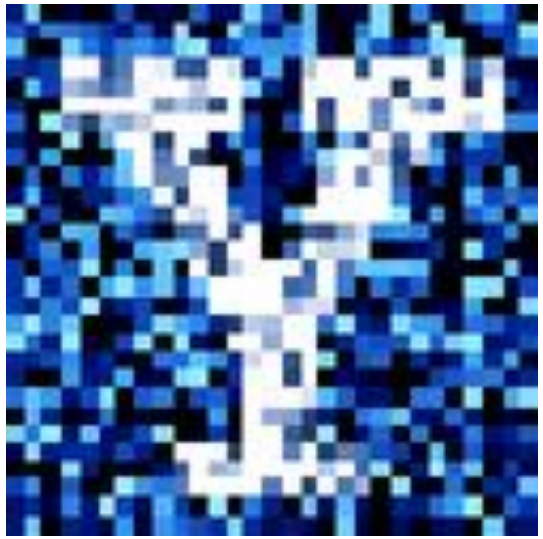
YData: Introduction to Data Science

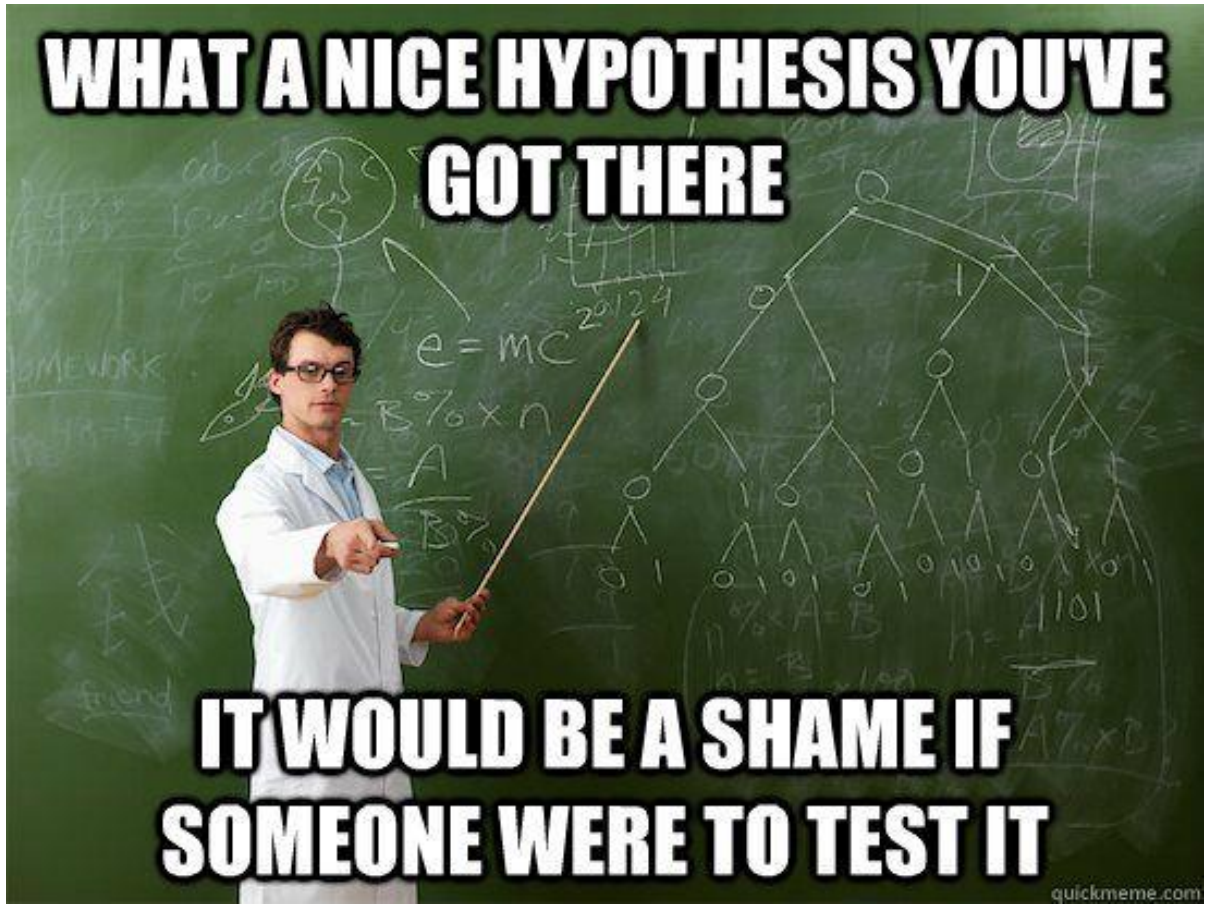


Lecture 19: A/B testing

Overview

Hypothesis test continued

- Review of terminology and concepts
- "A/B" testing



Review

Null and Alternative hypotheses

Null hypothesis

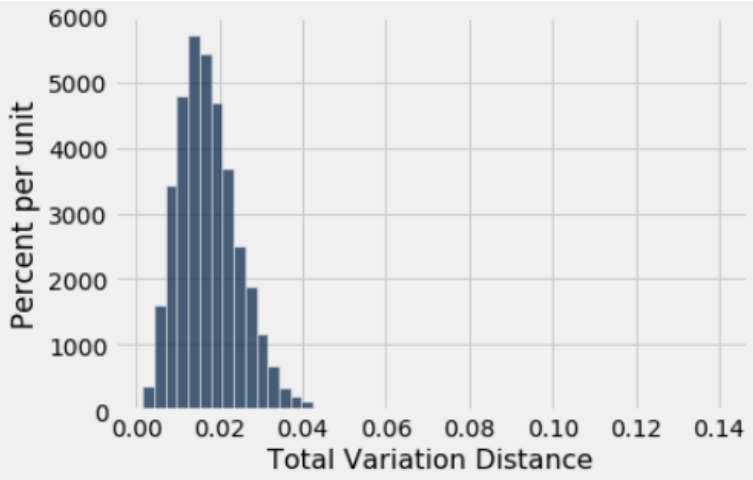
- A hypothesis that gives a well-defined chance model about how the data were generated
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics

Alternative hypothesis

- A different view about the origin of the data

