Lecture 6: Reproducibility and Replication II
Modeling Social Data, Spring 2019
Columbia University

March 1, 2019

Notes from sj2736

usepackageamsmath

1 Introduction

Jake began the lecture by pointing out that our Homework 1 submissions were generally non-reproducible, and provided an overview of general practices to ensure reproducibility in Homework 2. We began our discussion in class with two main questions: **How should one evaluate research results?** and **Will the result hold up with new data but the same analysis?**.

2 Crisis

We looked at the paper **Estimating the reproducibility of psychological science** which was a broad research project undertaken by a large number of scientists. 100 experimental and correlational studies were replicated. Of the replicated studies, 36% had statistically significant results whereas 97% of the original studies had significant results in their original publications. Importantly, 47% of the original effect sizes were in the 95% confidence interval of the replication effect size, which leads us to conclude that we should believe about half of what we read in the psychological literature.

3 What is an Effect Size?

So, what exactly is an effect size?

In statistics, an effect size is a quantitative measure of the magnitude of a phenomenon. Examples of effect sizes are the correlation between two variables, the regression coefficient in a regression, the mean difference, or even the risk with which something happens, such as how many people survive after a heart attack for every one person that does not survive. Source: Wikipedia: Effect Size

We looked at https://rpsychologist.com/d3/cohend/ to visualize Effect Sizes. The visualization is designed to show the real meaning of a type of effect size popularly cited in psychology. The main takeaways of this animation is summarized in the following statement, "Factors like the quality of the study, the uncertainty of the estimate and results from previous work in the field need to be appraised before declaring an effect "large"." Importantly, when comparing a control and test group with each other, we should consider not only the mean difference, but also the difference in variance between the distributions being studied.

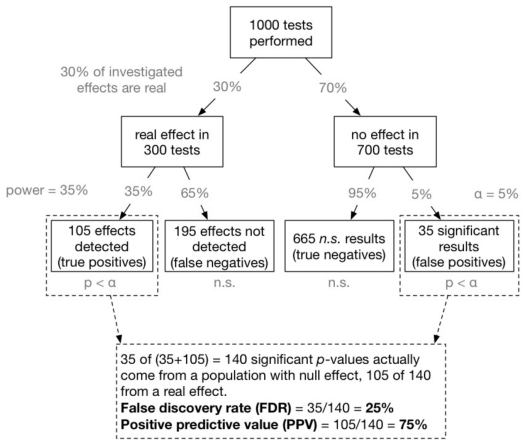
4 Misunderstandings in Statistics

We looked at the paper, **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations** and worked on the following problem to clarify our understanding of the probability of real effects utilizing Bayes Rule.

- 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 $1-\beta$ is 35%
- Given that one of these experiments shows statistical significance, what's the probability that it's a real effect?

Figure 1: problem from class

The solution is provided below:



bit.ly/fdrtree

Figure 2: solution from class

When doing this power it's important to remember the following when utilizing Bayes' Rule:

$$1 - \beta = \text{Power} = \text{P}(\text{stat sig} \text{ — } \text{real effect})$$
$$\alpha = \text{P}(\text{stat sig} \text{ — } \text{not real effect})$$
$$\text{False Discovery Rate} = \text{P}(\text{not real effect} \text{ — } \text{stat sig})$$

5 Area Under the Curve and Effect Sizes

We then analyzed the Area Under the Curve (AUC) which is also referred as the probability of superiority, which counts the fraction of times that the treatment is greater than the control over samples.

6 P-Hacking

We took a look at two studies on Musical Contrast and subjective age as well as musical contrast and chronological rejuvenation. We discussed the studies weaknesses which included poorly defined variables, lack of well defined controls, and strange adjustments to the data to justify normality. We highlighted the studies' P-Hacking issues. P-Hacking is well explained with an interactive graphic in https://fivethirtyeight.com/features/science-isnt-broken/#part1 and suggests some methods for better scientific research practices in the future. Broadly speaking, p-hacking is the process of looking at as many sources of data as possible and searching for any statistically significant result and publishing the results accordingly. Within the FiveThirtyEight applet, we can see that changing the factors can lead to broad generalizations such as Democrats or Republicans having large effects on the economy, when in reality, the situation is more complicated and can't be understood by just one data set with perfectly selected data to meet publishing criteria.

We considered the XKCD comic xkcd.com/882. In this hypothetical, researchers looked at acne trends and jelly bean consumption.

7 Publication and Citation Bias

Lastly, we looked at the effects of a culture where scientists are icentivized to only pubish positive results. Specifically, even though 50% of FDA registered studies find positive results, 95% of publications report positive findings. More information is here: bit.ly/depressionspin.

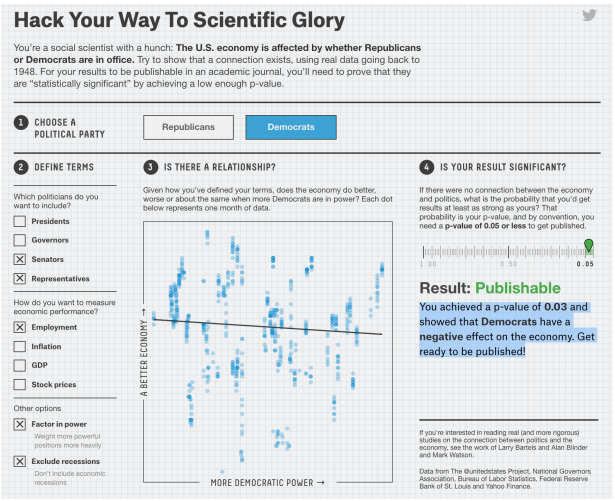


Figure 3: FiveThirtyEight p-hack

8 Conclusion

Think about the following questions:

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

Notes from sl4447

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.

