Comparing Multiple Samples

A/B Tests, Randomized Control Trials and Causality

LECTURE 20

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 8th Week At A Glance

Mon 3/4	Tues 3/5	Wed 3/6	Thurs 3/7	Fri 3/8
Attend & Participate in Class Lesson: Hypothesis Tests with a Single Sample	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class Lesson: Hypothesis Tests with Multiple Samples: A/B Tests, Randomized Controlled Trials and Causality	HW 7 Due 11:59pm	Attend & Participate in Class Lesson: Errors in Hypothesis Tests QUIZ 5: Scope: L13-L16, nb 7, HW 6 HW 8 Released

Lesson Learning Objectives:

- Implement A/B Hypothesis tests and explain conclusions
- Explain how and when to use a permutation test and implement in Python
- Conduct hypothesis test with Randomized Control Trial data and explain conclusions

Today’s Roadmap

CSCI 3022

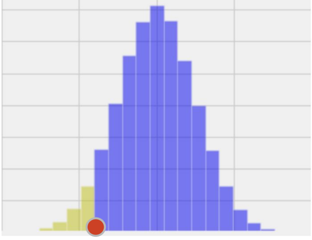
- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - A/B Testing
    - Permutation Tests
    - Randomized Control Trials
    - Causality

Recap: Steps in Hypothesis Testing

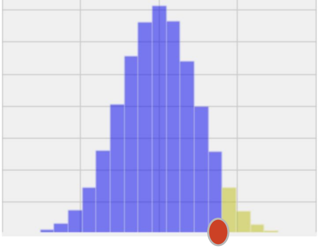
- **Define the null hypothesis and the alternative hypothesis**
- **Choose a significance level** (cutoff tail probability after which you will decide the null hypothesis is inconsistent with the observed data)
- **Choose a test statistic** to measure "discrepancy" between null hypothesis and data
- **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 using the data
- **Conclusion:**
    - If the p-value is less than the significance level:
        - Reject Null and Accept Alternative.
    - Otherwise Fail to Reject Null.

Review: Definition of the *p*-value

Formal name: observed significance level



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



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

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

The *p*-value is the chance (probability),

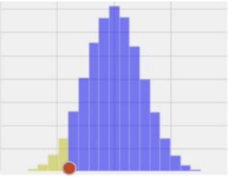
- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

The p-value is actually a Conditional Probability!
P(observed data or more extreme | null hypothesis)

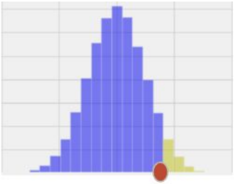
Review: P-value cutoff vs P-value

- P-value (You Compute It)
  - Depends on the observed data and simulation
  - Probability under the null hypothesis that the test statistic is the observed value or more extreme
  - P(data you observed or more extreme | null hypothesis)

- Significance level (i.e. P-value cutoff): You Pick It
  - Does not depend on observed data or simulation
  - “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”



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



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

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

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

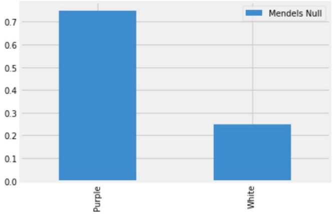
Recap: Hypothesis Testing So Far

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

- Did the pea plants that Mendel grew have colors that were consistent with the chances he specified in his model?

Null Hypothesis

Distribution of Qualitative Variable with
2 categories



In this case Mendel **DIDN'T know the population distribution**, but he made a guess (null hypothesis) about the distribution and wanted to see if it was supported by the random sample he collected.

Alternative Hypothesis

The chance of purple flowers is not 75%

Choose Significance level

5%

Gather Data (Sample of Size N) and Calculate Observed Test Statistic

Grew N=929 plants of which 76.32% had purple flowers

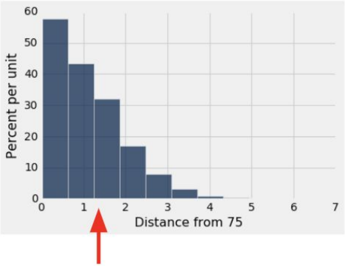
Simulate Random Sample of Size N using null model

```
np.random.binomial(N, null_hyp)
```

Calculate Test Statistic

```
abs(simulated_percentage - null_percentage)
Or abs(simulated_count  - null_count)
Or abs(simulated_proportion - null_proportion)
```

Simulate Distribution of Test Statistic Under Null Hypothesis



Observed Distance (1.32)

Conclusion

Empirical p-value = ____

__________ Null

"Data is consistent with the null"

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

Alternative Hypothesis

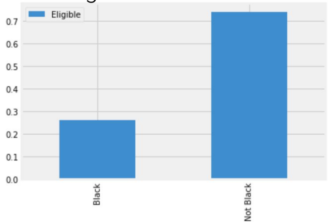
Simulate Random Sample of size N

Test Statistic

Distribution of Test Statistic Under Null Hypothesis

Conclusion

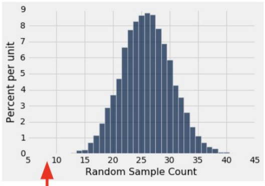
Dist of Qualitative Variable: 2 categories



Selection biased against Black people

```
np.random.
binomial(N, null_prop)
```

Proportion (or count) of 1 category



Observed Count (8)

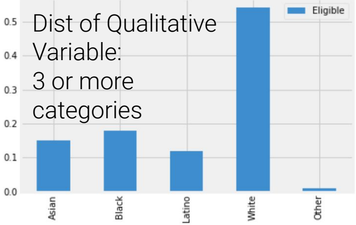
Gather Data and Calculate Observed Test Statistic

Empirical p-value = ____

______ Null

"Data is consistent with the alternative"

