Lecture 5 : Reproducibility, replication, etc.
Modeling Social Data, Spring 2019
Columbia University

Patrick Alrassy

February 25, 2019

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

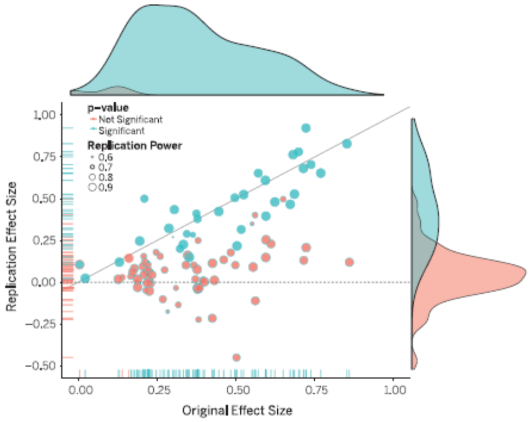
- Literate programming(Jupyter,Rmarkdown)
- Automated build scripts
- Containers(Docker,Code Ocean)

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

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 1: NIPS Badges



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

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

3 Role of Statistics

Then, we talked in class through three quizes 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 quizes:

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

$$\sigma_{se} = \sigma_{sd}/\sqrt{N} \tag{1}$$

   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.

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.

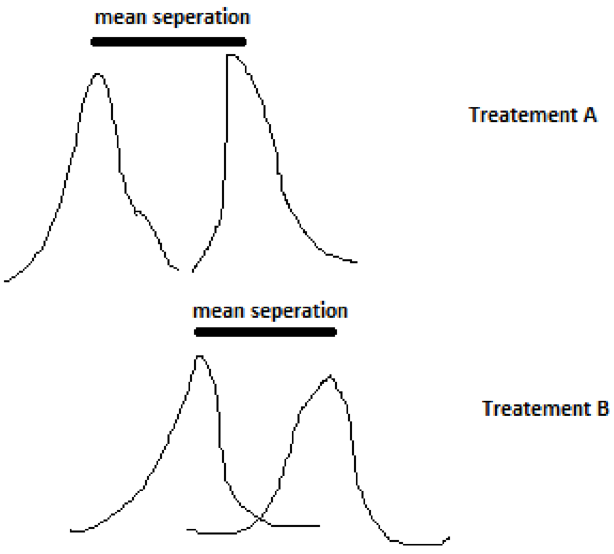


Figure 3: Effect size

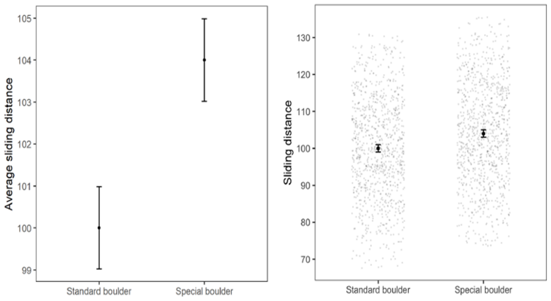


Figure 4: Left: Bad visualization Right: Correct visualization

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 (H0 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 H0. 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 Unlikeliness.

P-value tells you probability of data you saw given that you are in a boring world. (Null H0 is True)

$$p(D/H_0True) \tag{2}$$

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

$$\alpha \tag{3}$$

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

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

But what about

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

- Assume exciting world where effect exists (H0 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

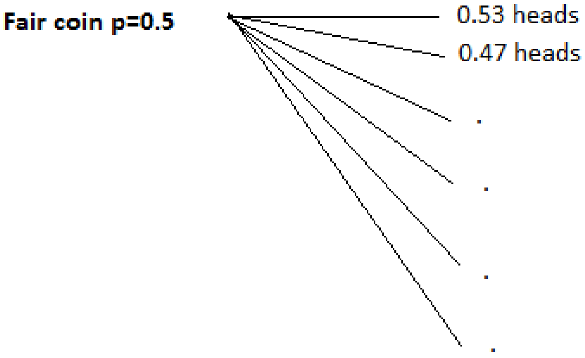


Figure 5: Boring world (H0 is true)

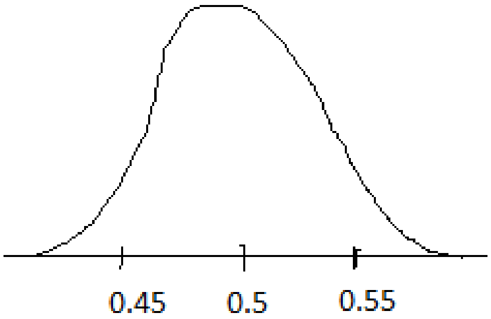


Figure 6: Distribution of outcomes from repetition in a boring world

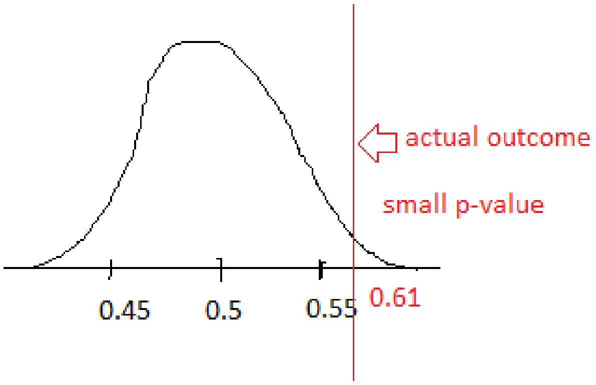


Figure 7: Reject the Null H0

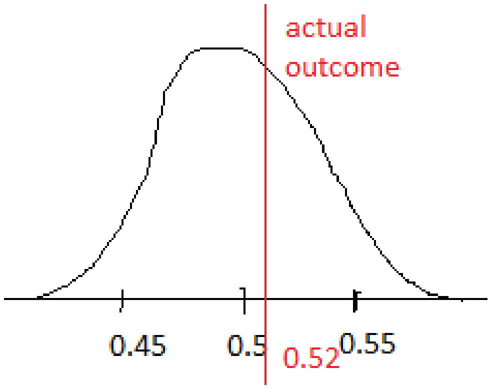


Figure 8: Cannot Reject the Null H0

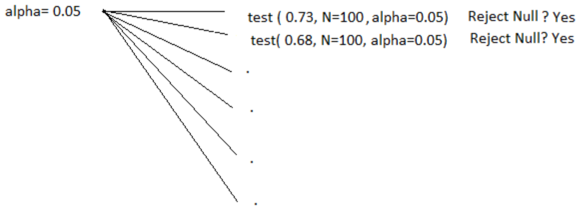

alpha= 0.05
test (0.73, N=100 , alpha=0.05)
Reject Null ? Yes
test(0.68, N=100, alpha=0.05)
Reject Null? Yes
.
.
.
.

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