control Scores (unknown distribution)

Scores of treatment group

The group distributions are observed.

Scores of control group

Example 2:

- Null:
  - The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
  - Summary: the treatment has no effect

- Alternative:
  - The distribution of 31 potential “treatment” outcomes is different from that of the 31 control outcomes.
  - Summary: the treatment does something different than the control

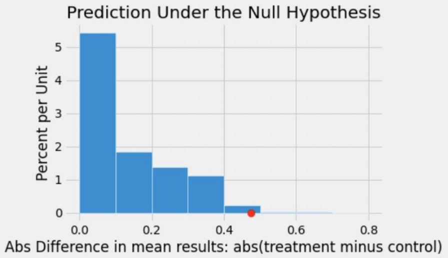
(Demo)

Recap of Results from Randomized Control Trial: Back Pain and Botox

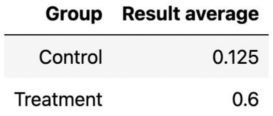
- Null: The treatment has no effect
- Alternative: The treatment does something different than the control
- Significance Level (p-value cutoff): 0.01
- Test Statistic: absolute value of the difference between group proportions

- Simulated null distribution (using permutation test)
- Empirical p-value: 0.009 (area of histogram to right of red dot)

- Test Conclusion: Since p<0.01 we can reject the null and accept that the treatment has an effect.
  - Using the observed data (which shows a positive effect) we can conclude more than just that the treatment had an effect- we can conclude that it had a positive effect (i.e. that it led to pain relief)
  - Because the trials were randomized, the test is **evidence that the treatment causes the difference.** The random assignment of patients to the two groups ensures that there is no confounding variable that could affect the conclusion of causality.
  - But it is **only a conclusion about the 31 patients in the study**. To make conclusions in greater generality, more and larger studies are needed.