Choose Significance Level: 1%

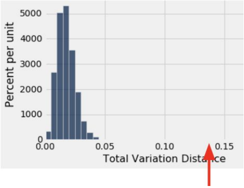


Dist of Qualitative Variable: 3 or more categories

Selection biased

```
np.random.
multinomial(N, null_prop)
```

Total Variation Distance (TVD)



Observed TVD (0.14)

```
sum(abs(simulated_prop -null_prop))/2
```

Empirical p-value = ____

______ 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

Today: Comparing Two Samples using A/B Testing

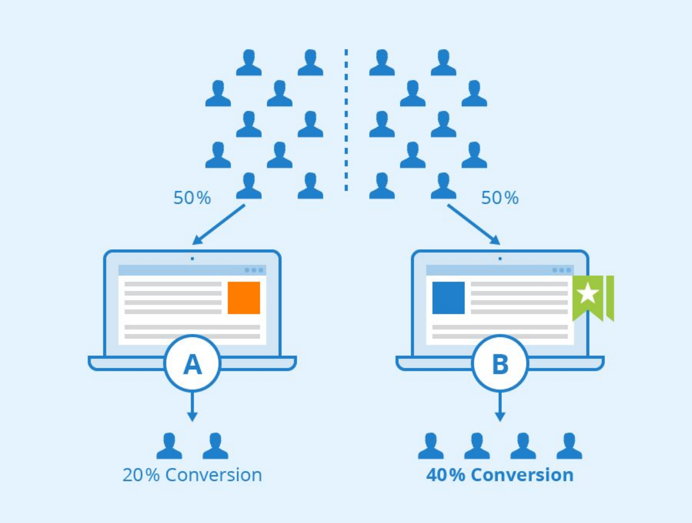
1). Test whether a **single sample** looks like **random** draws from a **specified chance model (null)**.

- Did the pea plants that Mendel grew have colors that were consistent with the chances he specified in his model?
- Do jury panel demographics look like a **random** sample from the known population demographics of eligible jurors?

2). Test whether **two samples** looks like **random** draws from the **same underlying distribution.**

- Ex: Compare number of purchases from 2 different versions of a website

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



A/B Testing

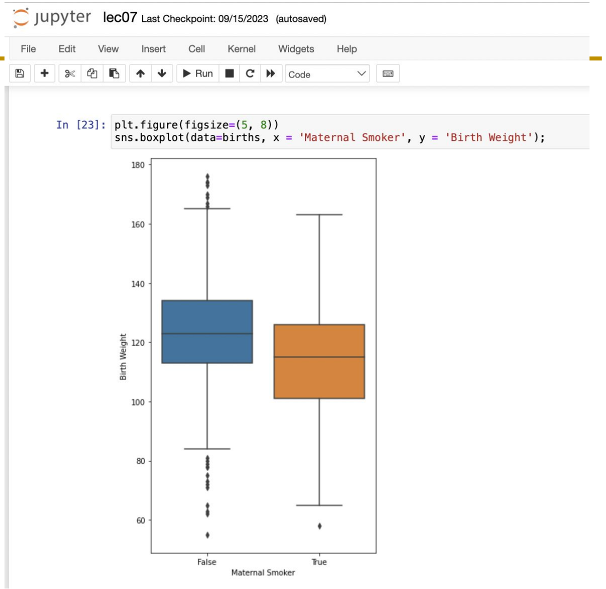
CSCI 3022

- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - **A/B Testing**
    - Permutation Tests
    - Randomized Control Trials
    - Causality

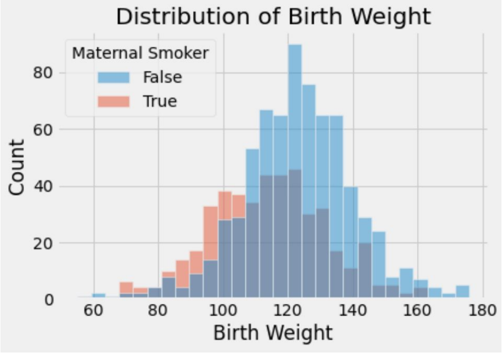
The Groups and the Question

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



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



Question: Could the underlying distributions of birth weights be the same for both groups and the difference we see in these samples just be due to random chance?

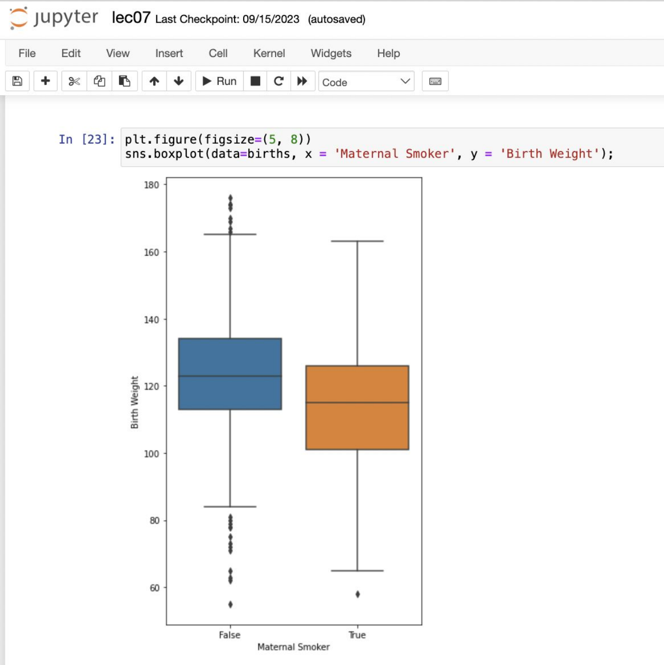
https://onlinestatbook.com/stat_sim/sampling_dist/

