CMPT 733 – Big Data Programming II

Hypothesis Testing

Instructor Steven Bergner

Course website https://sfu-db.github.io/bigdata-cmpt733/

Slides by: Jiannan Wang

Why Hypothesis Testing?

We want to make a claim from our data

But, data is just a sample

How to prove our claim in this situation?

Using Hypothesis Testing

Example

- Claim: A data scientist earns more money than a data engineer
- Data: A sample of 50 data scientists and 50 data engineers
- Result: 100K vs. 70k

Can we use this result to prove that our claim is correct?

Hypothesis Testing

Equivalent Terms

- Hypothesis == Claim
- Hypothesis Testing == Claim Proving

Key Idea

- Prove by contradiction

Analogy

- How to prove: There exists no smallest positive rational number.
- Hint: a rational number is any number that can be expressed as the fraction a/b of two integers

Alternative and Null Hypotheses

Alternative Hypothesis (H_a)

- This is the claim that you want to prove it's correct

Null Hypothesis (H_0)

- The opposite side of H_a

Possible Outcomes

- Reject H_0 (a contradiction is found) ➔ Accept H_a
- Fail to reject H_0 (no contradiction is found)

Example

Alternative Hypothesis (H_a)

- A data scientist earns more money than a data engineer

NULL Hypothesis (H_0)

- A data scientist earns less (or equal) money than a data engineer

If H_0 is true, what's the probability of seeing:

- ~~Data Scientist (100 K) vs. Data Engineer (70 K)~~
- Salary(Data Scientist) – Salary(Data Engineer) >= 30 K

This is called P-value

Make a decision based on p-value

We hope that

- p-value is as low as possible so that we can reject H_0 (i.e., accept H_a)

Level of Significance (e.g., $\alpha = 0.01$)

- How low do we want p-value to be?

Level of Confidence (e.g., $c = 1 - \alpha = 99\%$)

- How confident are we in our decision?

P-Hacking (Cheating on a P-Value)

Common Mistakes

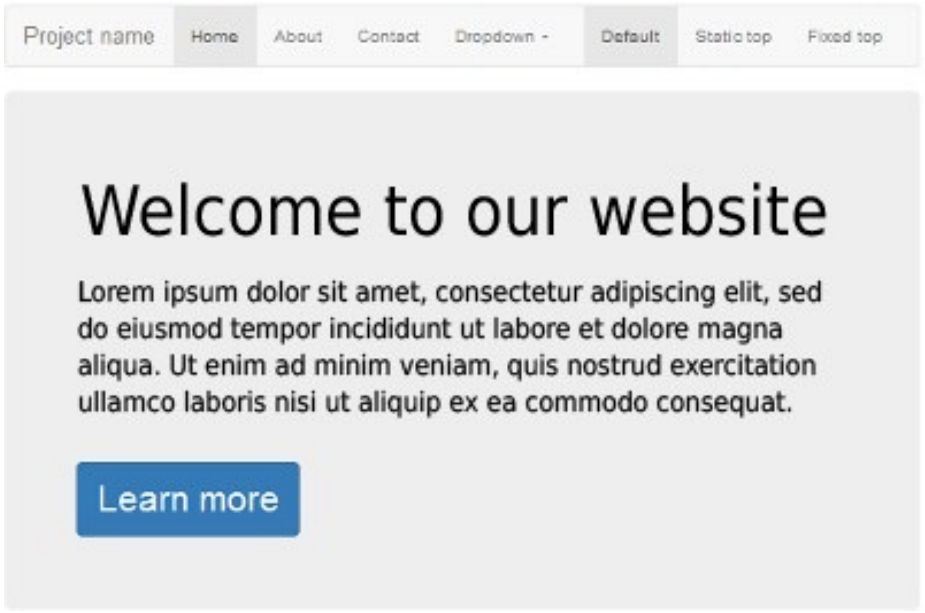
1. Collect data until the hypothesis testing is passed
2. Keep doing analysis on the same data until you find something significant

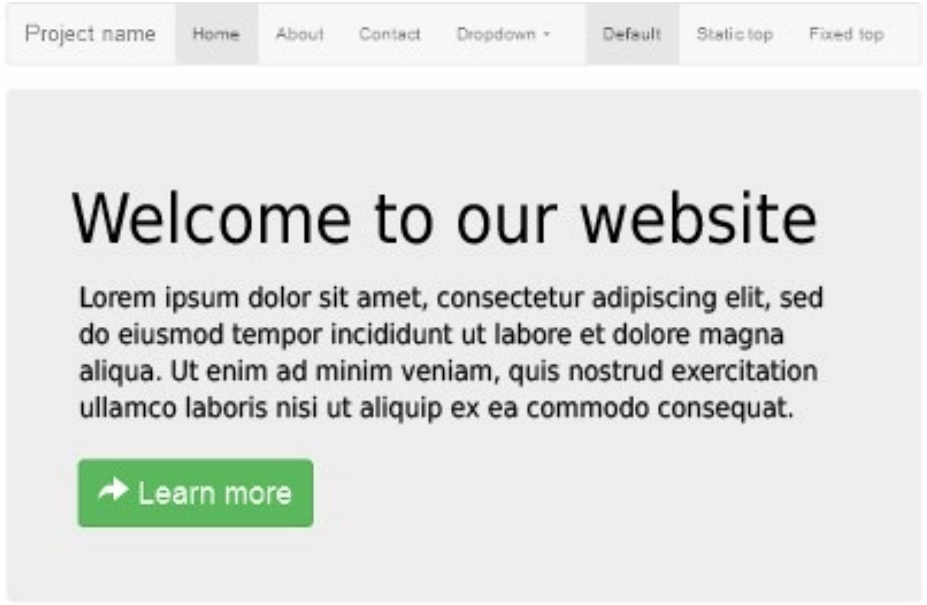
Solution

- You should know what you're looking for (H_0 and H_a) before you start
- Decrease the level of significance (e.g., $\alpha/2$ for two hypothesis tests on the same data)

