Lecture 6: Reproducibility, replication, etc., Part 2
Modeling Social Data, Spring 2019
Columbia University

Lina Tian

March 1, 2019

1 Introduction

This lecture is a continuation from the last one. Last class, we went through the first three questions of "how one should evaluate research results",

- Was the research done and reported honestly / correctly?
- Is the result "real" or an artifact of the data / analysis?
- Will it hold up over time?
- How robust is the result to small changes?
- How important / useful is the finding?

And this class, we will keep exploring these questions through:

- Common misunderstandings of α
- Effect size
- P-hacking

2 Reminder of definitions from last class

- Honesty
  - Was the data accurately collected and reported?
- Reproducibility
  - Can you independently verify the exact results using the **same data** and the **same analysis**?
- Replicability
  - Will the result hold up with **new data** but the **same analysis**?

3 Quiz—Clarifying common misunderstandings of α

- Question:
  - You do 1,000 experiments for 1,000 different research questions
  - Only 30% of these experiments investigate real effects
  - You set your significance level α to 5%
  - You use a small sample size such that your power 1 - β 35%
  - Given that one of these experiments shows statistical significance, what's the probability that it's a real effect?
- Solution:
  - Standard (yet confusing) way—Bayes rule
  - Easier way—Draw a tree

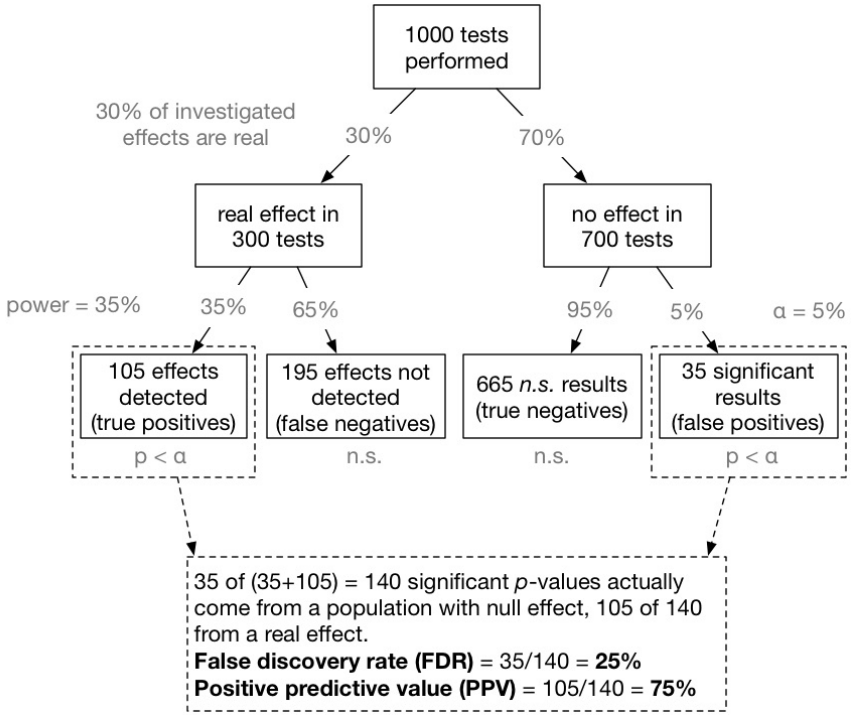

Figure 1: Tree diagram for this quiz

  - The answer we need is the 75%.
- Explanation & typical confusion:
  - **False positive rate:** α = P(show significance | no effect) = *5% in this problem*
  - **True positive rate:** $1 - \beta$ (power) = P (show significance | real effect) = *35% in this problem*
  - **False discovery rate:** P(no effect | show significance) = *25% in this problem*
  - Now, many people think that alpha means that 95% (a.k.a 1 - α) of the time we detect the real effect, **but 1 - α does not mean that**.
    - Given the definition of **false positive rate**, we see that only **75%** of the time we detect the real effect (very small compared to the 95% we thought)

    * To make it better—up the power or down the alpha (power of 35% is stupid–should be at least more than 50%, otherwise we might as well flip a coin and save our time)