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:

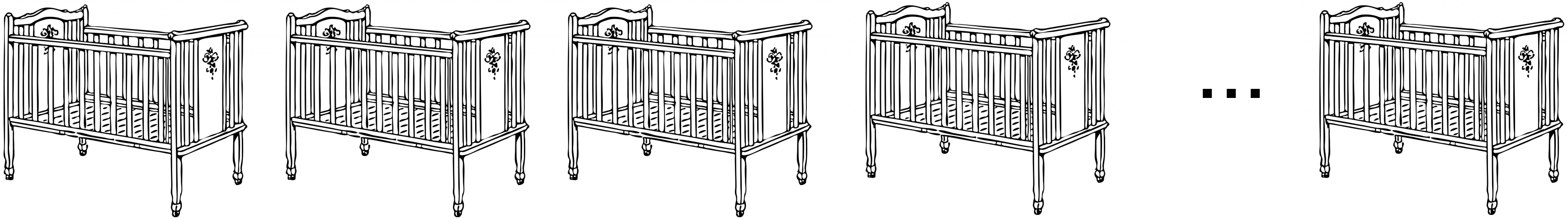
Simulate Random Sample(s) Under the Null



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

Simulate Random Sample(s) Under the Null

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.)



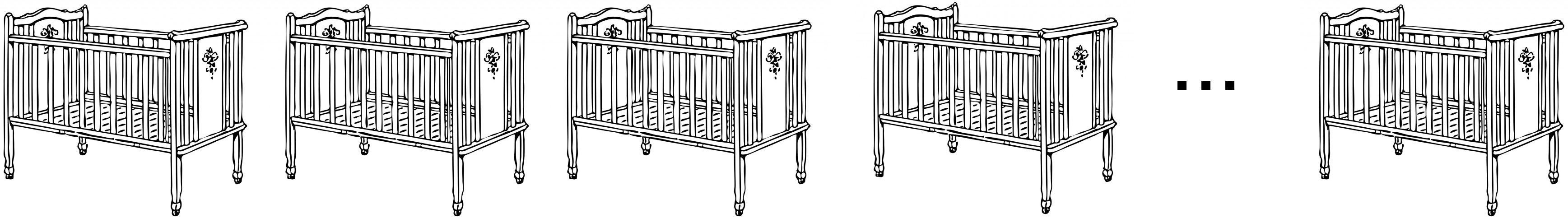
Non-smoker	Non-smoker	Smoker	Smoker	...	Non-smoker
120 oz	113 oz	128 oz	108 oz	...	117 oz

- If the null is true, all rearrangements of labels (i.e. smoker vs non-smoker) are equally likely

Permutations: Shuffling Labels Under the Null

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.)

- If the null is true, all rearrangements of labels (smoker vs non-smoker) are equally likely



Smoker	Non-smoker	Non-smoker	Smoker	…	Smoker
120 oz	113 oz	128 oz	108 oz		117 oz

- Shuffle group labels (smoker vs non-smoker), but don’t shuffle the baby weights.

Using Permutations to Simulate Under the Null

- If the null is true, all rearrangements of labels are equally likely
- Use Permutations:
  - Shuffle values in the Maternal Smoker column but keep other columns fixed
  - Calculate the **test statistic***.
  - Repeat

What's a Possible Test Statistic?

- 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.

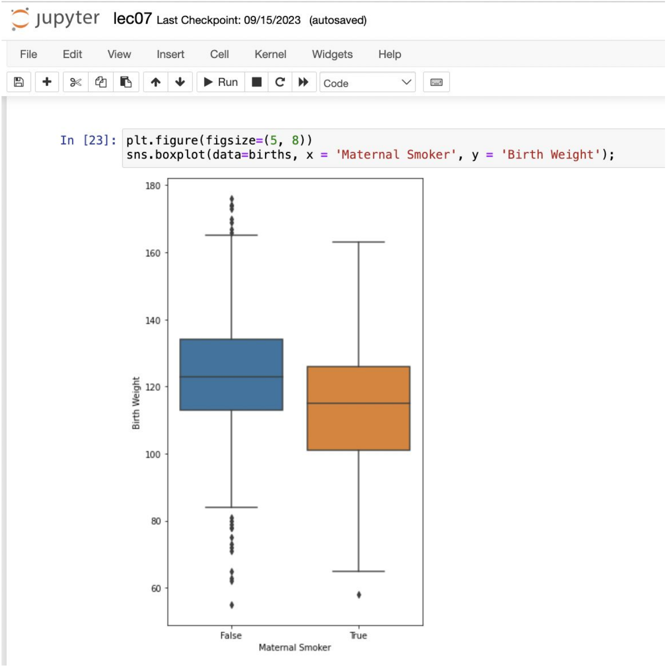
One Possible Test Statistic

- 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



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

Permutation Tests

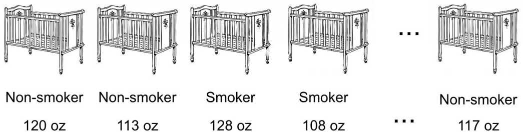
CSCI 3022

- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - A/B Testing
  - **Permutation Tests**
  - Randomized Control Trials
  - Causality

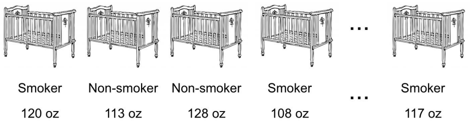
Permutation Test

- **Tests based on permutations of data are called permutation tests.** Thus we use permutation test to conduct an A/B hypothesis test.To do this you must have two (or more) samples of data, and your null hypothesis is that the **samples** are **random** draws from the **same underlying distribution.**

  Creating Permutations:
  - Shuffle values in the label column (i.e. Maternal Smoker) but keep other columns fixed
  - Calculate test statistic (i.e. difference in mean weights of babies born to non-smokers and smokers), now for the shuffled data. This constitutes one permutation iteration.
- Repeat N times and plot an empirical distribution of the test statistic



Shuffle labels



Recalculate test statistic:

Smoker average weight - non-smoker average weight

Permutation Test in Python

Useful Pandas method:

- **`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: Permutation Tests in Python

Recap: A/B Hypothesis Testing

2). Test whether **two samples** looks like **random** draws from the **same underlying distribution.**

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.)