A/B Testing

What UI is better?





Surprising A/B Tests

A. Get $10 off the first purchase. Book online now!

B. Get an additional $10 off. Book online now.

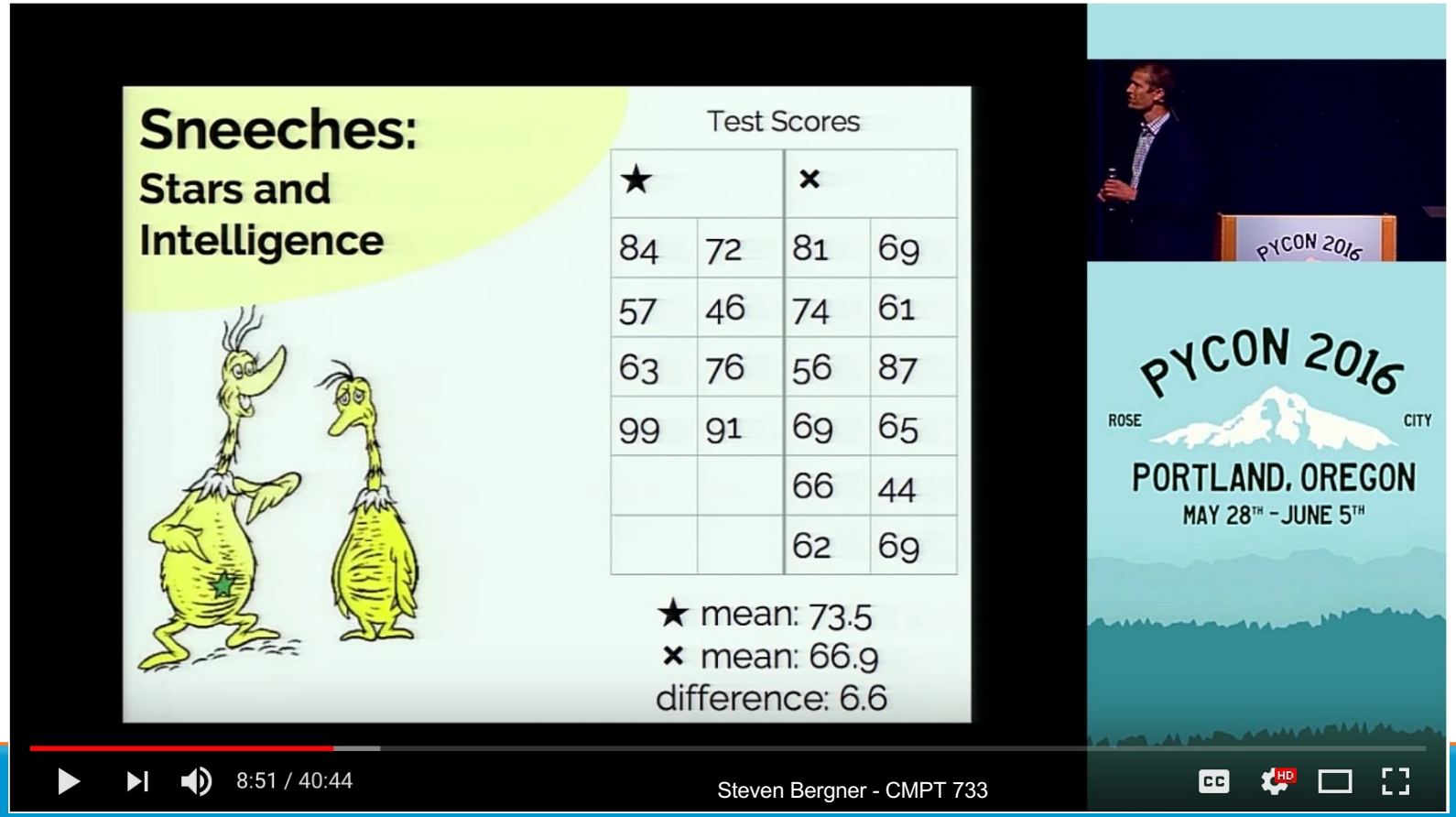
Control Button

Experiment Button

https://www.wordstream.com/blog/ws/2012/09/25/a-b-testing

Permutation Test

https://youtu.be/lq9DzN6mvYA?t=8m9s



Conclusion

- Hypothesis Testing
  - Null Hypothesis (H_0) and Alternative Hypothesis (H_a)
  - P-value and P-hacking
  - A/B Testing