A test statistic is the statistic we choose to simulate in order decide between the two hypotheses

Alameda county jury "null distribution"

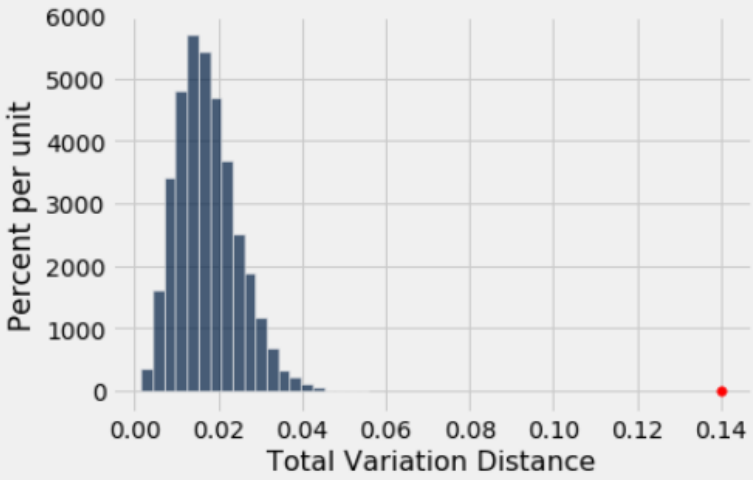


Testing the null hypothesis

To resolve choice between null and alternative hypotheses:

- We compare the observed test statistic to the statistic values in the null distribution
- If the observed statistic is not consistent with the null distribution, then we can reject the null hypothesis
  - And we accept the alternative hypothesis

