

Lecture 5: Reproducibility and Replication I
Modeling Social Data, Spring 2019
Columbia University

February 22, 2019

Notes from pa2492

1 Introduction

This lecture is about evaluating a research results. The main evaluation questions that one should ask are the following

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

2 Reproducibility and Replicability

We usually take the optimistic view that most researchers who publish their results are honest. However, some exceptions were reported, although few, in social science Literature available [here](#).

The two main criteria for evaluating credibility of a research is the reproducibility and replicability of the results. Reproducibility is the ability to independently verify the exact results using the same data and the same analysis. This is improving with better software engineering practices among researchers like

- Literate programming(Jupyter,Rmarkdown)
- Automated build scripts
- Containers(Docker,Code Ocean)

The well renowned journals like NIPS started to encourage researchers by attaching acknowledgement badges on their published papers as shown in Figure 1.



Figure 1: NIPS Badges

Replicability is the question of whether the result holds up with new data under the same analysis. Since it's easy to be fooled by randomness, and Noise can dominate signal in small datasets and asking too many questions of the data can lead to overfitting. Open science collaboration conducted replications of 100 experimental and correlation studies and deduced that 97 percent had significant results, but 36 percent had statistically significant results, 47 percent of original effect sizes were in the 95 percent confidence interval of the replication effect size. This leads to a Crisis where one has to believe half of what he reads.

Figure 2 depicts the original study effect size versus replication effect size.

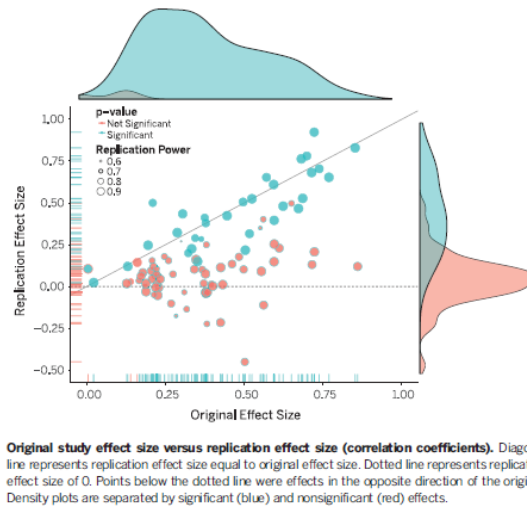


Figure 2: Original study effect size vs replication effect size.

3 Role of Statistics

Then, we talked in class through three quizzes about the role of statistics specifically the hypothesis testing, p-values, statistical significance, confidence intervals and effect sizes. We draw the below conclusion on the these quizzes:

1. Quiz 1 : One should choose Treatment A , even if it is with small numbers of participants , because following the below formula:

$$\sigma_{se} = \sigma_{sd} / \sqrt{N} \quad (1)$$

Under the same uncertainty about the mean (standard error) , the lower the N in the denominator of equation 1, the lower the standard deviation (numerator of equation 1), as shown in Figure 3.

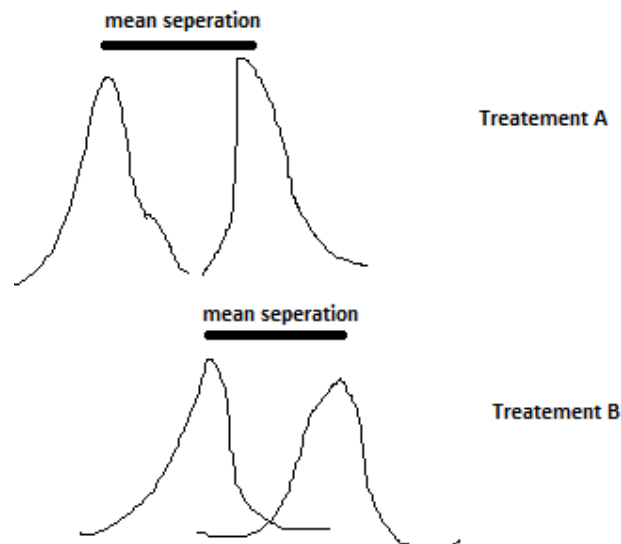


Figure 3: Effect size

2. Quiz 2 : All answers are false, the p-value given here is not a sufficient indicator of the null hypothesis, experimental hypothesis.
3. Quiz 3: Conclusion to draw here is one should not fool people with better visualization as it could be the case in Figure 4.

