LESSON 19

Hypothesis Testing - A/B Testing

Comparing 2 Samples To Determine if They’re From the Same Distribution

Roadmap

- Hypothesis Tests with Multiple Samples
- Causality
- Recap of Steps in Hypothesis Testing

A/B Testing

- Hypothesis Tests with Multiple Samples:
  - **A/B Testing**

This Lesson: Comparing Two Samples using A/B Testing

Often we're interested in testing whether **two samples** looks like **random** draws from the **same underlying distribution.**

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

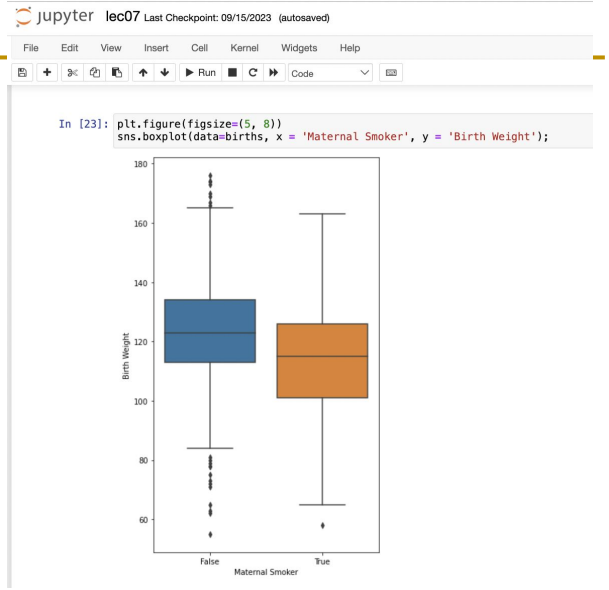
Answering this question by performing a statistical test is called **A/B testing**.

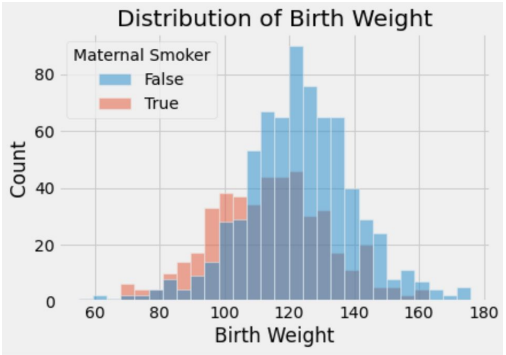
Example: A/B Testing Example

- Recall our babies data set with a 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

- 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?





Example: A/B Hypothesis Testing

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

Null Hypothesis	Alternative Hypothesis	Test Statistic	Simulate Test Statistic Under Null Hypothesis	Gather Data and Calculate Observed Test Statistic	Calculate p-value and make Conclusion
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.)	***In the population***, the babies of the mothers who smoked **weigh less,** on average, than the babies of the non-smokers.				

In these cases there was an observed sample of weights with two different categories (smoking vs non-smoking) and we do 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.

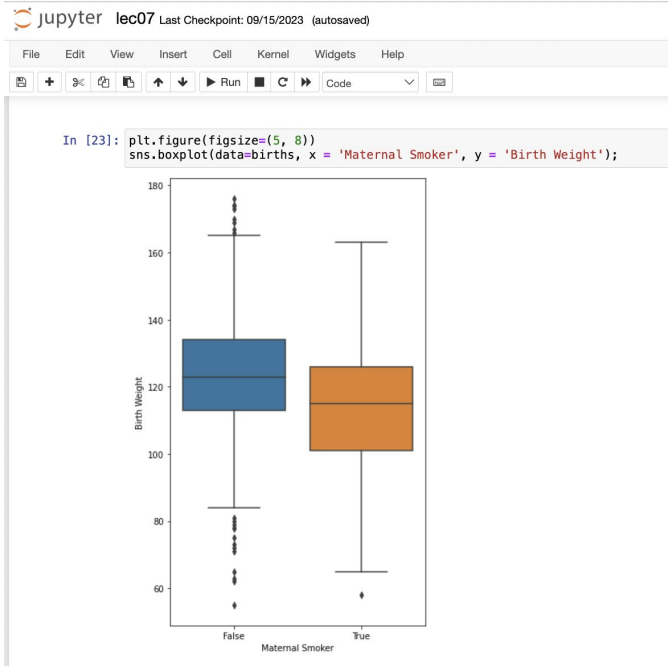
One Possible Test Statistic

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

- Statistic: Difference between average weights

  Group B average - Group A average

- What values of this statistic will favor the alternative?



