Unit 2: Foundations for Inference

2. Hypothesis Testing

10/11/2017

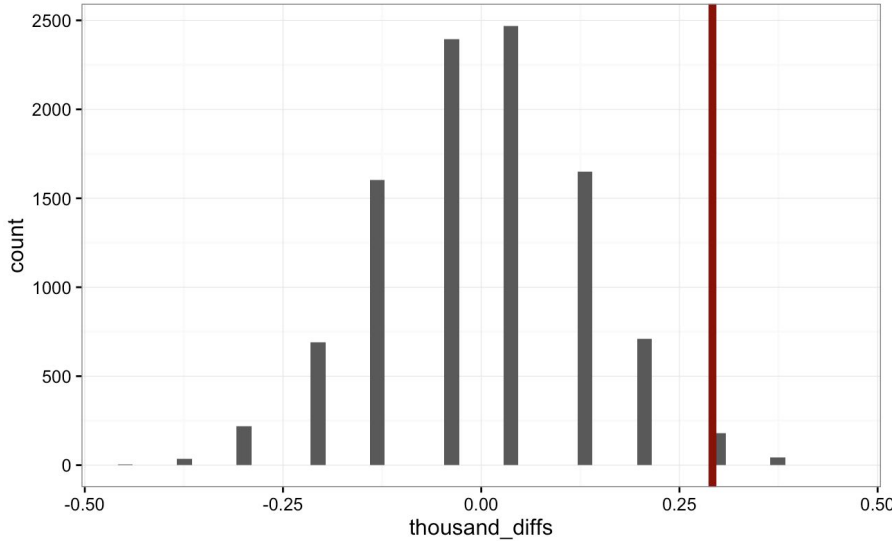
Recap from last time

1. We generally don't want to make claims about samples, we want to make claims about populations (or the processes that generated the samples)
2. We can use randomization to ask what inferences our sample license about the population
3. We are always talking about degrees of evidence. We can never have certainty.

Today: Formalizing the inferential process

1. Null hypothesis testing is a framework for quantifying evidence
2. Whenever we pick a standard of evidence that trades off Type I and Type II errors
3. We generally want to use two-sided tests, increasing our standard for evidence

Are women less likely to receive promotions than men?



If promotion is independent of gender, we should see a difference like the one we observed l*ess than 1% of the time*.

The Null Hypothesis Testing framework

1. "There is nothing going on" (**Null Hypothesis**)
   The *process* of promotion is independent of gender
   We observed results that *look* dependent due to chance

2. "There is something going on" (**Alternative Hypothesis**)
   The *process* of promotion is dependent of gender
   We observed results that *look* dependent because they *are dependent*

A hypothesis test is like a jury trial

H_0: Defendant is innocent

H_A: Defendant is guilty

We then collect the data, and present the evidence. Then we make a judgment.

"Could we plausibly see these data by chance?" (Null Hypothesis is true)

Ultimately we must make a decision. How unlikely is unlikely enough?

Picking a standard of evidence

In our court system, the defendant is innocent until proven guilty.

How much evidence do we need to conclude guilt?

Probable cause	Search warrant
Some credible evidence	
Substantial evidence	
Preponderance of evidence	Civil cases involving only money
Clear and convincing evidence	
Beyond a reasonable doubt	Criminal proceedings

Inferential decision errors

Inference

Truth

	Do not reject H_0	**Reject H_0 in favor of H_A**
H_0 True	Correct	Type I Error
H_A True	Type II Error	Correct

Increasing our standard of evidence yields fewer Type I Errors, but more Type II Errors.

You can't avoid this! You just have to decide how important each type of error is.

A hypothesis test as a jury trial

If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty".

- The jury **does not say that the defendant is innocent**, just that there is not enough evidence to convict.
- The defendant may, in fact, be innocent, but the jury has no way of being sure.

Said statistically, we fail to reject the null hypothesis.

- We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
- Therefore we never “accept the null hypothesis.”

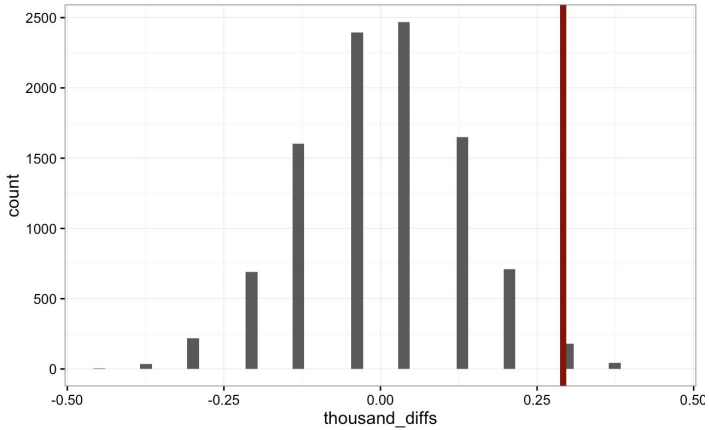
All together: The jury trial metaphor for hypothesis testing