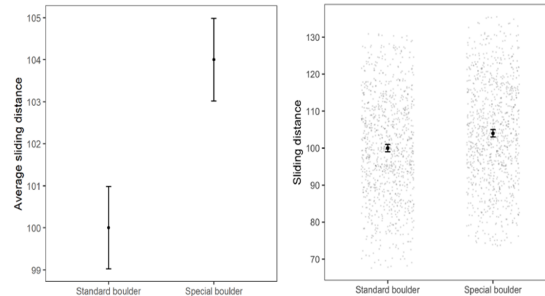


Figure 4: Left: Bad visualization Right: Correct visualization

4 Some Statistics Thoughts...

We cannot prove theories are right. Sometimes we can find contradictions to prove things false. Most often we have to settle for ruling things that are unlikely. How unlikely are my results in a boring world?

Assuming boring world (H_0 is true) Imaging running study many times in a boring world as shown in Figure 5. Look at distribution of outcomes (test statistics) from repetition in a boring world. (Figure 6) . Compare outcome (100 flips) from actual world to this distribution. For instance, if actual outcome is 0.61 , as shown in Figure 7, It is unlikely that we live in a boring world and thus one could reject the Null H_0 . However, if actual outcome is 0.52, as shown in Figure 8 one cannot reject the Null. And thus the world could be boring.

As a conclusion :

1. One need to put a threshold on the p-value
2. Need to quantify Unlikelihood.

P-value tells you probability of data you saw given that you are in a boring world. (Null H_0 is True)

$$p(D/H_0\text{True}) \quad (2)$$

You have to decide for p-value on a threshold alpha below which you reject boring world.

$$\alpha \quad (3)$$

Alpha is false positive rate and the convention value is 0.05 (1 in 20)

$$\alpha = p(\text{reject } H_0 / H_0 \text{ true}) \quad (4)$$

But what about

$$p(\text{reject } H_0 / H_0 \text{ false}) \quad (5)$$

- Assume exciting world where effect exists (H_0 is false)
- Run a study in exciting world many times, execute test and look how often to reject the Null as in Figure 9. Convention: Want power about 80

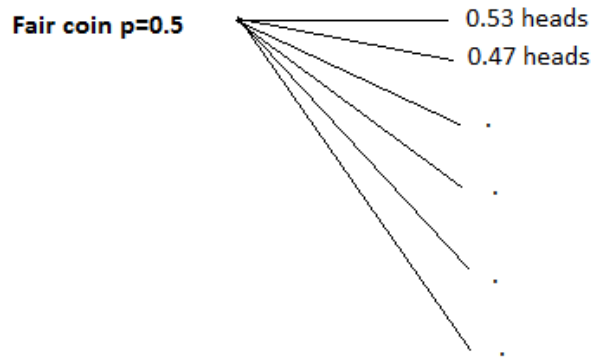


Figure 5: Boring world (H_0 is true)

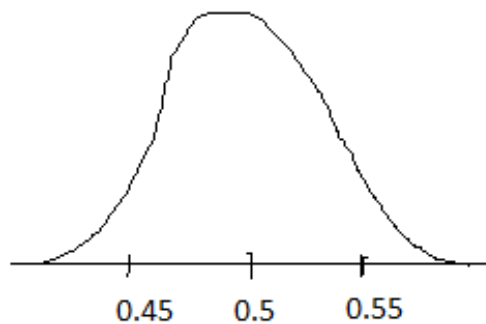


Figure 6: Distribution of outcomes from repetition in a boring world

Notes from sl3946

1 Introduction

1.1 Motivation for the lecture:

- Currently, we are seeing lots of issues with reproducibility and replication in Social Science academia, as well as in astronomy and the life sciences.
- This seems to happen because professors would get tenure from shiny results, so they occasionally had dubious approaches to achieve these results.
- Up until now we (students) have only been producing results, now we are going to look at analyzing/critiquing results to hopefully get rid of bad science and only get good scientific papers published.

1.2 Key Questions to consider in evaluation research:

- Was the research done and reported honestly/correctly?
- Is the result "real" or an artifact of the data?