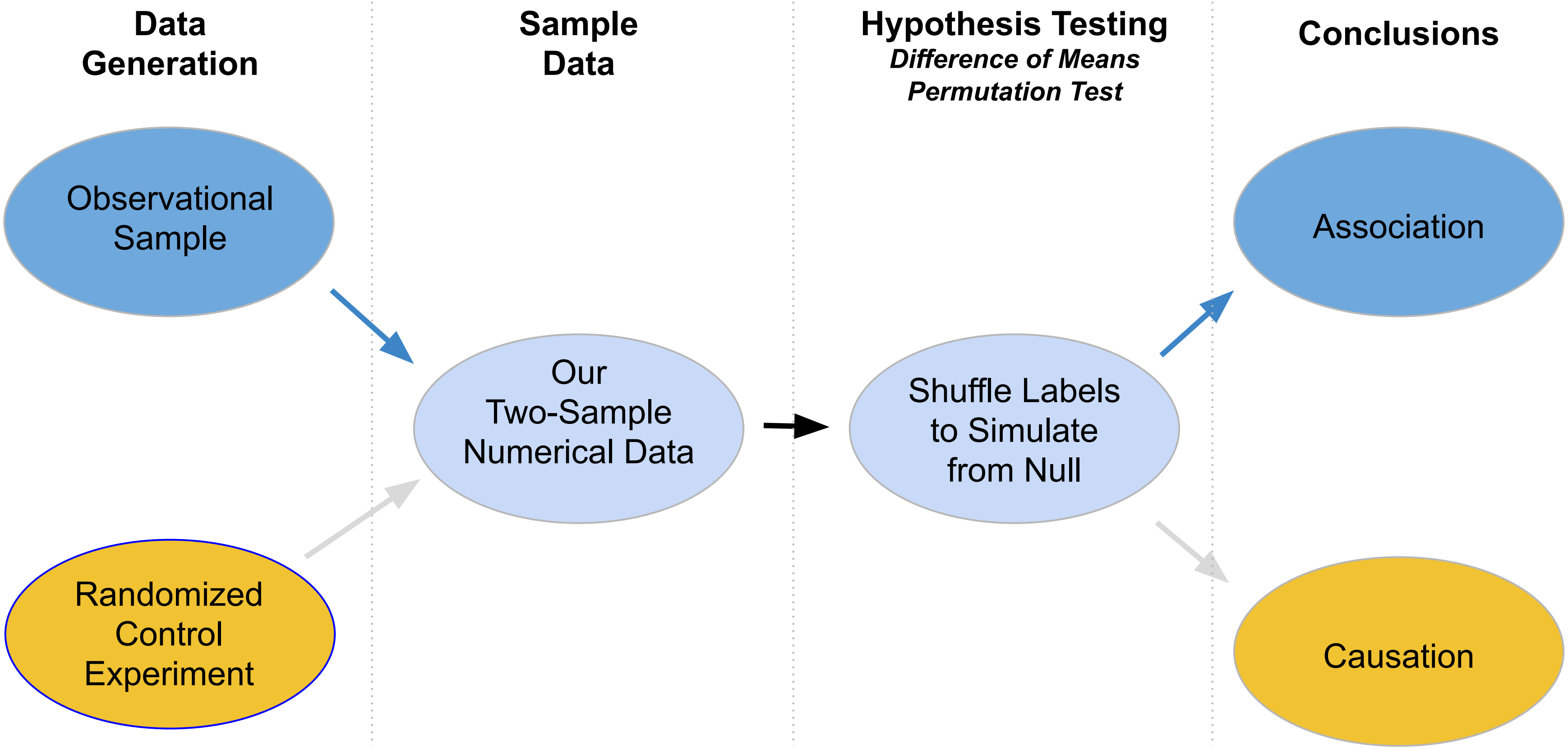


- Observed data: Treatment improved result by 0.475 compared to control



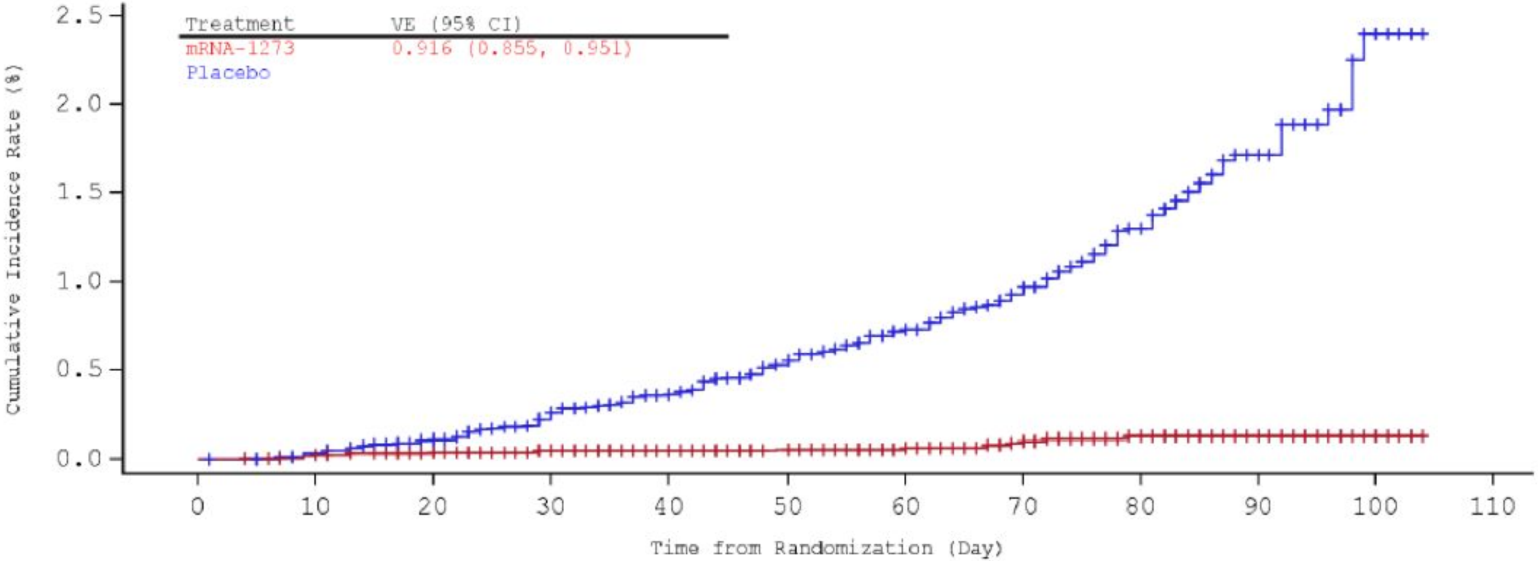
Group	Result average
Control	0.125
Treatment	0.6

Random Assignment & Permutation Tests



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

Causality in the Real World



No. at risk

14312	14306	13964	13490	12981	12284	10742	8327	5705	2621	583	0
14370	14363	14000	13515	12972	12225	10657	8283	5663	2594	586	0

Source: FDA