A/B Hypothesis Testing

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

Null Hypothesis	Alternative Hypothesis	Test Statistic	Simulate Test Statistic Under Null Hypothesis	Gather Data and Calculate Observed Test Statistic	Calculate p-value and make Conclusion
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.)	***In the population***, the babies of the mothers who smoked **weigh less,** on average, than the babies of the non-smokers.	Difference between mean weights			

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.

Simulate Random Sample(s) Under the Null Using Permutations

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.)

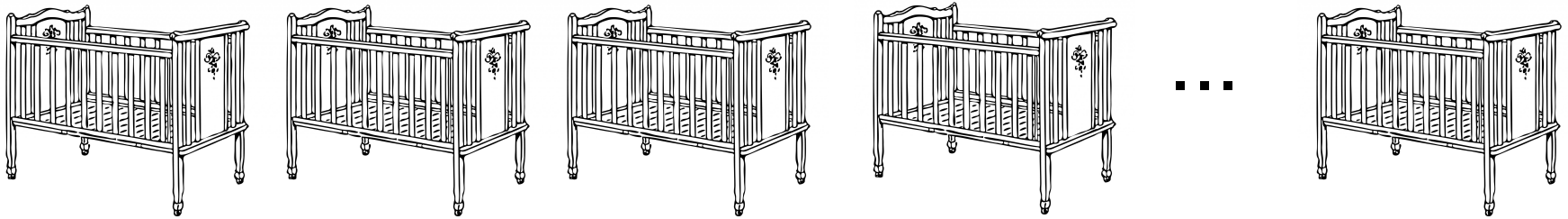
Non-smoker	Non-smoker	Smoker	Smoker	...	Non-smoker
120 oz	113 oz	128 oz	108 oz	...	117 oz

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

Simulate Random Sample(s) Under the Null Using Permutations

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.

Permutation Test

- **Hypothesis tests conducted using permutations of data are called permutation tests.**
  - **Our A/B tests are an example of a permutation test**

- Permutation tests: Need 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

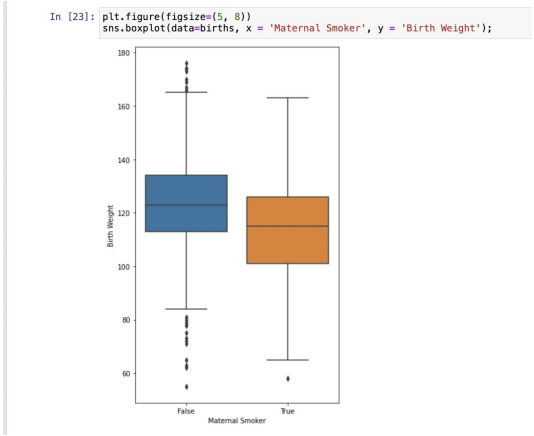
Recall Where We Left Off Last Time; Example: A/B Testing Example

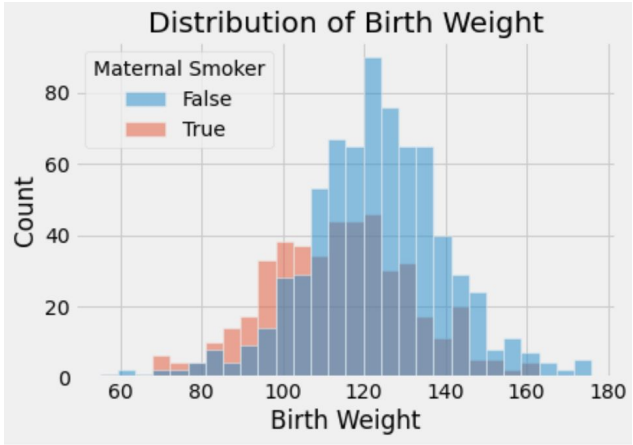
- Recall our babies data set with a random sample of mothers and newborns.

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

- 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?