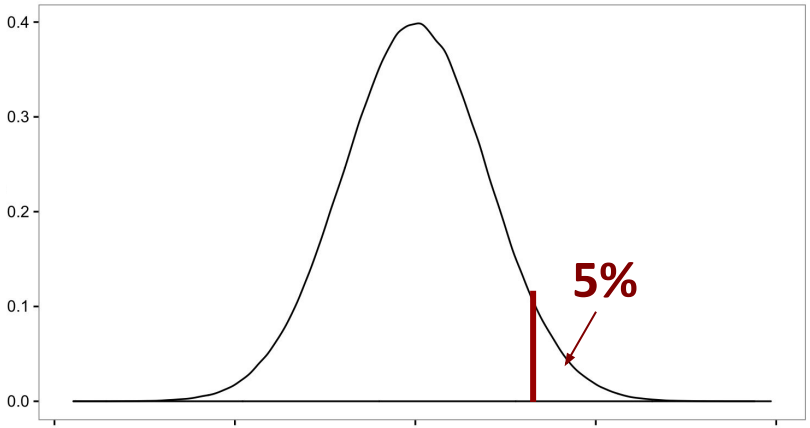
- We start with a null hypothesis (H0) that represents the status quo.
- We also have an alternative hypothesis (HA) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation (today) or theoretical methods (later in the quarter).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.
- The burden of proof is on the alternative hypothesis!

How do we pick our standard of evidence?

The standard criterion for many questions is **p = .05**

- We reject the null hypothesis if the probability of observing the empirical data under the null hypothesis is less than 5% (1/20 times)





Case study: Cardiac arrest at the wrong time of the year

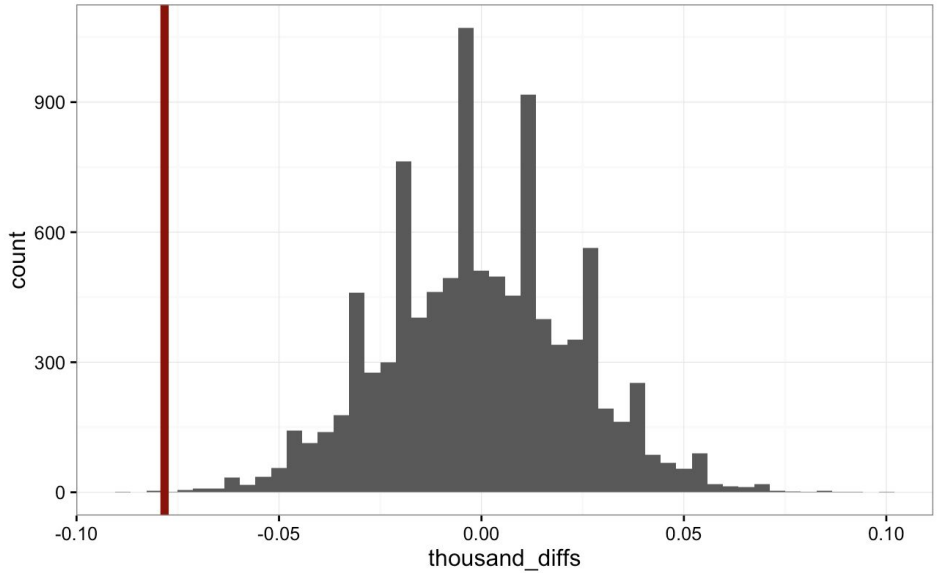
- Doctors at teaching hospitals sometimes attend international conferences to catch up on the latest scientific discoveries in their area
    - E.g. in 2006 ~19,000 cardiologists attended the American Heart Association annual meeting
- But what happens to patients while these doctors are away? Are they more likely to have negative outcomes?
- Jena et al. looked at 30-day mortality rates among patients admitted during the dates of national cardiology meetings compared to non-meeting dates.

Is this an observational study or an experiment?

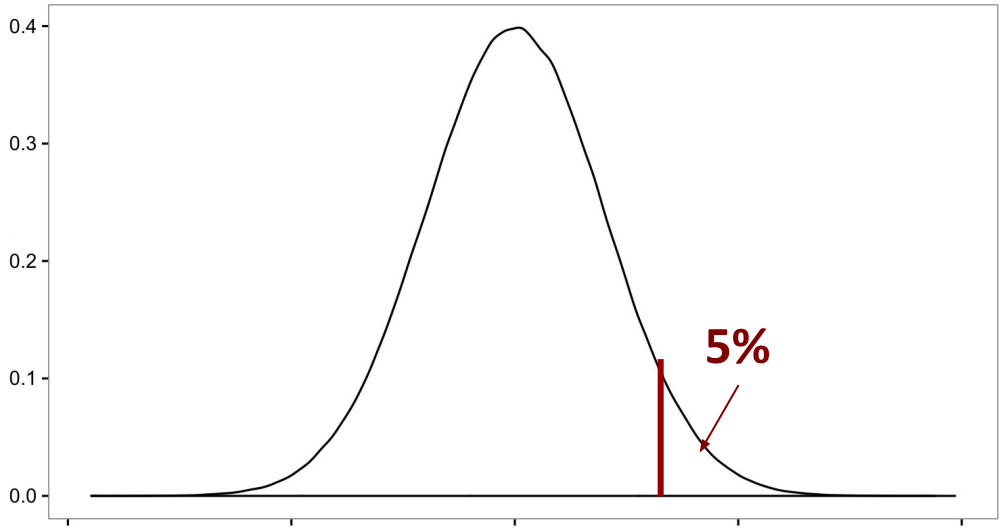
Observational study

Jena et al. (2015, JAMA Internal Medicine)

Let’s look at the data

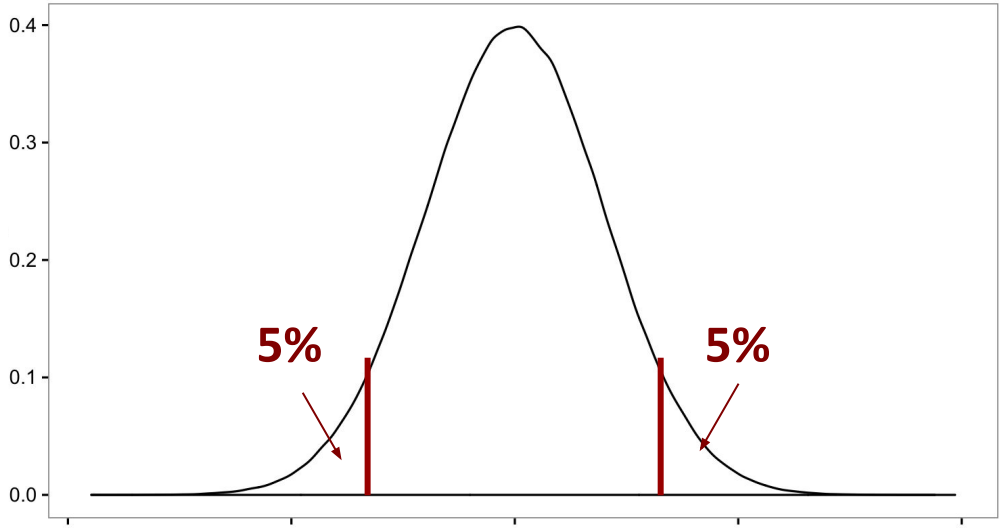


Why we need two-sided hypothesis tests



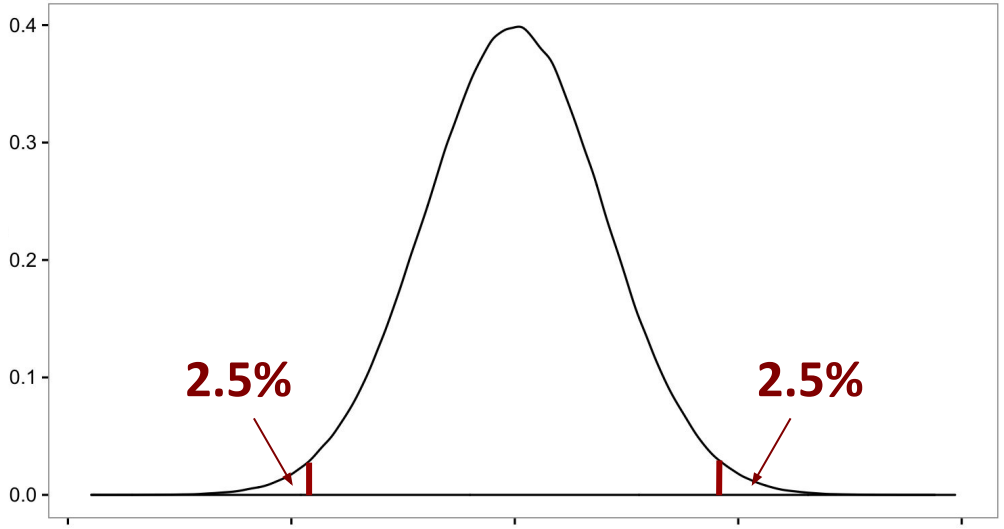
If we had defined our Alternative Hypothesis as “mortality increases,” we would have never found this effect!

Practice question: Why is this not the solution?



This increases our Type I Error rate! We now reject twice as many null hypotheses!

Solution: Raise the standard of evidence on both sides of the distribution



This keeps our total Type I error rate the same.

Practice question: How do we know it was the conferences?

What if people are just sicker during conference days for some reason? How do we know it was the doctors being absent?

What if we compare teaching hospitals to non-teaching hospitals?

Key ideas

1. Null hypothesis testing is a framework for quantifying evidence
2. Whenever we pick a standard of evidence that trades off Type I and Type II errors
3. We generally want to use two-sided tests, increasing our standard for evidence