Hypothesis Testing I: Making Decisions

Introduction to Data Science Algorithms
Jordan Boyd-Graber and Michael Paul
OCTOBER 4, 2016
Point Estimates Lie
Point Estimates Lie
So how can you make a decision?

- Error bars help, but not systematic
- Make the point that decisions need to not just look at single estimates but *distributions*
- Statistical Test: Deciding whether a hypothesis is true or not
Statistical Test Lingo

- Null hypothesis
- Test statistic
- p-value
- p-hacking
Null hypothesis

Null Hypothesis
A statement that can be validated through a statistic derived from observations.

- Often status quo
- Goal prove false: “reject the null”
- Phrased in terms of distributions

Examples
- Average body temperature 98.6?
- Voting republican and education independent?
Body temperature

$n = 130, \bar{x} = 98.249, \text{ standard deviation } s = 0.7332.$

- Not exactly equal (but wouldn’t expect that)
- Is the difference meaningful?
- Null hypothesis, $H_0 : \mu = 98.6$
- Alternative hypothesis, $H_a : \mu \neq 98.6$
What can happen

- **Correct** (True, True)
- **Type I False Positive** (True, False)
- **Type II False Negative** (False, True)
- **Correct** (False, False)
Boy who cried wolf

- Null hypothesis (status quo): no wolf
Boy who cried wolf

- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn’t)
Boy who cried wolf

- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn’t)
- Second error, Type II: villagers believed there was no wolf (when there was)
Boy who cried wolf

- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn’t)
- Second error, Type II: villagers believed there was no wolf (when there was)
- Type I and Type II in that order
Test Statistic

- Measurement of how far observations deviate from null hypothesis (e.g., $\bar{x}$ far from $\mu$)
- Test statistic is paired with a distribution that measures deviation
- Lower probability test statistics let you reject the null
$p$-value

- Probability of null hypothesis being true
- Lower is better
- Common critical values $\alpha$: 0.05, 0.01
- We’ll see examples in a bit
p-hacking

- Rerunning / changing experiments to reject the null
- Discuss at the end of today