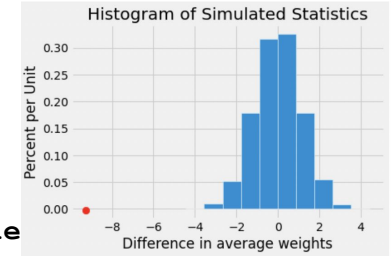
https://onlinestatbook.com/stat_sim/sampling_dist/





A/B Hypothesis Testing

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

Null Hypothesis	Alternative Hypothesis	Test Statistic	Simulate Test Statistic Under Null Hypothesis	Gather Data and Calculate Observed Test Statistic	Calculate p-value and make Conclusion
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.)	***In the population***, the babies of the mothers who smoked **weigh less,** on average, than the babies of the non-smokers.	Difference between mean weights	Use Permutations to simulate random samples: `observed_df.sample (frac=1, replace=False)` Calculate the difference between mean weights		Empirical p-value = 0 < 1% REJECT Null Data is consistent with 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.

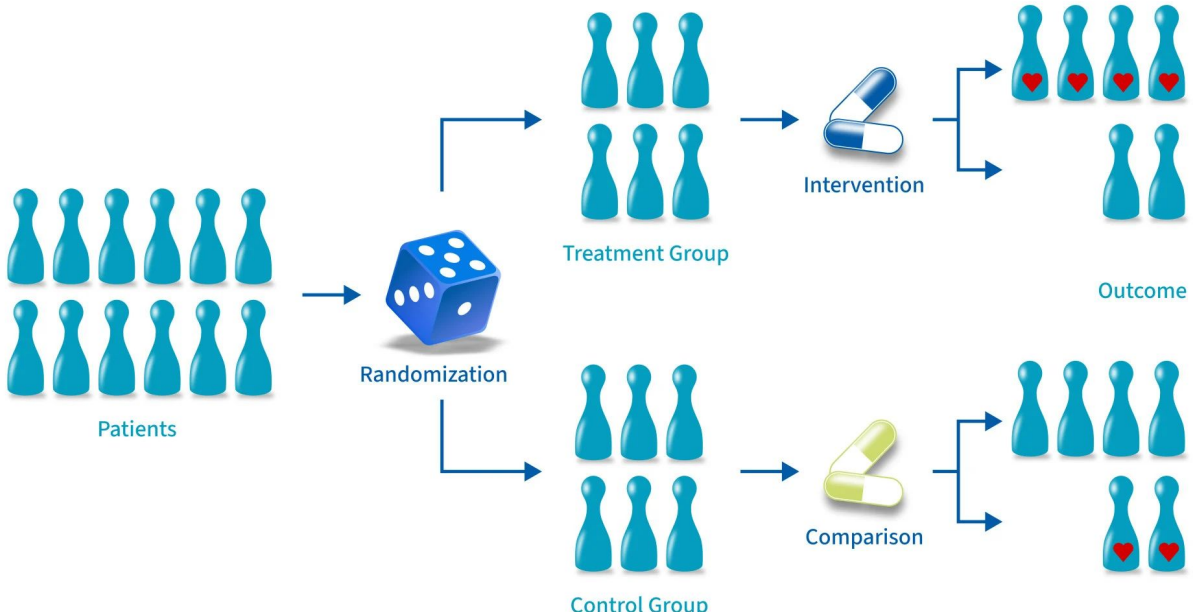
Significant Does Not Mean Causal

We've concluded the data is consistent with the alternative hypothesis that 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)

