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

Sameer Jain

March 1, 2019

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.

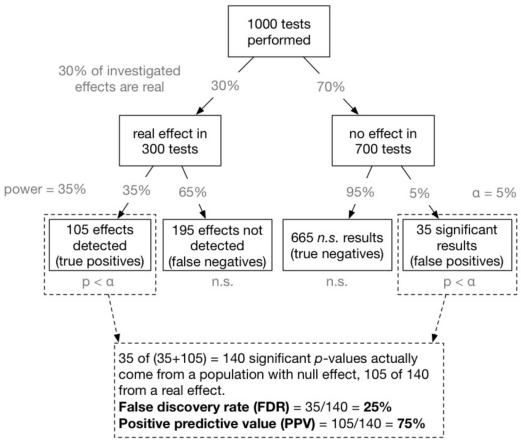
4 Misunderstandings in Statistics

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.

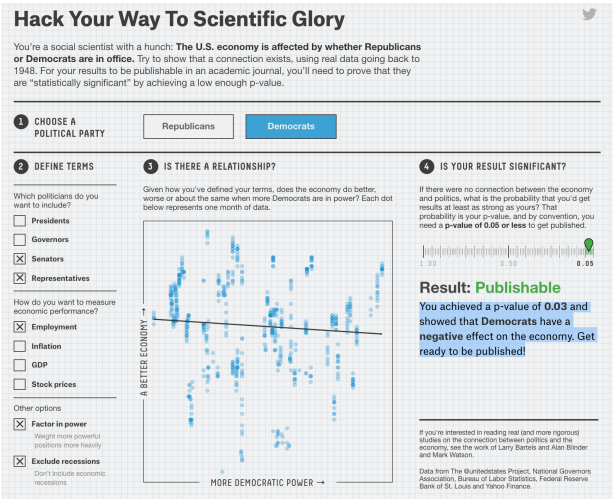


Figure 3: FiveThirtyEight p-hack

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.

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?