Alternative Hypothesis

In the population, the babies of the mothers who smoked **weigh less,** on average, than the babies of the non-smokers.

Simulate Random Sample of size N

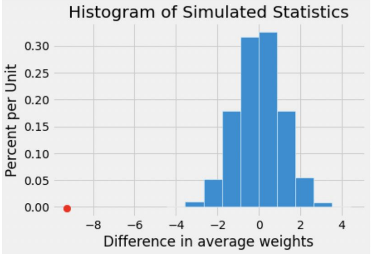
```
Use Permutations:

observed_df.sample(frac=1,
replace=False)
```

Test Statistic

Difference between means

```
group_a_mean -
group_b_mean  OR
group_b_mean -
group_a_mean
```

Distribution of Test Statistic Under Null Hypothesis



Gather Data and Calculate Observed Test Statistic

Conclusion

Empirical p-value = 0 < 1%

REJECT Null and accept alternative

In these cases there was an observed sample of weights with two different categories (smoking vs non-smoking) and we did NOT know the underlying population distribution of weights. We were testing whether the observed samples could have been randomly chosen from the same underlying distribution.

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** <u>caused by</u> 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)

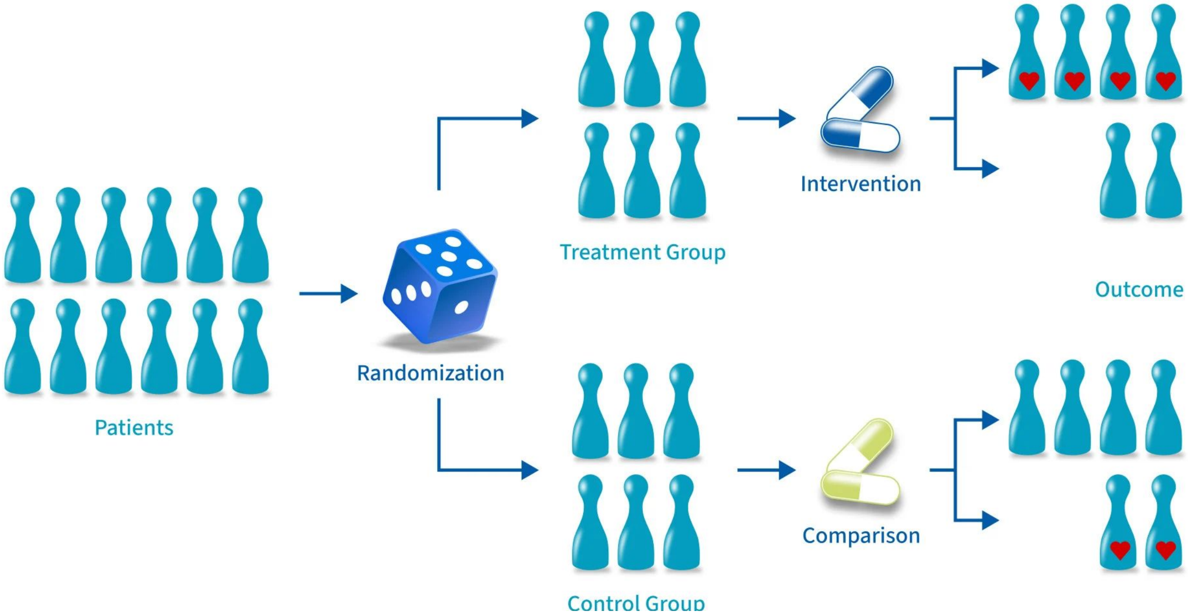
Randomized Control Trials

CSCI 3022

- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - A/B Testing
  - **Permutation Tests**
  - **Randomized Control Trials**
    - Causality

Randomized Controlled Trials

Common in health/medical studies



Randomized Controlled Experiments & Causality

- 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

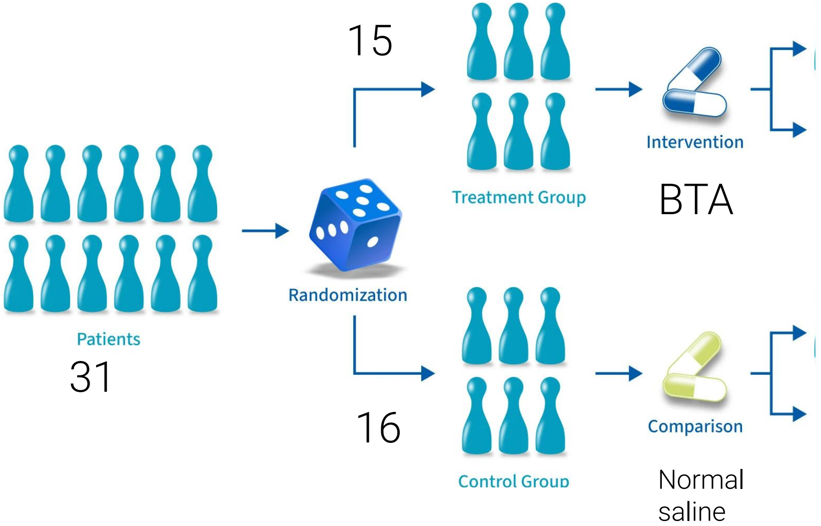
Demo:

CSCI 3022

- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - A/B Testing
  - **Permutation Tests**
  - **Randomized Control Trials**
    - Causality

Example: Treating Back Pain

A randomized controlled trial (RCT) examined the effect of using Botulinum Toxin A (BTA) as a treatment for low-back pain.