Causality

Randomized Controlled Trials



Common in health/medical studies

Randomized Controlled Experiments & Causality

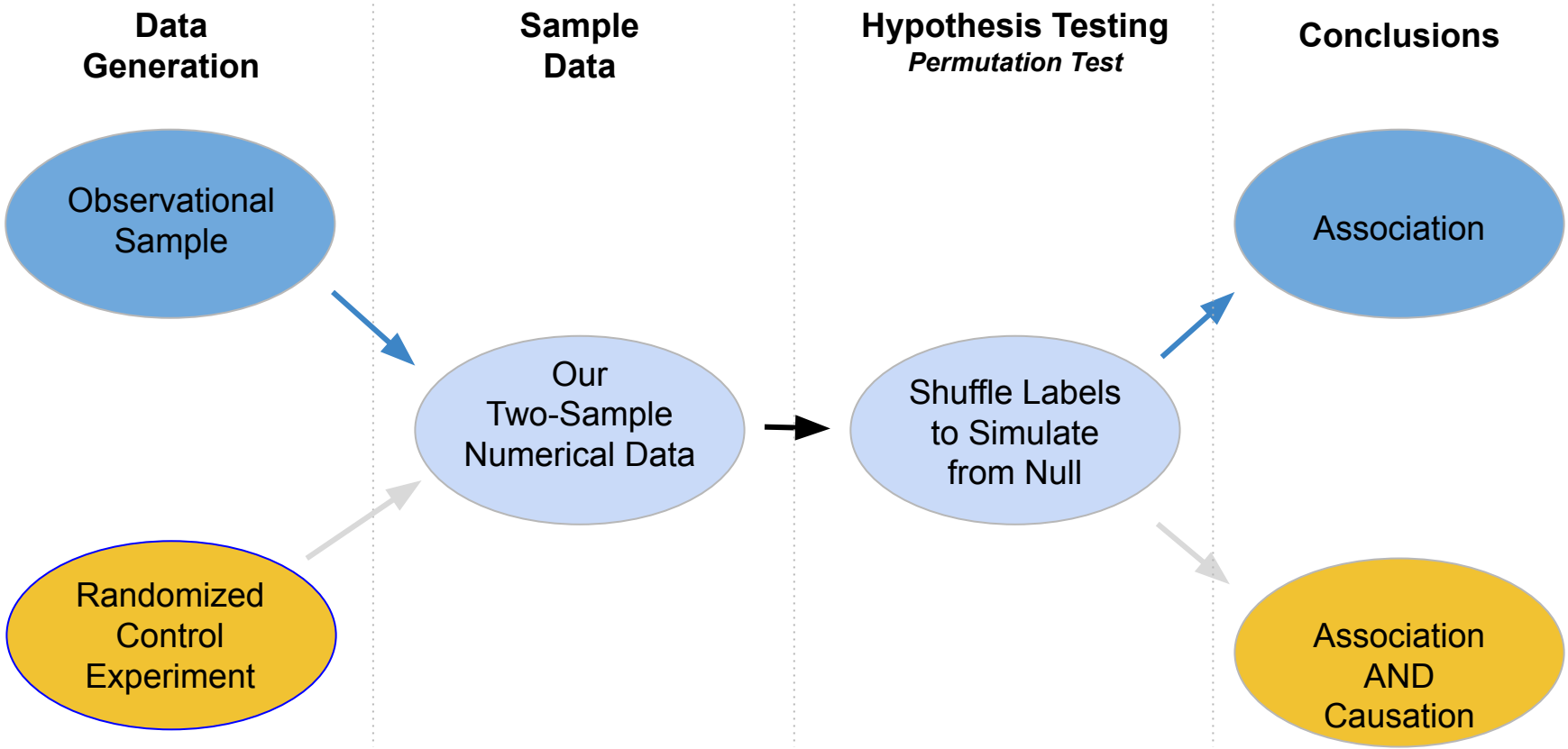
- Sample A: control group
- Sample B: treatment group

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

By randomly assigning cases to different conditions, if we get a significant result from our hypothesis test, a causal conclusion can be made;
in other words, we can say that differences in the response variable are caused by differences in the explanatory variable.

Without randomization, an association can be noted, ***but a causal conclusion cannot be made***.

When Can we Make Causal Conclusions?

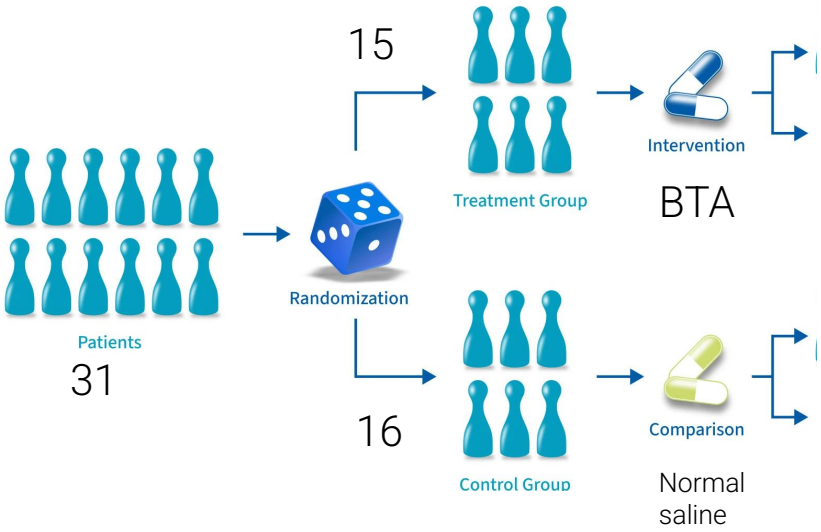


Supplementary Material: More Practice:

https://youtu.be/R8zjn4s-cP0\

Video 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

12.5% of control group reported pain relief

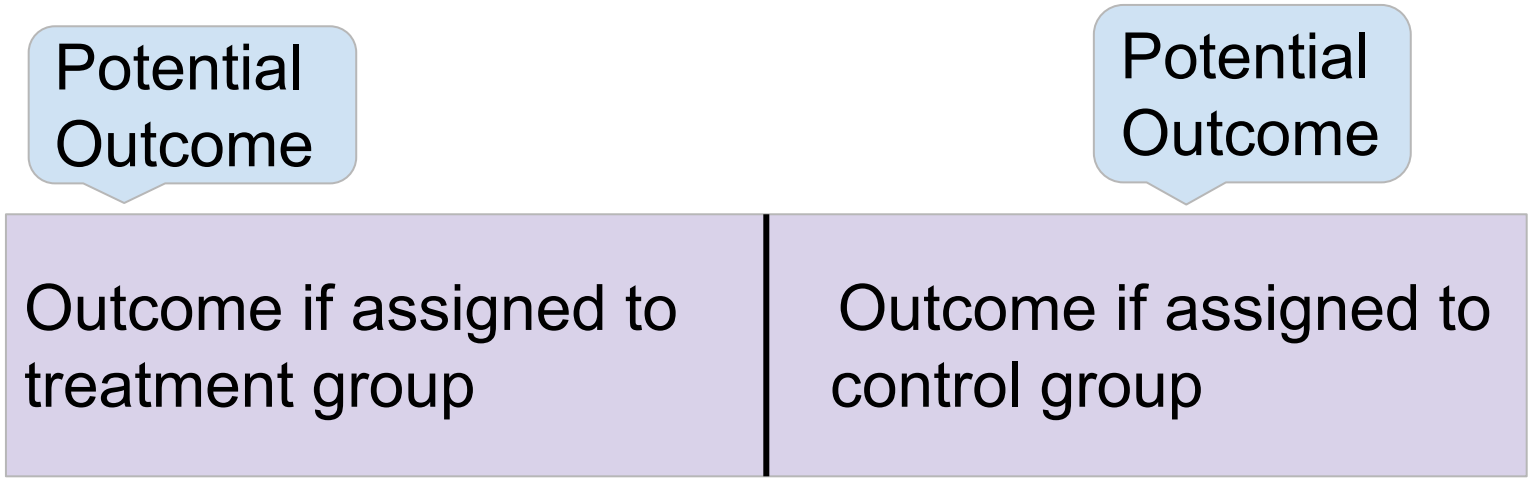
Could this difference be due just to chance?

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:



The Data

16 randomly picked tickets show:

	Outcome if assigned to control group

The remaining 15 tickets show:

Outcome if assigned to treatment group	

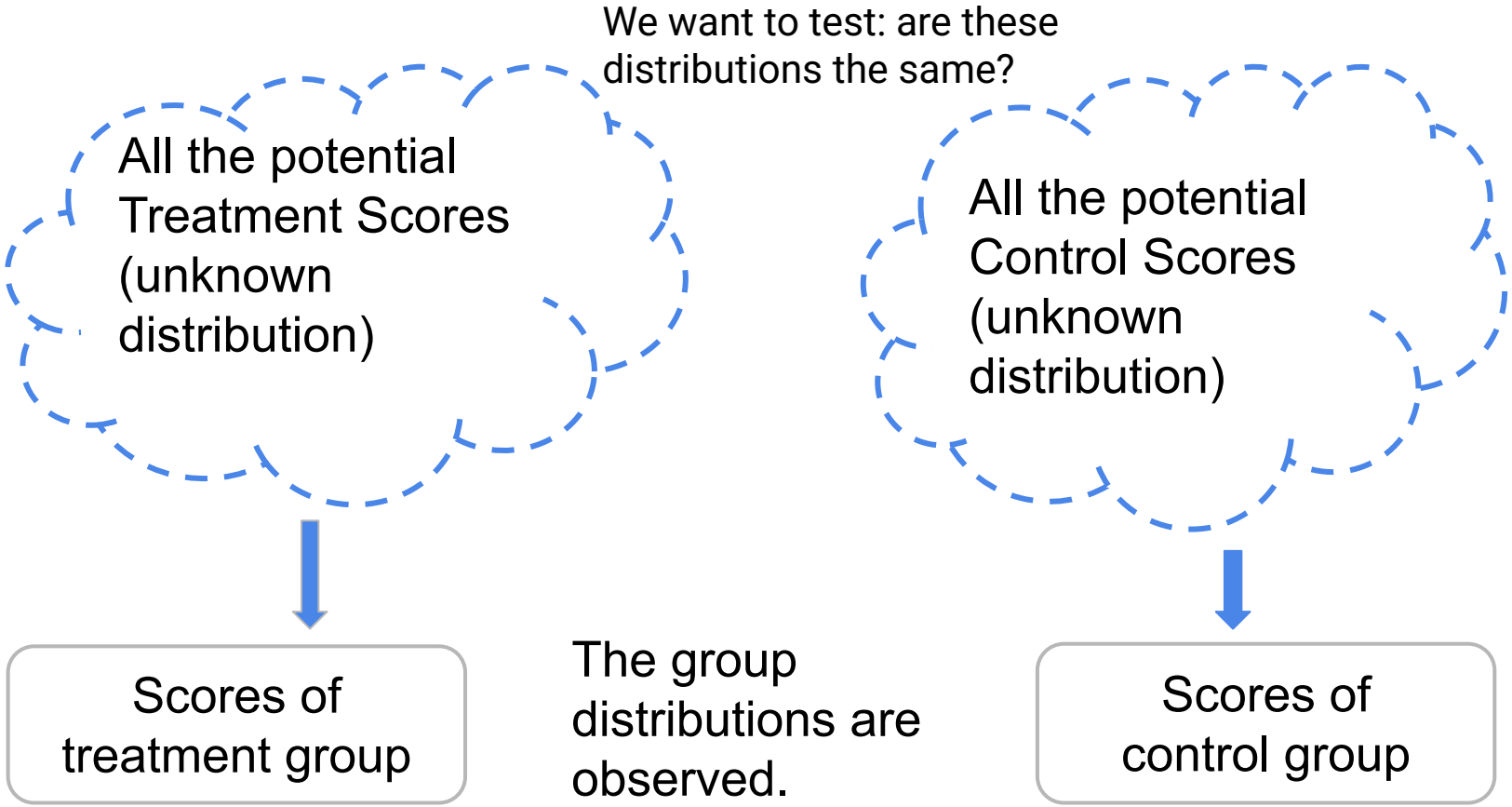
Determining if Data Came from Same Underlying Distribution

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)



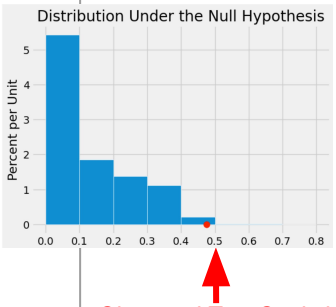
Example

- 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**

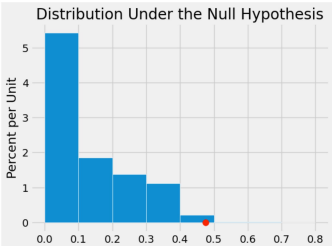
Hypothesis Test for Randomized Control Trial: Back Pain and Botox

2). Test whether a **multiple samples** look like random draws from the **same distribution**.

Null	Alternative	Significance	Test Statistic	Simulate Distribution of Test Statistic Under Null	Calculate Observed test statistic	Calculate p-value and make conclusion
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.	The distribution of all 31 potential "treatment" outcomes is **NOT** the same as that of all 31 potential "control" outcomes.	1%	X= absolute value of the difference between proportions with outcome 1 in each group (notice this is equivalent to the TVD)	Use Permutations to simulate random samples: `observed_df.sample(frac=1, replace=False)` Calculate the TVD between the 2 samples	Observed Test Statistic = 0.475 	Empirical P-value = P(X>=0.475) = .009 Conclusion: Since 0.009<0.01 We reject the null. The data is consistent with the alternative and the result is highly statistically significant.

