YData: An Introduction to Data Science

Lecture 19: A/B Testing

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Reminders

- Topics this week: Hypothesis testing and causality
- Assignment 06 due on Thursday
- Break day next Wednesday (no class or OH)
- Midterm next Friday
  - Practice midterm posted; another will follow
  - Review session on Monday (no new topics next week)
  - Exam available Friday during normal class time (no class)
- Assignment 07 posted Friday; due April 1

Review

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.

Using the P-value

- If the P-value is small, this is evidence against the null hypothesis
- Conventions about "small":
  - Less than 5% (result is called statistically significant)
  - Less than 1% (result is called highly statistically significant)

Testing errors

- At the 5% level, if the null hypothesis is true we will reject it 5% of the time.
- This is sometimes called a "false discovery" (or Type 1 error)
- 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 (unless special care is taken)

Discussion Questions

Suppose the P-value of a test comes out to be about 0.5%.

(a) Fill in the blanks: The test supports the _____ hypothesis more than it supports the _____ hypothesis.

(b) True or false: There is about a 0.5% chance that the null hypothesis is true.

A/B Testing

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.
- An A/B test is just a particular type of hypothesis test

(DEMO)

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 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

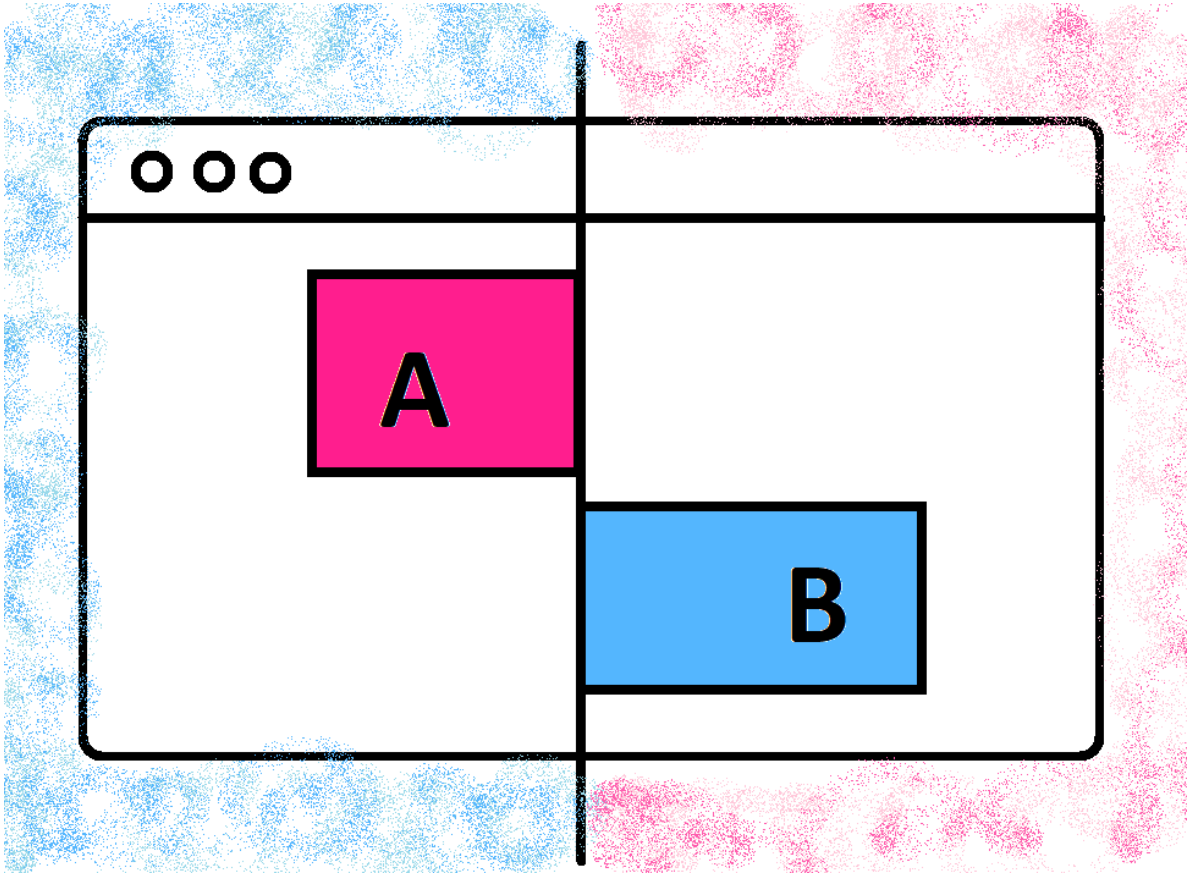
Simulating Under the Null

- 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

(DEMO)

A/B Testing of Web Sites



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

Deflategate

2015 AFC Championship Game

Deflategate

Wikipedia:
The 2015 AFC Championship Game football tampering scandal, commonly referred to as Deflategate, or Ballghazi

...

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

(DEMO)

AstraZeneca Vaccine

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)

- On March 15th 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–and similarly small—reports have come 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.

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:

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)

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