GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS, PART 2

K. Nathaniel Tucker

Quant, Google

LEARNING OBJECTIVES

- Explain what a confidence interval is
- Test a hypothesis within a sample case study
- Use hypothesis tests and confidence intervals with scipy

COURSE

PRE-WORK

PRE-WORK REVIEW

- Use descriptive statistics to understand your data

OPENING

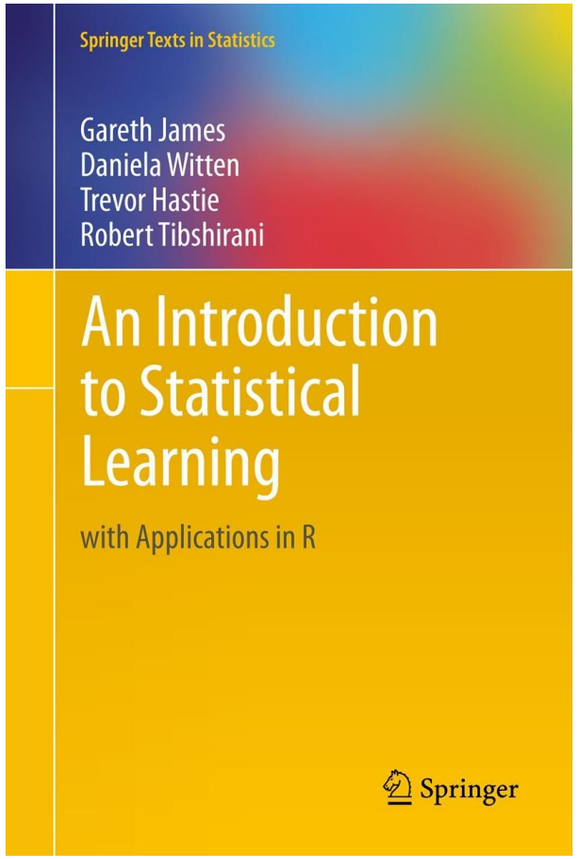
STATISTICS FUNDAMENTALS, PART 2

LAST SESSION

- Any questions from last class?
- Exit tickets

DATA SOURCE

- Today, we'll use advertising data from an example in *An Introduction to Statistical Learning*.



Confidence Intervals

Onto the Whiteboard

- How can we estimate a range around a true value?
- Sampling Distributions
- Remember what is random!!

HYPOTHESIS TESTING

HYPOTHESIS TESTING

- How can we tell the difference between two groups of observations (e.g. smokers vs. non-smokers)?

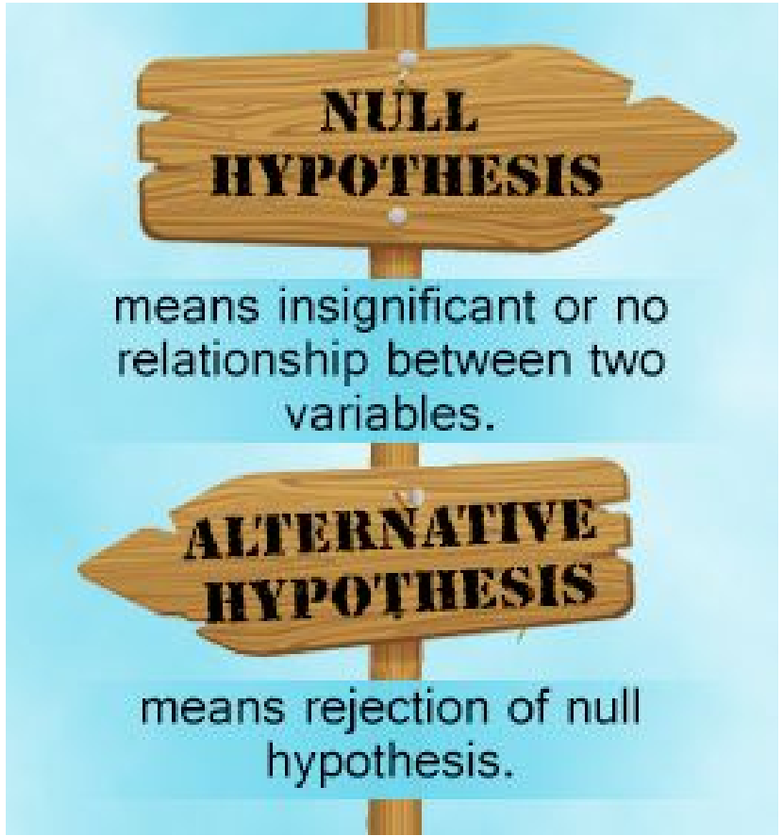
- Imagine we are testing the health of smokers vs. non-smokers. At a cursory glance, our results may show that smokers are marginally healthier than non-smokers.

- Are they healthier due to random chance or is there a statistically significant difference? Maybe we happened to assemble a strange group of smoking triathletes and a group of non-smoking couch potatoes.

- This is where hypothesis testing can help.

HYPOTHESIS TESTING STEPS

- First, you need a hypothesis to test, referred to as the *null hypothesis*. The opposite of this would be the *alternative hypothesis*.