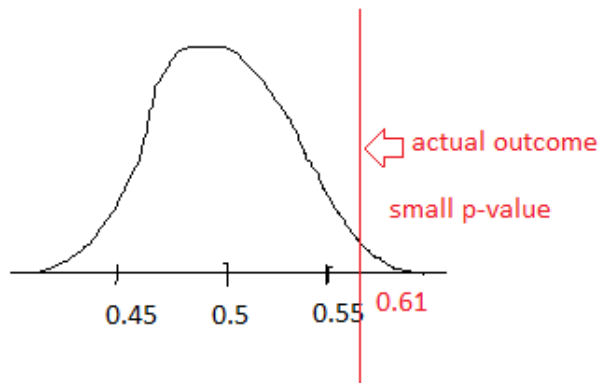


Figure 7: Reject the Null H_0

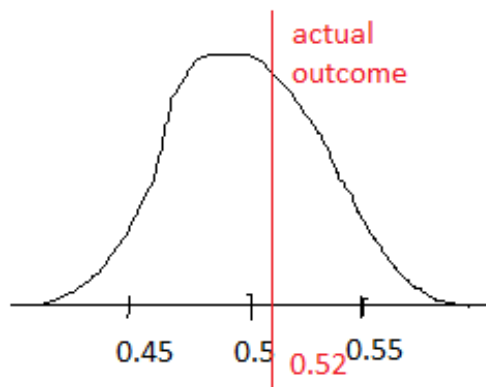


Figure 8: Cannot Reject the Null H_0

- Will it hold up over time?
- How robust is the result to small changes?
- How important/useful is the finding?

2 Honesty

Was the data reported and collected honestly: We'll take the optimistic view point that most researches are honest, and mistakes are due to stupidity, but that is not always the case.

2.1 When Contact Changes Minds: A Social Science Example of honesty failing

2.1.1 The experiment

- Ask question: Can a single conversation change people's belief on gay marriage
- Someone knocks on your door to talk to you about talking gay rights (might present yourself as straight/gay whether you are or not)

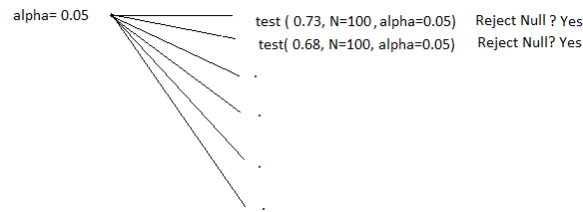


Figure 9: Study run in an exciting world with $\alpha=0.05$

- This conversation changes your view in the moment. Furthermore if the person who knocked on the door said they were gay, the results were more persistent
- The idea was that this meant that more people could knock on doors to shift public perception

2.1.2 Questioning of Results and Reproducibility

- Irregularities in LaCour (2014) showed that the results were pretty dubious
- Looked at public opinion polls that were so different from the papers results
- But they thought question was interesting so they went to analyze it with transgender
- Found that conversations decreased transphobia at a greater rate and that the results persisted for 3 months, regardless if canvasser was trans or not.

3 Reproducibility

Can you independently verify the exact results using the same data and the same analysis

3.1 Though a low bar, most research doesnt currently pass this test

- Data or code arent available/complete
- Code is difficult to run/understand
- Complex software dependencies
- Full reproducibility is not part of the current pipeline. Due to dependencies or code being complex even if its given to you its tough run

3.2 This is improving with better software engineering practices among researchers

- Literate programming (Jupyter, Rmarkdown)
- Automated build scripts (Makefiles)
- Containers (Docker, Code Ocean)
- Currently, the ACM (association for computing machinery which is main CS society) puts badges on papers if example code is provided and is functional to use.

3.3 Practical Constraints that make this not a hard rule

- [Open AI last week published](#) a better language model that can be misused so they only published a fraction of their research.
- For example: Private company wants to publish paper that used a 50k GPU etc, doesn't want to give out their own private repository and not everyone has access to this GPU
- Jake can't publish the Twitter data that he worked with due to privacy concerns as well as general data access decision by Twitter
- [A practical taxonomy of reproducibility for ML research](#) is a paper from 2018 by UW professor and Kaggle team. They analyzed the NIPS conference (where 6k submissions only 700 accepted) only 40% of accepted papers provided links to work

4 Replicability

4.1 Will the result hold up with new data but the same analysis?

- It's easy to be fooled by randomness
 - For example seeing that sending people to the door to talk about gay rights changes people's opinions.
 - You get results and it seems that there is an effect, but not persist because you're just looking at noise.
 - Small datasets of size of 50 are common in political science, which can make it especially difficult to draw conclusions
- Noise can dominate signal in small datasets
- Asking too many questions of the data can lead to overfitting
 - For example you keep y the same, but you add more and more x s to see how they impact y and eventually find a perfectly predictive model but it might not hold up
 - Tangential relationship to bias-variance trade off
 - [XKCD election replicability comic](#)

4.2 Replicability Crisis Example

- People at open science in Virginia (Brian Nosek) wanted to redo 100 different psych experiments to check to see if the results were replicable. They conducted 100 psych experiments
- Results
 - Replication effects were half the magnitude of original effects
 - 97% of original studies had statistically significant results, while in this study 36% of replicated studies had statistically significant results
 - Only 47% of original results were within the 95% confidence interval
 - The effect size versus replication shows how the independent variable effect size was rarely as strong as stated. In fact some times the effect size was negative when reported positive.
 - * Effect size indicates the difference in mean of two distributions, ie the impact of a drug on health compared to that of placebo on health
- There is a clear crisis going on that suggests that you should basically believe about half of what you read for some scientific journals
 - Jake believes it's more bad science rather than malevolence
 - World is struggling with bad science/bad statistics, so let's relearn statistics