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

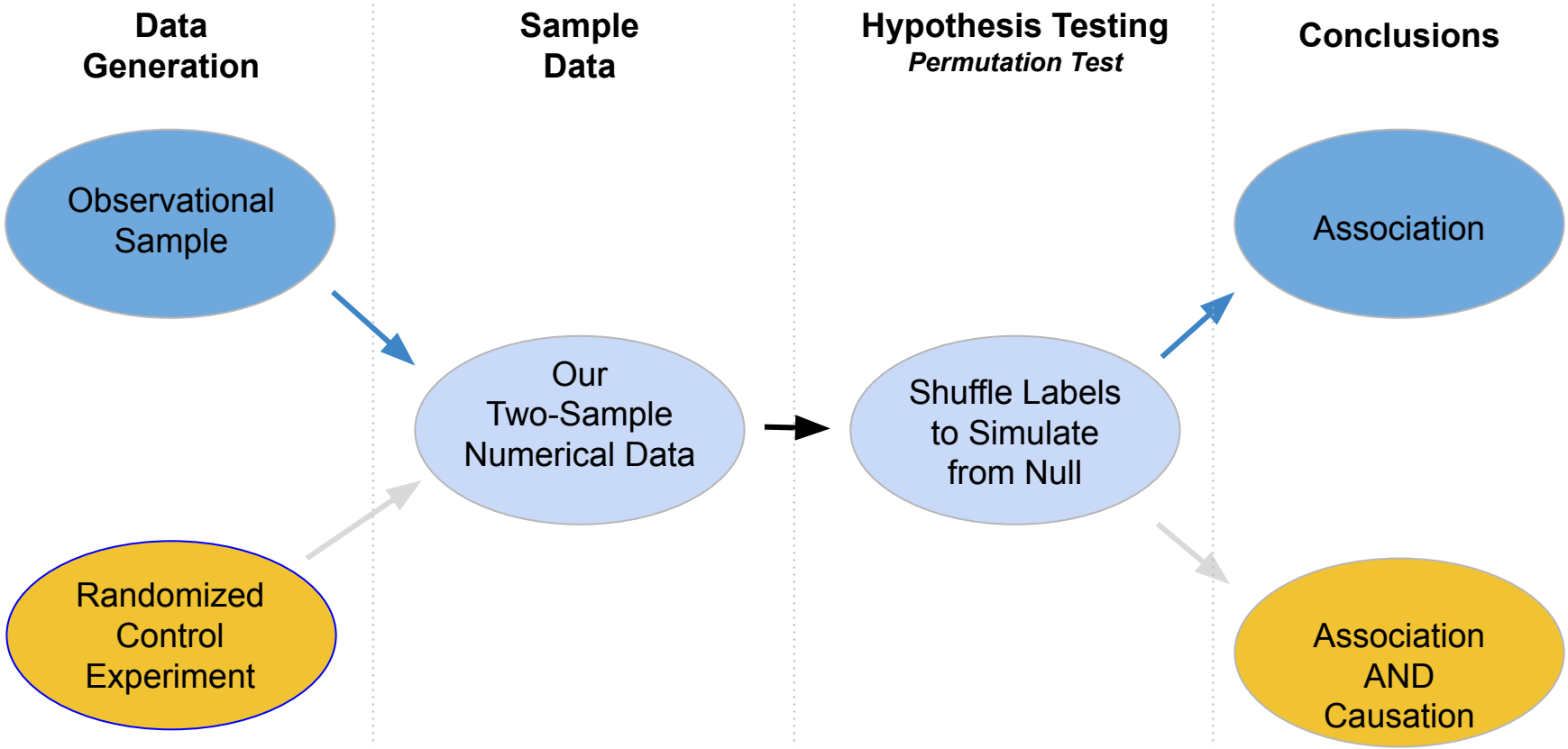


- Test Conclusion: Since p<0.01 we can reject the null that the difference was due to chance. Thus we accept the alternative that the difference we observed is due to the treatment.
  - Because the trials were randomized, the test is evidence that the treatment causes the difference. The random assignment 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

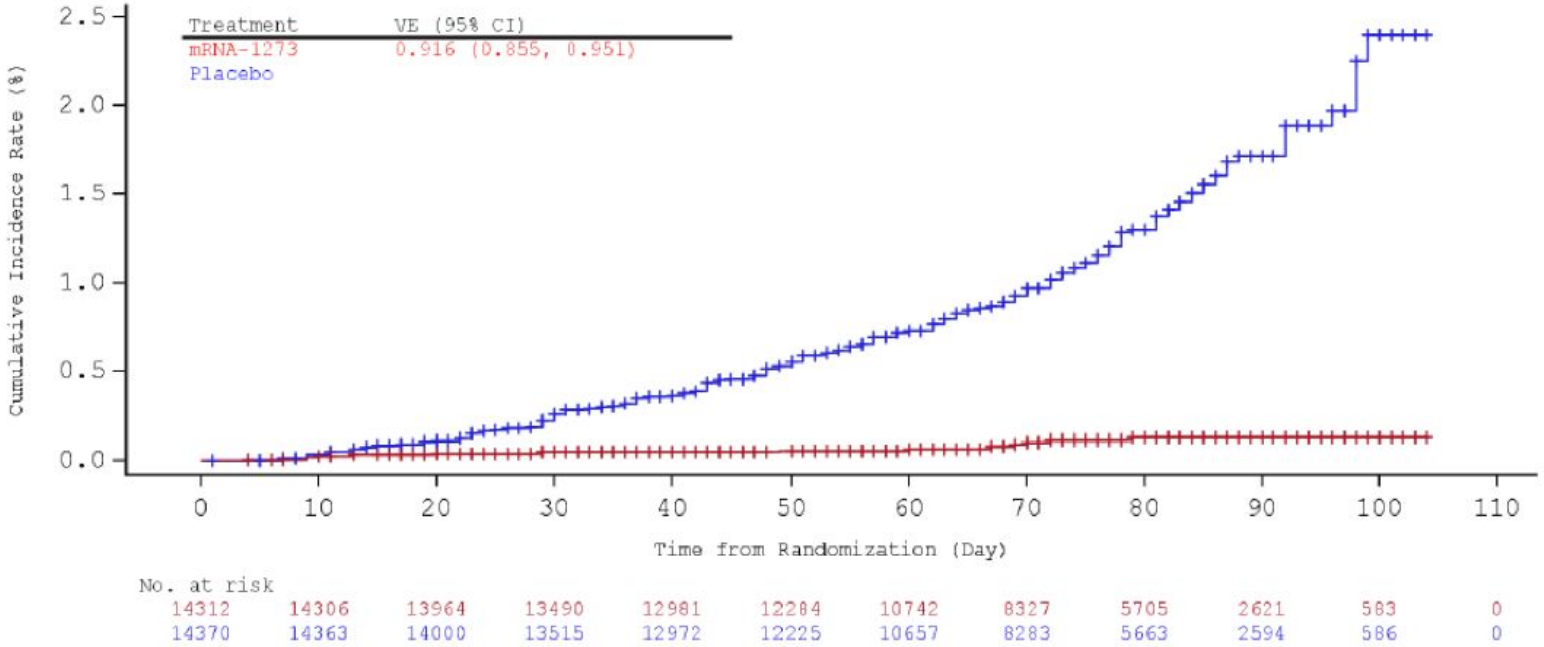
But Does This Generalize to All Possible Patients? A Meta-Analysis

- In 2011, a group of researchers performed a meta-analysis of the studies on the treatment.
- There were several studies but not many could be included in a scientifically sound manner: “We excluded evidence from nineteen studies due to non-randomisation, incomplete or unpublished data.” Only three randomized controlled trials remained, ***one of which is the one we have studied in this section***.
- Conclusion of the meta-analysis:
  - “There is low quality evidence that BoNT injections improved pain, function, or both better than saline injections and very low quality evidence that they were better than acupuncture or steroid injections. … Further research is very likely to have an important impact on the estimate of effect and our confidence in it. Future trials should standardize patient populations, treatment protocols and comparison groups, enlist more participants and include long-term outcomes, cost-benefit analysis and clinical relevance of findings.”

When Can we Make Causal Conclusions?



Causality in the Real World: Covid Vaccines



Source: FDA