4 Effect size–how big is that effect?

- definition on Wikipedia: a quantitative measure of the magnitude of a phenomenon
- **Cohen's d**
  - Imagine you have two distributions
  - The effect could be measured by only the *difference between the mean* of the placebo group and experiment group
  - However, if we want to take into account *variance*:
    * **Cohen'd** $= \frac{\mu_1 - \mu_2}{\sigma}$
    * μ_2 , μ_2 are the means of the 2 distributions
    * σ the same (pooled) standard deviation of both groups
- Implications
  - In real world, when we apply testing to a very big population, we can potentially have a very small distance in mean (very commonly around or below 0.2). However, since we have a big population, the variance is very small. Therefore the effect size can be big.
  - The fact that I reject the null does not tell me how big a difference the means are
  - Related term: AUC, the probability of superiority P(treatment ¿ control)

5 P-hacking—what happens when we get fooled by randomness

- Read this paper that discusses these studies and talk about the danger of p-hacking
- Study 1 (music from a time with high age contrast make people feel older)
- Study 2 (music from a time with high age contrast make people actually older)
- The problem with this study at first glance:
  - father's age is a weird and unreliable measurement
  - songs are different from one study to another
  - the sample size is too small (20)
- What they actually did:
  - they had 34 students in the sample, but they threw out a bunch when analyzing
  - they only had three songs
- What is wrong with it:
  - When doing exploratory data analysis and something shows up, and they decided its interesting immediately
  - they are doing selective overfitting
  - tested after every 10 students surveyed, and stopped surveying when found significance
  - Eventually, if you ask enough questions about the data–then you are gonna get significance with some question

- So how do we know that we are not fooling ourselves:
  - tell the difference between **confirmatory analysis** and textbfexploratory data analysis
  - when we write our paper, need to disclose everything that we did in the experiment
  - do not select and report
- A site that helps with study before you go into it:
  - they have couple check mark questions
  - Last question: are the data for the study ready? If the answer is yes, then not good. (experiment vs. observation)

6 Predicting employee satisfaction in Microsoft—Jake's "failed?" study

- Take a look at this article and the study at Microsoft that the author discussed.
- So what happened?
  - Data: all the emails, and a survey to all employees
  - Method:Random forest and logistic models
  - Predict: people would be happy with work-life balance
  - Result: Good results–good prediction
  - However, after a year, when they did a **pre-registering predictive models**
    - they pre-wrote the exact process and question they want to explore
    - they used the exact same survey result, but with email data updated
    - It failed quite badly, using the same method
    - why: the way people use email has completely changed over a year
- what can we learn from this?
  - It is important to do a pre-registering predictive models
  - So we do not just fall into trap of testing a bunch of hypothesis until one works
  - And it force you to think hard on what exactly you wanted
  - It avoids p-hacking
  - And this is what people should be doing

7 How to do R-markdown and makefile

R-markdown is similar to jupyter notebook. You can export a pdf or html from a structured block format r code. Makefile allows you to organize all the files needed to run your code and produce the result. It has dependencies of your files clearly laid out, and makes everyone's life easier.
If you want to check out the template Jake gives. See the R markdown write up and makefile write upin git-hub.

8 Homework 2 explanation (if needed)

- Problem 1
  - should be a one-line solution with r code
- Problem 2

  - do them on the markdown file
  - Problems are from bit by bit book

- Problem 3
  - we are forgetting the past faster and faster
  - Take the string 1883 and each year after that, measure the frequency of the appearance of the word compare to all other words
  - $Freq = \frac{numtimes\ 1883\ appeared\ in\ some\ year}{total\ num\ words\ printed\ in\ that\ year}$
  - Small plots: compare the half life of the words over all years (the bottom graph only showed 3 years)
  - Tool: Google books ngram viewer
  - Data: need to get it from the website, compresses, confusing, weird format, many extra data, super big file size (but most do not match what we need)
  - Only use the ones start with 1 in one-gram, instruction on github, and use shell to process the data first–use the fancy egrep stuff to clean (make sure we do not match 195743, but only 1957)
  - One line solution for all shell stuff
  - Submit all through github