60% of treatment group reported pain relief

Could this difference be due just to chance?

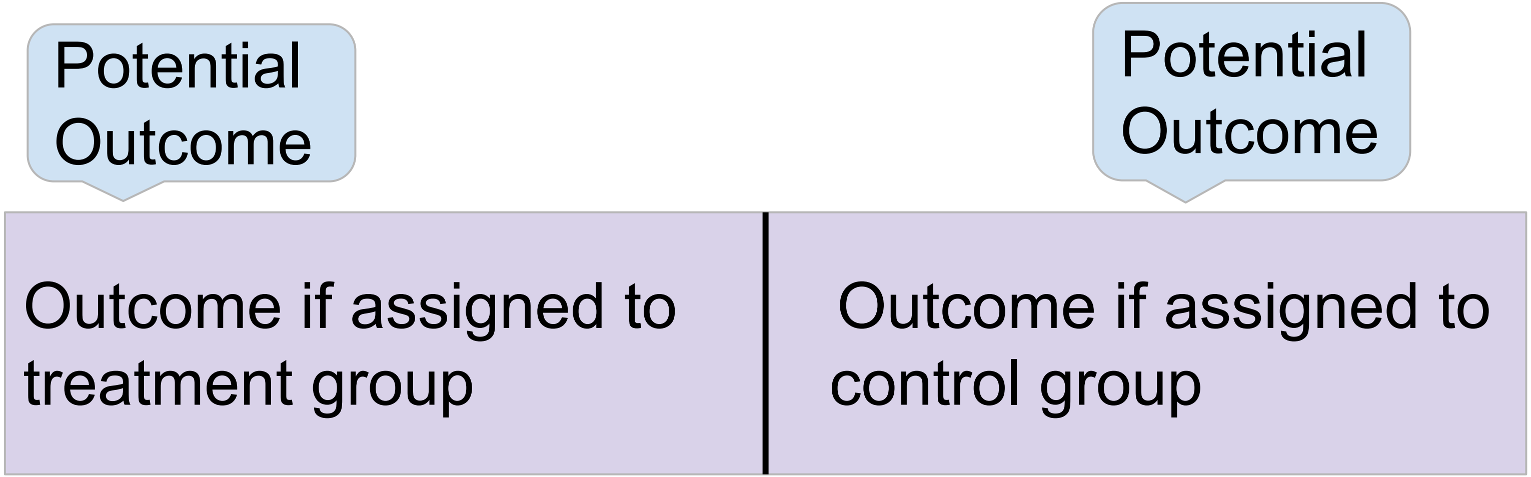
12.5% of control group reported pain relief

The trials were run double-blind so that neither doctors nor patients knew which group they were in.

Wait: Where is the “Chance” in this Scenario?

Analogy for Understanding the Chance Model used to test RCT Results:

- 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

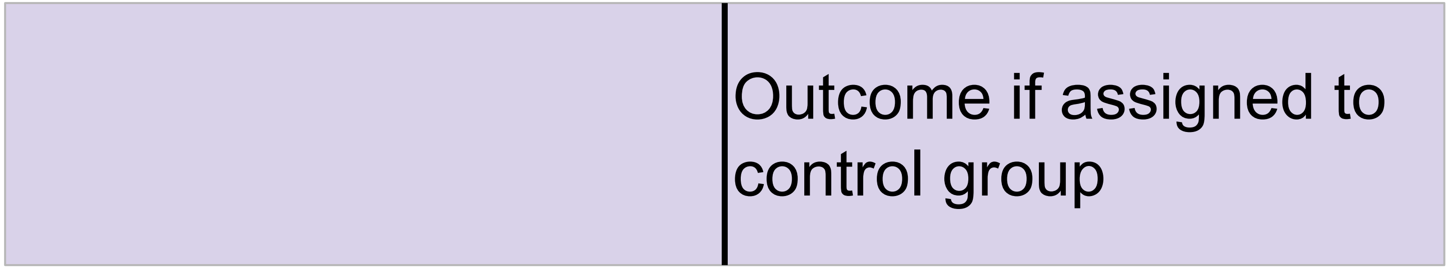
Potential Outcome

Outcome if assigned to treatment group

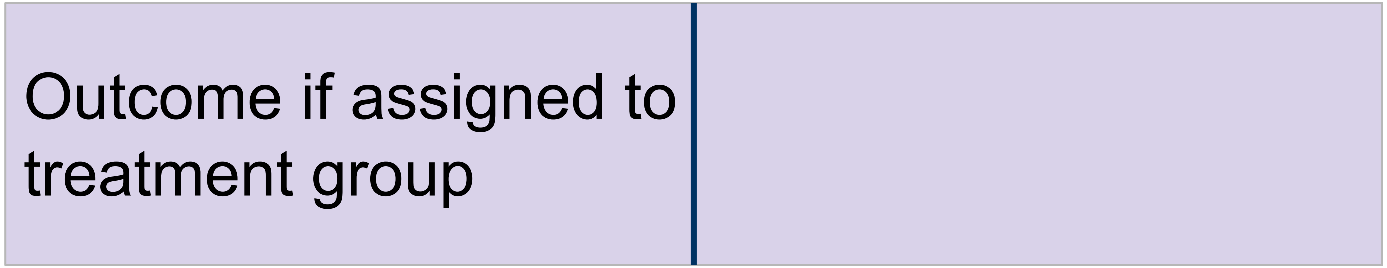
Outcome if assigned to control group

The Data

16 randomly picked tickets show:



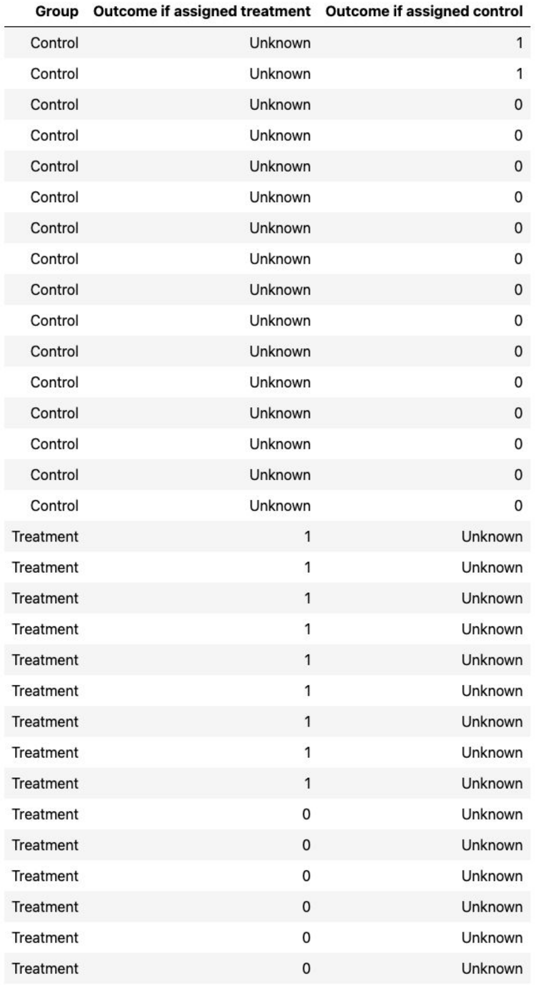
The remaining 15 tickets show:



Determining if Data Came from Same Underlying Distribution

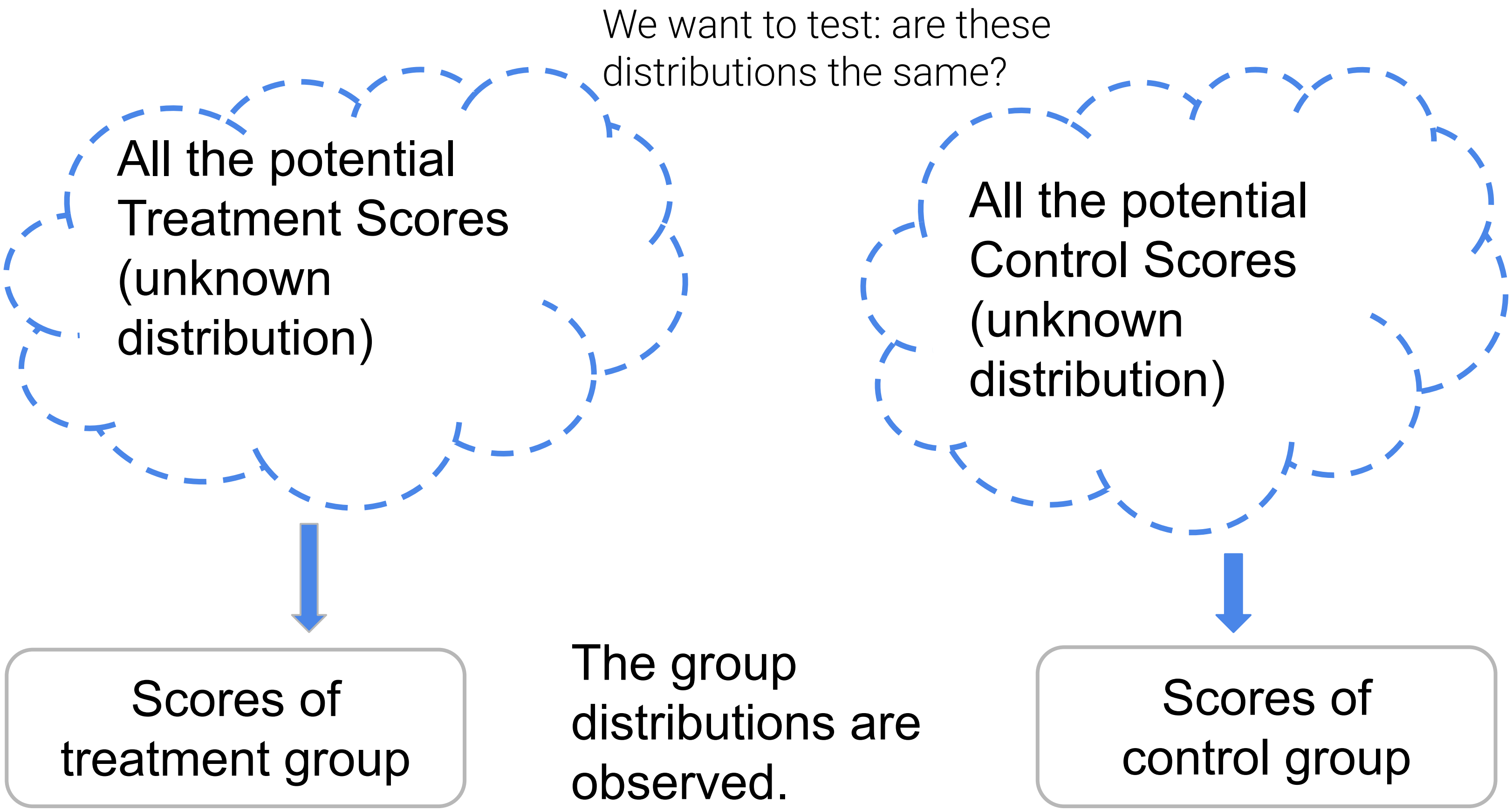
Question:

Is the distribution of the 31 “treatment” values (including the unknown ones), the same as distribution of the 31 “control” values (including the unknown ones)?



Group	Outcome if assigned treatment	Outcome if assigned control
Control	Unknown	1
Control	Unknown	1
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown

The Question in the Randomized Control Trial (RCT)



We want to test: are these distributions the same?

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

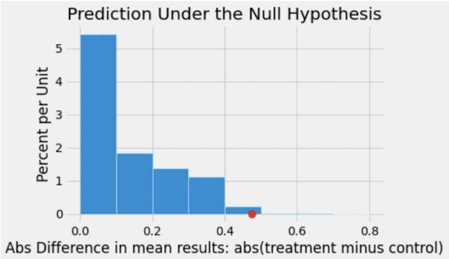
Example

- 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**

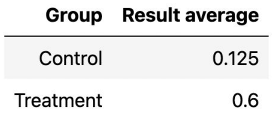
- Alternative:
  - ○ The distribution of 31 potential “treatment” outcomes is different from that of the 31 control outcomes.
  - ○ **Summary: the treatment does something different than the control**

Hypothesis Test for Randomized Control Trial: Back Pain and Botox

- Null: The treatment has no effect. Any difference we see is due to chance.
- Alternative: The treatment does something different than the control
- Significance Level (p-value cutoff):
- Test Statistic:
- How to Simulate Under Null Hypothesis:
- Observed test statistic:
-
- Empirical p-value:
- Test Conclusion:



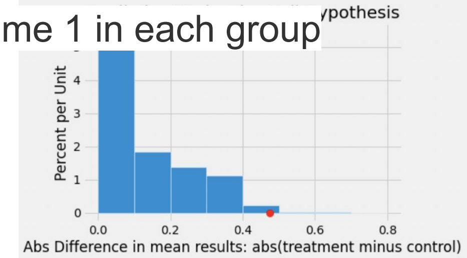
○ Observed data: Treatment improved result by 0.475 compared to control



Group	Result average
Control	0.125
Treatment	0.6

Hypothesis Test for Randomized Control Trial: Back Pain and Botox

- Null: The treatment has no effect. Any difference we see is due to chance.
- Alternative: The treatment does something different than the control
- Significance Level (p-value cutoff): 0.01
- Test Statistic: absolute value of the difference between proportions with outcome 1 in each group



- How to Simulate Under Null Hypothesis: Use permutations. Shuffle “Group” column (control vs treatment) but leave other columns fixed and re-calculate test statistic
- Observed test statistic: 0.475
- Empirical p-value: 0.009 (area of histogram to right of observed test statistic)
- Test Conclusion: Since p<0.01 we can reject the null and accept that the treatment has an effect.

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

Group	Result average
Control	0.125
Treatment	0.6

Causality

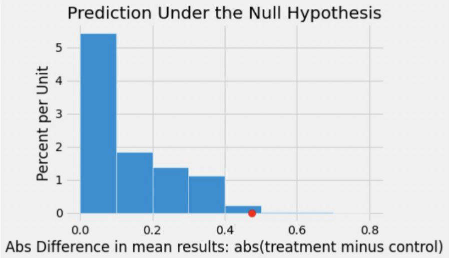
CSCI 3022

- Review: Hypothesis Testing with a single sample
- Hypothesis Tests with Multiple Samples:
  - A/B Testing
  - **Permutation Tests**
  - **Randomized Control Trials**
    - **Causality**

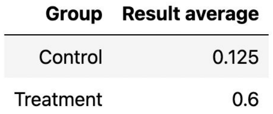
Causality and Hypothesis Tests

- Recall: If the treatment and control groups are selected at random any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

- Test Conclusion: Since p<0.01 we can reject the null that the difference was due to chance. Thus we accept the alternative that the difference we observed is due to the treatment.
  - 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.

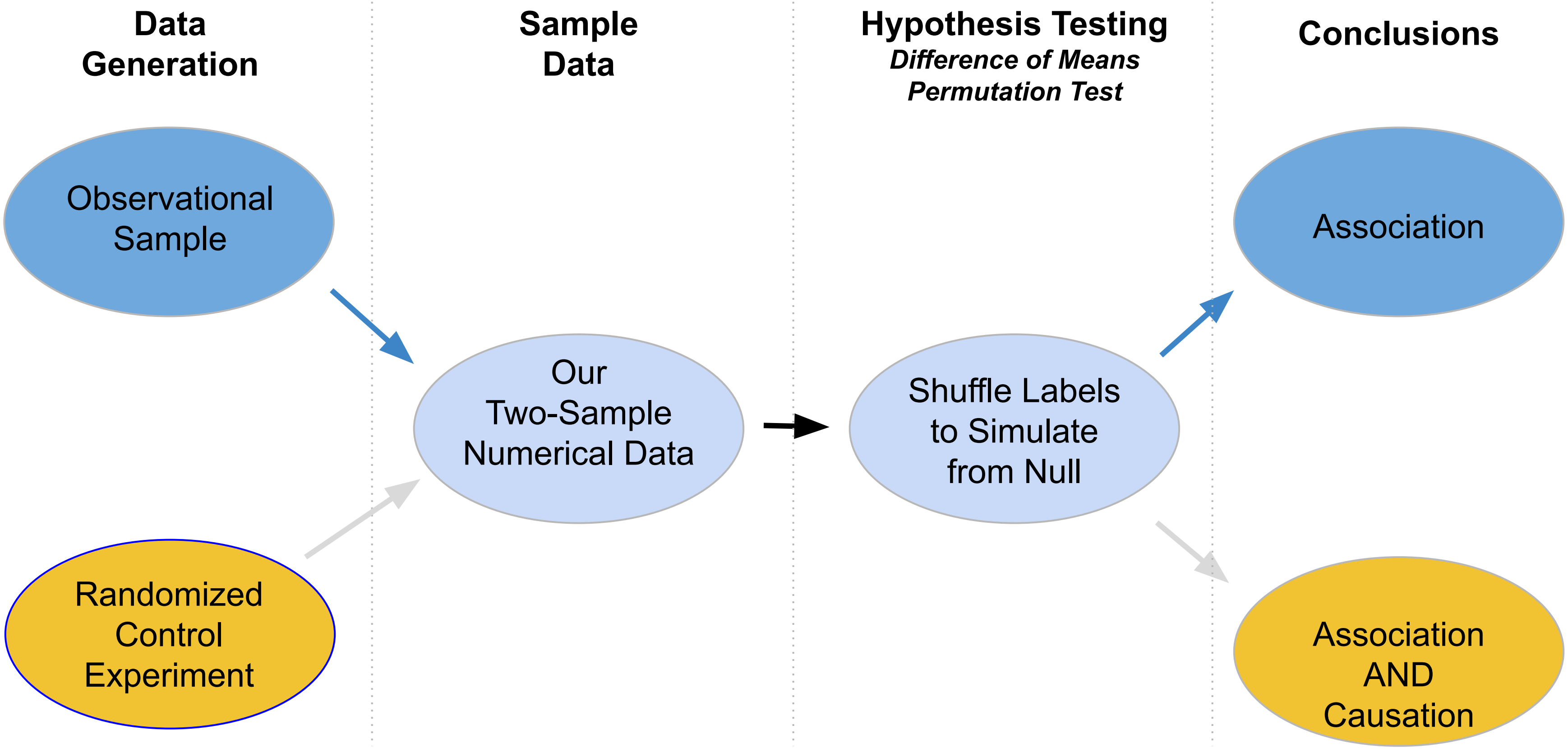


- Observed data: Treatment improved result by 0.475 compared to control



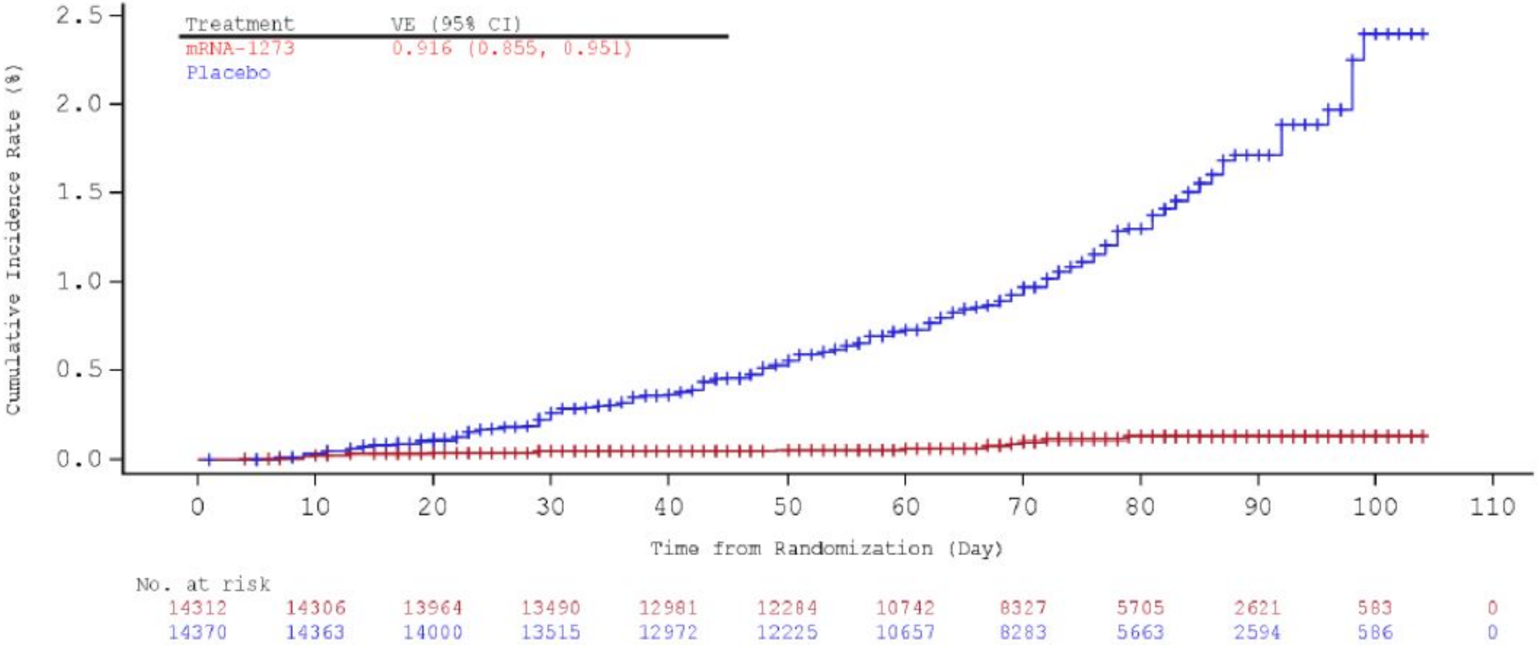
Group	Result average
Control	0.125
Treatment	0.6

When Can we Make Causal Conclusions?



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
Association AND Causation

Causality in the Real World: Covid Vaccines



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