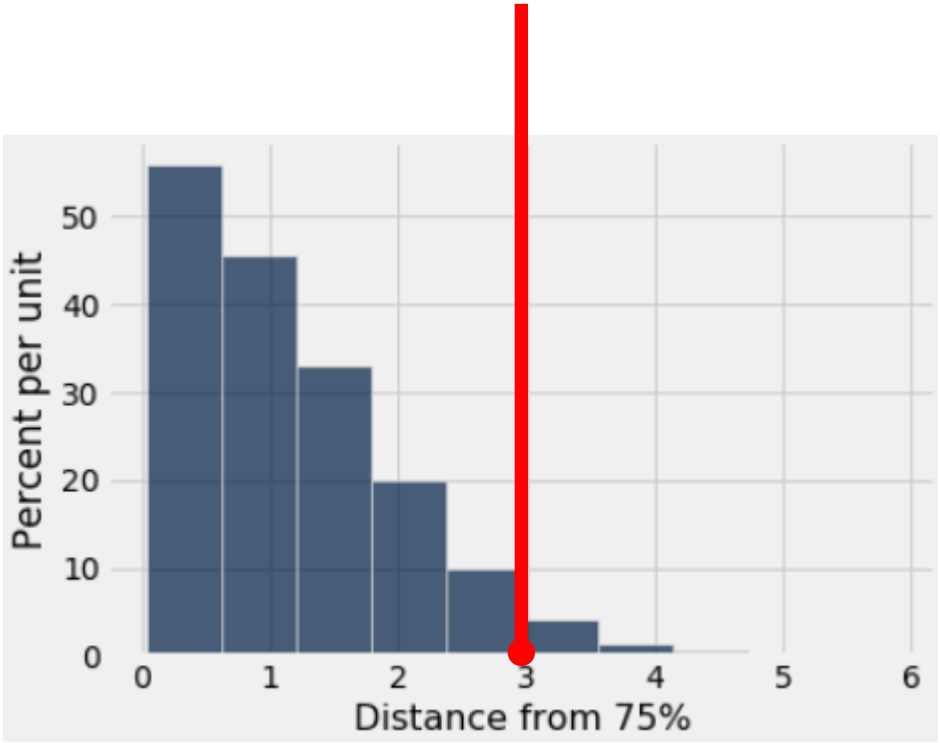
Alameda county jury "null distribution"



The p-value

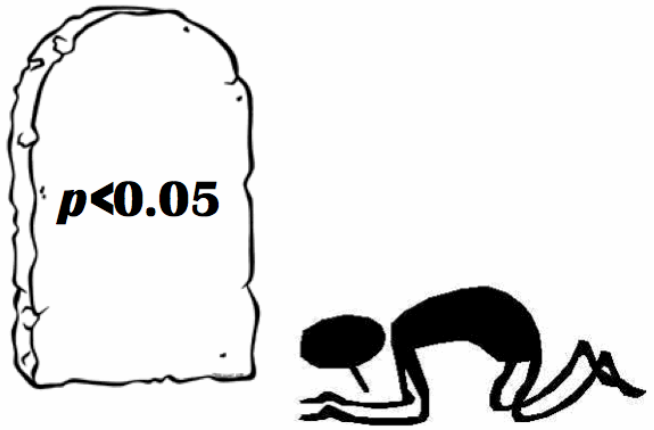
The p-value is the probability, that we get a statistic as or more extreme than the observed statistic from the null distribution

- P(Null_Stat ≥ obs_stat | H_0)



If the P-value is small, this is evidence against the null hypothesis and the results are often called "statistically significant"

- Convention, p-value < 0.05



Testing errors

If we always reject the null hypothesis when the p-value is less than 0.05, then:

- If the null hypothesis is true, we should make type I errors 5% of the time
- This is sometimes called the "false discovery rate"

In these cases, the test statistic will seem extreme just by chance

If we run many tests, we can expect to make some false discoveries

	Null is true	Alternative is true
Test rejects the null	"Type I error"	✓
Test doesn't reject the null	✓	"Type II error"

A/B Testing

Comparing two samples

We can use tests to compare parameters for values from two groups:

- Group A vs. 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:

- By Data Scientists: "A/B testing"
- By Statisticians: "Two independent samples hypothesis test"

Let's explore this in Jupyter!

Baby birth weights

Example: 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 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 didn't smoke were heavier, on average, than the babies of the smokers.