HYPOTHESIS TESTING STEPS

- For example, if we want to test the relationship between gender and sales, we may have the following hypotheses.

- Null hypothesis: There is no relationship between Gender and Sales.

- Alternative hypothesis: There is a relationship between Gender and Sales.

HYPOTHESIS TESTING STEPS

- Once you have your hypotheses, you can check whether the data supports rejecting the null hypothesis or failing to reject the hypothesis.

- **Note**: Failing to reject the null is **NOT** the same as accepting the alternate. While the alternative hypothesis **might** be true, we don't have enough data to support that claim specifically.

- Keep this in mind so you don't overstate your findings.

DEMO

HYPOTHESIS TESTING CASE STUDY

HYPOTHESIS TESTING CASE STUDY

- We're going to walk through Part 1 of the guided-demo-starter-code notebook in the class repo for lesson 4.

- There are several questions to answer. We'll answer those questions in small groups and then discuss with the class.

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS

1. What is the null hypothesis?
2. Why is this important to use?

DELIVERABLE

Answers to the above questions

VALIDATE YOUR FINDINGS

VALIDATE YOUR FINDINGS

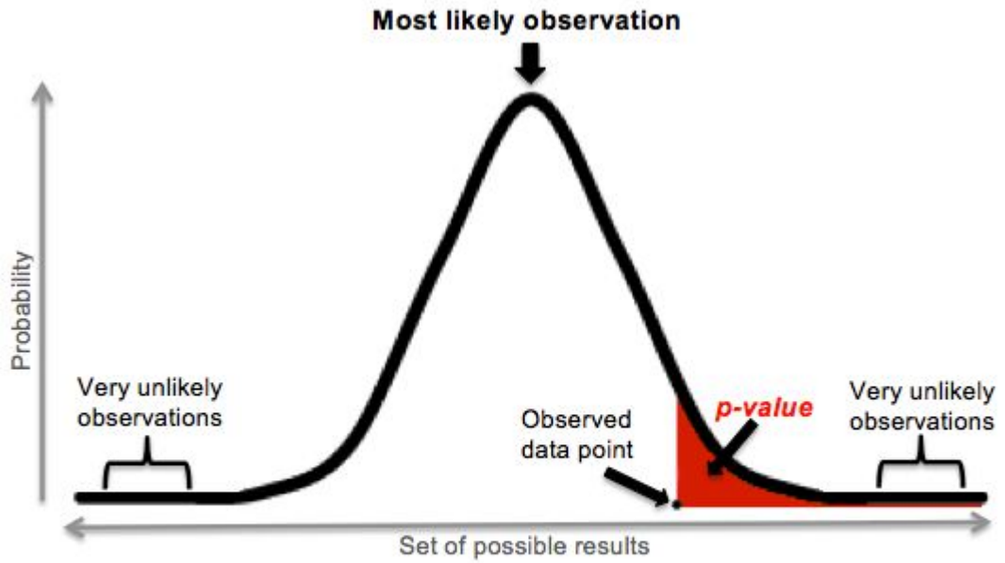
- We know how to carry out a hypothesis test, but how do we tell if the association we found is *statistically significant*?

- *Statistical significance* is the likelihood that a result or relationship is caused by something other than random chance.

- Statistical hypothesis testing is traditionally employed to determine if a result is statistically significant or not.

VALIDATE YOUR FINDINGS

- Typically, a cut point of 5% is used. This means that we say something is statistically significant if there is a less than a 5% chance that our finding was due to random chance alone.



A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

VALIDATE YOUR FINDINGS

TABLE 1

Relationship between Common Language and Hypothesis Testing

COMMON LANGUAGE	STATISTICAL STATEMENT	CONVENTIONAL TEST THRESHOLD
"Statistically significant" "Unlikely due to chance"	The null hypothesis was rejected.	$P < 0.05$
"Not significant" "Due to chance"	The null hypothesis could not be rejected.	$P > 0.05$

VALIDATE YOUR FINDINGS

- When we present results, we say we found something significant using this criteria.

- We will use an example to dive further into this and understand p-values and confidence intervals.

DEMO

P-VALUES AND CONFIDENCE INTERVALS CASE STUDY

P-VALUES AND CONFIDENCE INTERVALS CASE STUDY

- We're now going to walk through Part 2 of the guided-demo-starter-code notebook in the class repo for lesson 4.

- There are several questions to answer. We'll answer those questions in small groups and then discuss with the class.

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS

1. What does a 95% confidence interval indicate?

DELIVERABLE

Answers to the above questions

INDEPENDENT PRACTICE

INTERPRETING RESULTS

ACTIVITY: INTERPRETING RESULTS

EXERCISE

DIRECTIONS (35 minutes)

1. Using the lab-start-code-4, you will look through a variety of analyses and interpret the findings.
2. You will be presented with a series of outputs and tables from a published analysis.
3. Read the outputs and determine if the findings are statistically significant or not.

DELIVERABLE

Answers to the questions in the notebook

CONCLUSION

LAB REVIEW

LAB REVIEW

- Let’s review the answers to the questions in the labs.
- Any other questions?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project: Unit Project 2

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET