YData: An Introduction to Data Science

Lecture 19: A/B Testing

Jessi Cisewski-Kehe and John Lafferty
Statistics & Data Science, Yale University
Spring 2019

Credit: data8.org

Announcements

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, that 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)

Discussion Questions

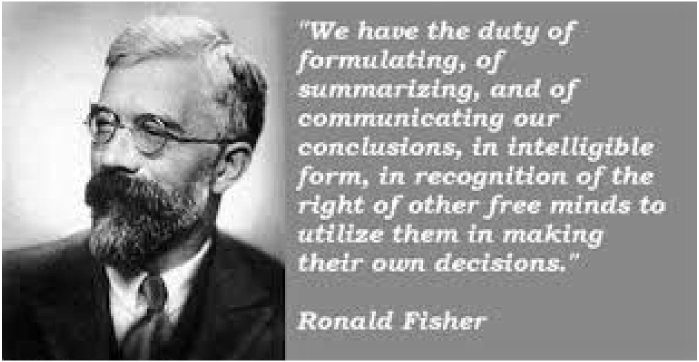
Suppose the P-value of a test comes out to be about 0.5%.

(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.

Origin of the Conventions

Sir Ronald Fisher, 1890-1962



"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

Ronald Fisher

Sir Ronald Fisher, 1925

“It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not.”

– Statistical Methods for Research Workers

Sir Ronald Fisher, 1926

“If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point ...”

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.

(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)

Deflategate

2015 AFC Championship Game

Syracuse, NY
11:04 AM ET

DEVELOPING STORY

PATRIOTS UNDER PRESSURE IN 'DEFLATEGATE' SCANDAL

Tim Green | Former NFL Player

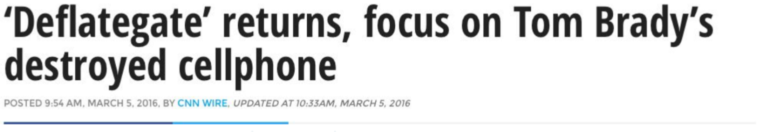
LIVE CNN 11:04 AM ET

R PARTS OF NEW YORK UP TO BOSTON, WITH WINDS OVER 60 MPH ▶ RELIABLE SOURCES

Deflategate

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

...



‘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*

(DEMO)

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)