Test statistic

Group A: smokers

Group B: non-smokers

Statistic: Difference between average weights

- Group B average - Group A average

Large values of this statistic favor the alternative

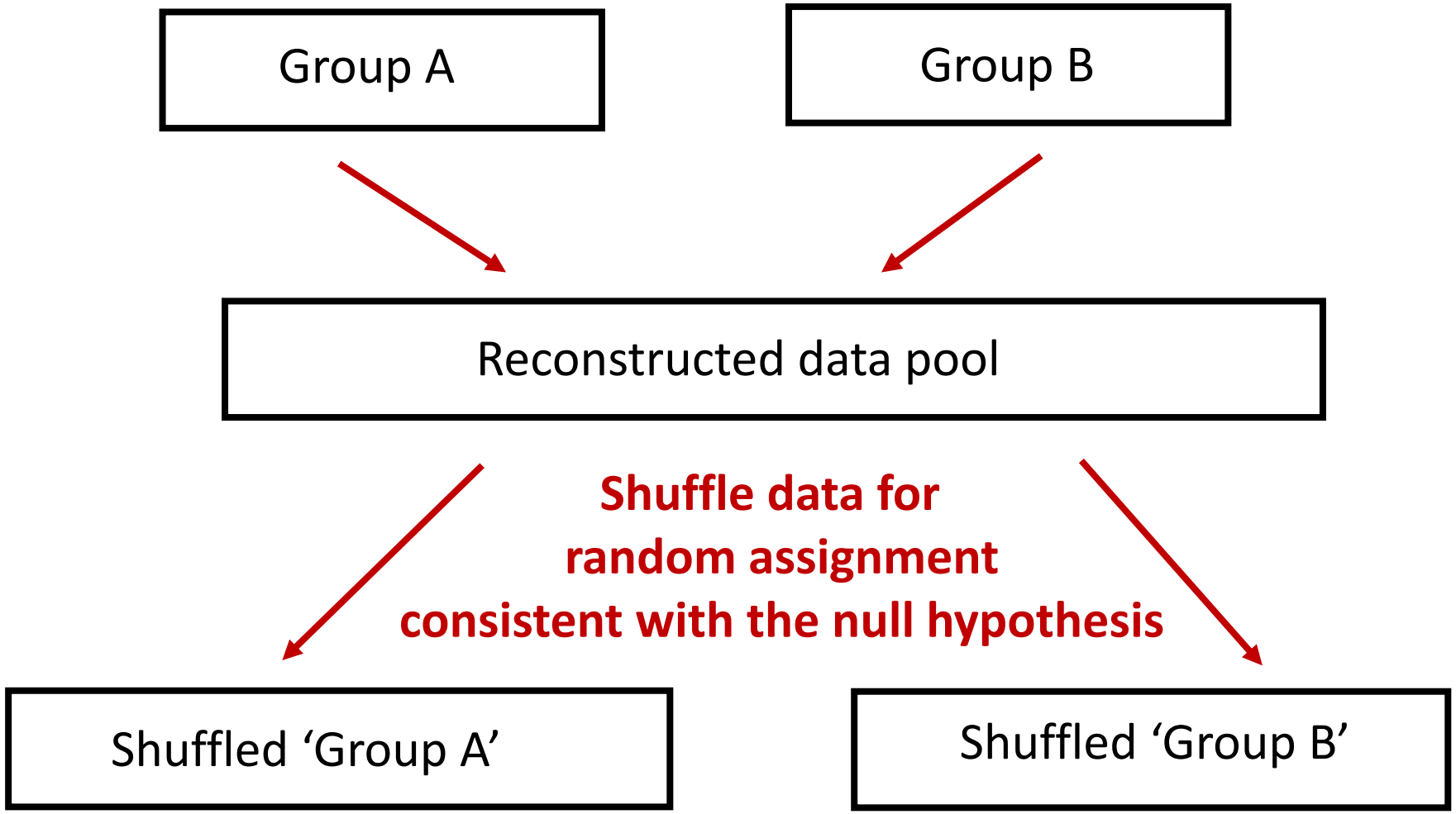
Simulating under the null hypothesis

If the null is true, all rearrangements of the birth weights among the two groups are equally likely

Plan:

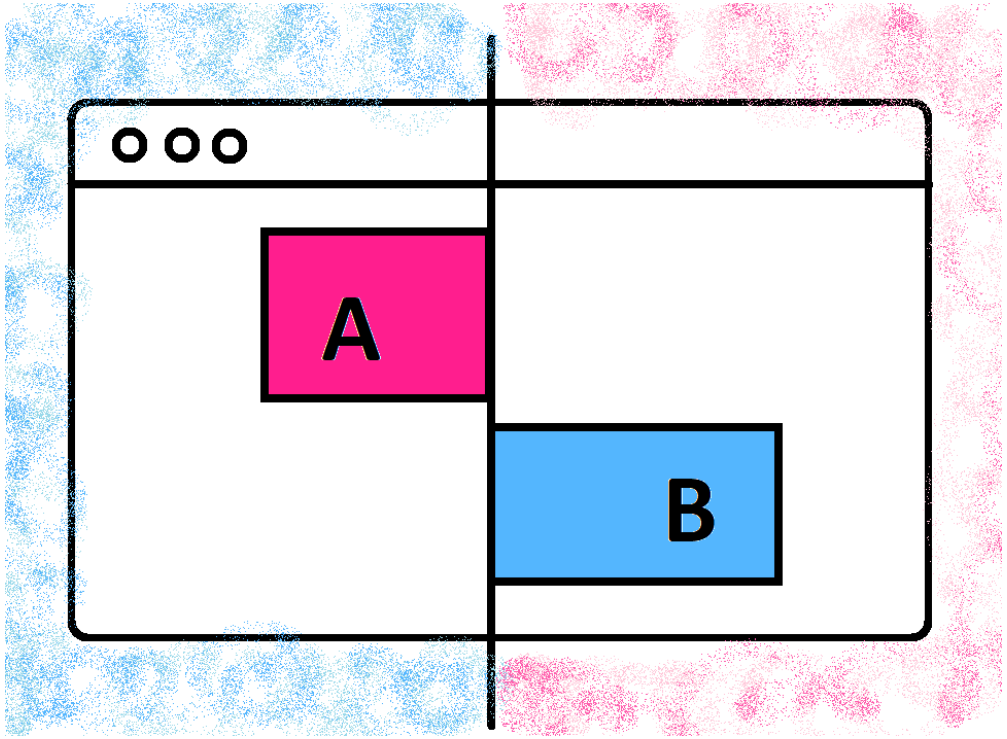
- Shuffle all the birth weights
- Assign some to "Group A" and the rest to "Group B", maintaining the two sample sizes
- Find the difference between the averages of the two shuffled groups
- Repeat

Create the null distribution!



One null distribution statistic: Average B – Average A

A/B testing of web sites



https://towardsdatascience.com/a-b-testing-the-basics-86d6d98525c9

Deflategate

2015 AFC Championship Game

Deflategate


'Deflategate' returns, focus on Tom Brady's destroyed cellphone

POSTED 9:54 AM, MARCH 5, 2016, BY CNN WIRE, *UPDATED AT 10:33AM, MARCH 5, 2016*

Wikipedia:

- The 2015 AFC Championship Game football tampering scandal, commonly referred to as Deategate, or Ballghazi

Claim: New England Patriots deflated footballs

Data:

- Pressure (PSI) was measured in 11 Patriots and 4 Colts footballs at halftime

Null hypothesis

The 4 Colts footballs are like a sample drawn at random without replacement from all 15 balls.

To test this hypothesis, repeat this process:

- Randomly permute all 15 balls
- Label 11 of them "Patriots" and the remaining 4 "Colts"
- Compare the averages of the two groups

Let's explore this in Jupyter!

AstraZeneca Vaccine

Science & technology

Vaccination vacillation

EU countries pause AstraZeneca's covid-19 jab over safety fears

An abundance of caution could well backfire

Example: AstraZeneca Vaccine

On March 15th 2021, France, Germany and Italy announced they were halting use of the AstraZeneca vaccine.

Why? A Norwegian medical regulator reported four cases of blood clotting in adults given the vaccine. Similar reports came from Denmark, Italy and Austria.

The World Health Organization (WHO) and European Medicines Agency (EMA) said they have no reason to believe the vaccine is unsafe.

Science & technology

Vaccination vacillation

EU countries pause AstraZeneca's covid-19 jab over safety fears

An abundance of caution could well backfire

AstraZeneca Vaccine (from The Economist)

The point at issue, as so often in medicine, is the disentangling of causation from correlation, especially when it comes to medicines given to millions of people. Blood clots are common. So, increasingly, are covid-19 vaccines. The EMA reckons there have been 30 “thromboembolic events” among around 5m people who have been given AstraZeneca’s vaccine. That some people with blood clots have also had a covid-19 vaccine is, by itself, no more remarkable than the fact that some of them will probably have taken vitamin supplements, or paracetamol, or breakfast. The question is whether the rates are higher than would otherwise be expected—which they do not seem to be. “The number of [blood clots] in vaccinated people is no higher than the number seen in the general population,” says the EMA.

Discussion Question: A/B Testing of Covid Data

Suppose that clinical trials of the AstraZeneca vaccine resulted in data of this form:

How would we perform an A/B test to decide whether or not there is excess risk of blood clotting, compared with random chance?

Treatment	Symptoms	Thrombosis
Placebo	False	False
Vaccine	False	False
Placebo	True	False
Vaccine	False	False
Vaccine	True	False
Placebo	False	False
Vaccine	False	False
Vaccine	False	False

... (9992 rows omitted)