5 Statistics Quiz

5.1 Question 1

5.1.1 Question:

- Treatment A was found to improve health over a placebo by 10 points on average (with standard error of 5 points) in a study with $N = 100$ participants.
- Treatment A was found to improve health over a placebo by 10 points on average (with standard error of 5 points) in a study with $N = 1,000$ participants.
- Which treatment would you prefer?

5.1.2 Answer:

$\sigma_{SE} = \frac{\sigma_{SD}}{\sqrt{n}}$ where σ_{SE} = the uncertainty about the mean, σ_{SD} = deviation in the population, and n = the sample size.

Clearly the difference is that treatment A has a smaller standard deviation than B therefore it has a bigger effect size. Thus **Treatment A is preferred over B.**

5.2 Question 2

5.2.1 Questions:

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant ($t = 2.7$, $df = 18$, $p = 0.01$). Which of the following are true?

- You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).
- You have found the probability of the null hypothesis being true.
- You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
- You can deduce the probability of the experimental hypothesis being true.
- You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
- You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number times, you would obtain a significant result on 99% of occasions.

5.2.2 Answers:

All false. See p-value notes in section 6

5.3 Question 3

A data visualization question that is based on the estimation of the probability of special boulder sliding further than a normal boulder based on some information from 1000 slides.

Key notes: Second visualization lowers people's confidence but still not the true result of 57%.

6 Understanding Statistics terms to better understand results

6.1 Key Ideas:

- Hypothesis testing?
- P-values?
- Statistical significance?
- Confidence intervals?
- Effect sizes?

6.2 Understanding P-Values

Assume a boring world where H_0 is true and that we have a fair coin with $p = 0.50$.

Imagine a running many trials of an experiment (where one experiment is flipping the coin 100 times). You might get $\hat{p}_0 = 0.53$ or 0.47 for example.

If you look at a distribution of the outcomes from the repetition of trials in the boring world you're likely to get a normal distribution with the mean of $\hat{p}_0 = 0.50$.

Compare the outcome from one real world experiment (ie 100 flips) to the distribution from the boring world.

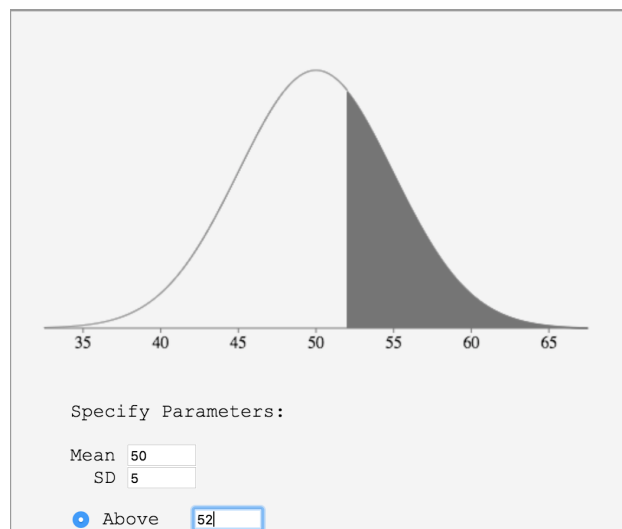


Figure 10: In this case the real world experiment had 52 head's tosses. Based on a normal distribution, the dark shaded area indicates the chance that you achieve 52 heads if $\hat{p}_0 = 0.50$. In this case it is 0.34

The p-value essential tells you the probability of seeing the data that you saw in the single test trial given that you're in a boring world (H_0 is true). Mathematically this means $p(\text{Data} \mid H_0 = \text{true})$.

Ideally before you run the experiment you decide on a p-value threshold below which reject the hypothesis of being in a boring world. This is known as α or the false positive rate. Traditionally this is known to be 0.05, ie you expect to see 1 in 20 false positives. Another way to express is alpha is $\alpha = p(\text{reject } H_0 \mid H_0 = \text{true})$.

But what about $p(\text{reject } H_0 \mid H_0 = \text{false})$ or the idea of power. It measures how likely is it that I reject a boring world when it is not a boring world? We want this to be high. If you want the value to be high even when H_0 is false for small differences from boring world, ie if think boring world is false when $\hat{p}_0 = 0.52$ then you need a relatively huge sample size.