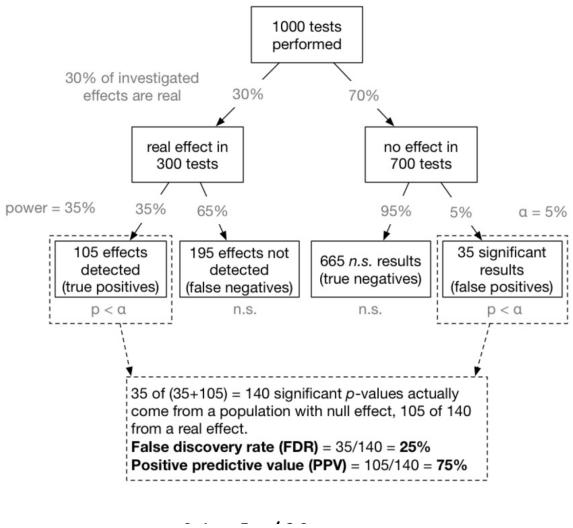


Figure 4: Answer provided during class.

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.

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

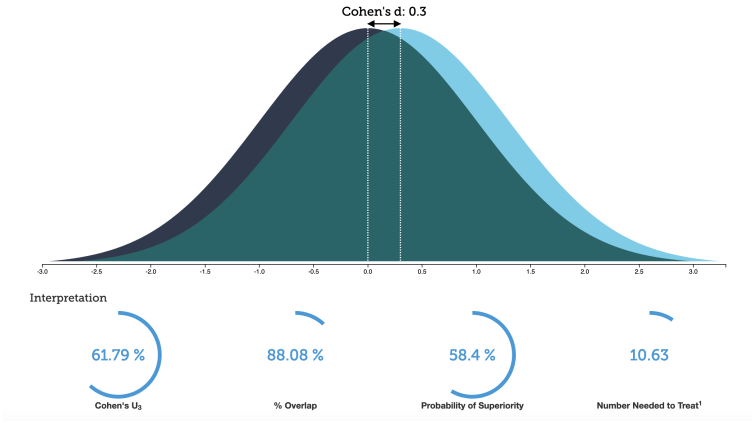


Figure 5: Cohen's d when it is 0.3.

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.

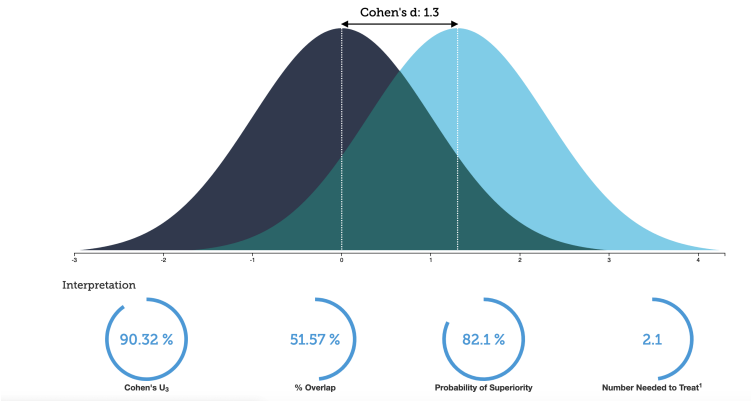


Figure 6: Cohen's d when it is 1.3.

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 7: 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.

Notes from yt2511

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
    - The answer we need is the 75%.

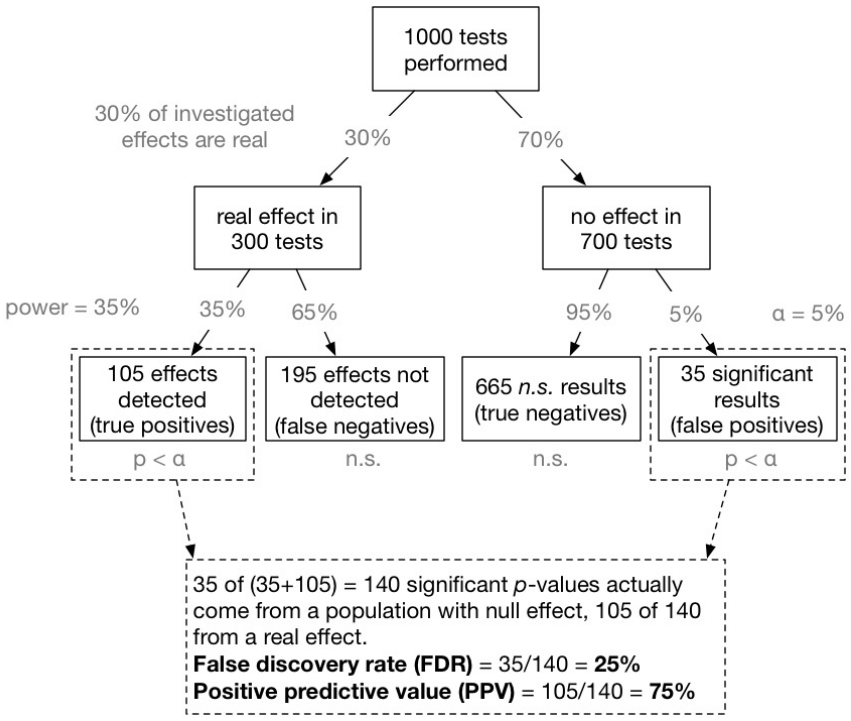


Figure 8: Tree diagram for this quiz

- 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** $= \dfrac{\mu_1 - \mu_2}{\sigma}$
    - 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