Lecture 6: Reproducibility, replication, etc., Part 2
Modeling Social Data, Spring 2018
Columbia University

Sang Won Lee

March 01, 2019

1 Quiz

- You do 1,000 experiments for 1,000 different research questions
- Only 30 percent of these experiments investigate real effects
- You set your significance level alpha to 5 percent
- You use a small sample size such that your power (1-beta) is 35 percent
- Given that one of these experiments shows statistical significance, what's the probability that it's a real effect?

Professor's comment: "High school teachers do relatively well compared to students who just took a statistics class."

Given

- alpha = P(significance given no effect) = 5 percent
- (1-beta) = P(significance given effect) = 35 percent

Solution
Bayes Rule

$$\alpha = 0.05 = P(\textit{significance given no effect}) = \frac{\textit{(no effect given significance)} * \textit{P(significance)}}{P\textit{(no effect)}} \tag{1}$$

$$(1-\beta) = 0.35 = P\textit{(significance given effect)} = \frac{\textit{(effect given significance)*P(significance)}}{P\textit{(effect)}} \tag{2}$$

OR

- Alpha is the false positive rate, and not false discovery rate.
- False discovery rate in the quiz above is 35/(35+105)=35 percent.

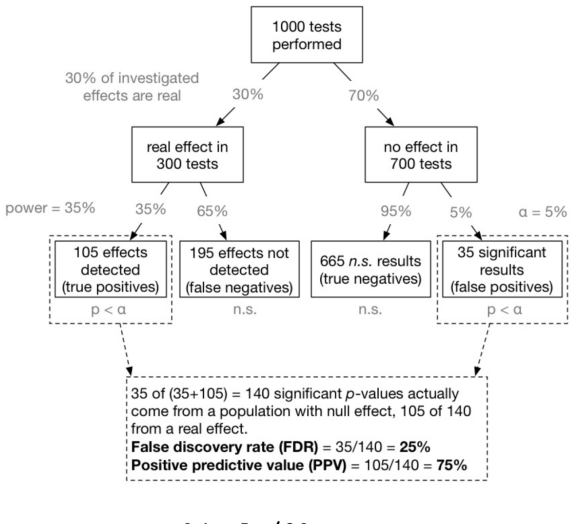

Figure 1: Answer provided during class.

2 Effect Size

- Effect size is anything that might be of interest.
- Effect sizes, in this case, are metrics that represent the amount of differences between two sample means.

The Cohen's d effect size is very popular in psychology. The following graphs show change in Cohen's d.

$$d = \frac{M1 - M2}{SD\ pooled} \tag{3}$$

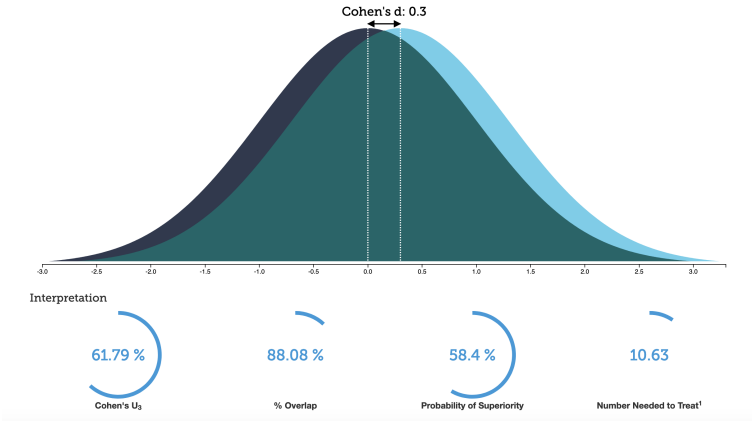

Figure 2: Cohen's d when it is 0.3.

Probability of superiority is when probability of treatment is greater than probability of control.
When effect size is larger, the probability of superiority increases.
Professor's comment: Rejecting the null hypothesis does not tell you difference, the effect size does.

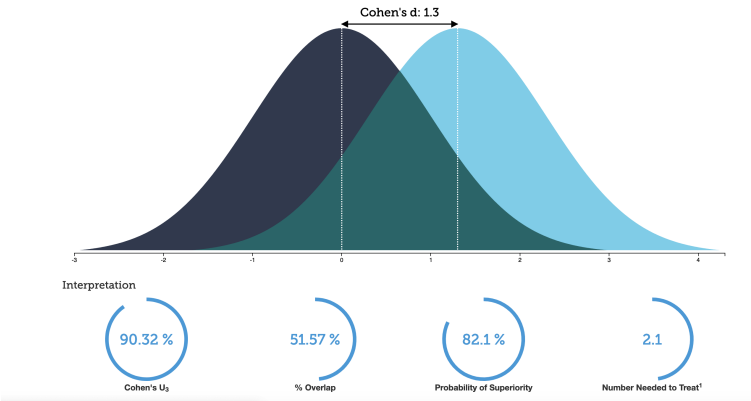

Figure 3: Cohen's d when it is 1.3.

3 P-Hacking

1. Study I: musical contrast and subject age
2. Study II: musical contrast and chronological rejuvenation

Study I: Reject null hypothesis. Listening to a children's songs does not make people feel older.
Study II: Investigated whether listening to old songs makes one really older.

Pretty sure listening to adult songs don't make one older. First of all, sample size is small. Songs are different in two studies. One can realize that there are small details that are weird. Actual title of the paper is "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant."

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked 20 34 **University of Pennsylvania undergraduates to listen** only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), $F(1, 18) = 1.01$, $p = .33$.

Figure 4: Paper without overfitting.

This paper is about manual overfitting, and selective presenting. These types of overfitting can be intentional or not intentional. Exploratory data analysis needs to be dealt very carefully. How do you know you are not fulling yourself? You can check if you are looking for patterns or finding a pattern. You should report everything you did in the paper (disclose all information). One can also pre-register what they are going to do at aspredicted.com. In this article, the authors accomplish two things. First, the authors show that despite empirical psychologists nominal endorsement of a low rate of false-positive findings (.05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. The authors present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and

report) statistically significant evidence for a false hypothesis. Second, the authors suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.