Lecture 5: Reproducibility, replication, etc.
Modeling Social Data, Spring 2017
Columbia University

Sai Srujan Chinta

February 22, 2019

1 Motivation

This lecture deals with challenges related to reproduciblity and replicability of scientific studies. The later part of the lecture gives an introduction to statistics and the correct way to interpret statistics.
While evaluating research results as a consumer, it is important to think critically about the study and formulate an informed opinion about whether or not the study can be believed. Here are a few questions to ask regarding the study:

1. Was the research honest?
2. How important is the result of the research?
3. How robust is the result to small changes?
4. Will the result of the research stand the test of time?
5. Is the result real or was the data manipulated in order to obtain this result?

For the rest of the lecture, we will assume that researchers are honest, although there is clear evidence which suggests the contrary.[1]

2 Reproducibility

A study is said to be reproducible if the same results are obtained as those mentioned in the study upon conducting an individual experiment using the same data and same analysis. There are several challenges related to reproducibility:

1. Data and code are not available publicly most of the time. Speifically, getting data from private companies to recreate ML/AI experiments is difficult since all the data and infrastructure will be needed. It becomes more problematic when something like reinforcement learning needs to be verified.
2. Sometimes, even when the code is available, it is difficult to understand it, which makes it difficult to run the code.
3. A major reason for this difficulty in running the code stems from the complex software dependencies of the programs.

However, recent efforts by researchers in the form of better software engineering practices is improving the situation:

1. Literate Programming
2. Automated Build Scripts
3. Containers

[1] https://en.wikipedia.org/wiki/List_of_scientific_misconduct_incidents

3 Replicability

Replicability seeks to answer the question: Will the result hold up with new data but the same analysis?
Real-life datasets tend to be very small and noise can dominate signal in small datasets. Even in the case of large-datasets, asking enough questions leads to overfitting.
For example. consider an experiment wherein every classroom in Columbia is split into two sides and the number of IPhones owned by each side of each classroom is counted. Eventually, a conclusion can be drawn that the side of the room has an effect on IPhone ownership (even though this is clearly false). A different version of the same experiment is to take one very big classroom and conduct the same experiment on that big classroom. It is possible to find a feature which perfectly classifies this particular classroom but will not work for other classrooms.
An open science collaboration based in Virginia re-conducted 100 psychological experiments to check for replicability. The results of this study can be summarised as follows:

1. Out of the 97% studies which reported significant results, only 36% reported statistically significant results.
2. Only 47% of the original results were in the 95% confidence interval.
3. In short, this study concluded that only half of what we read can be trusted.

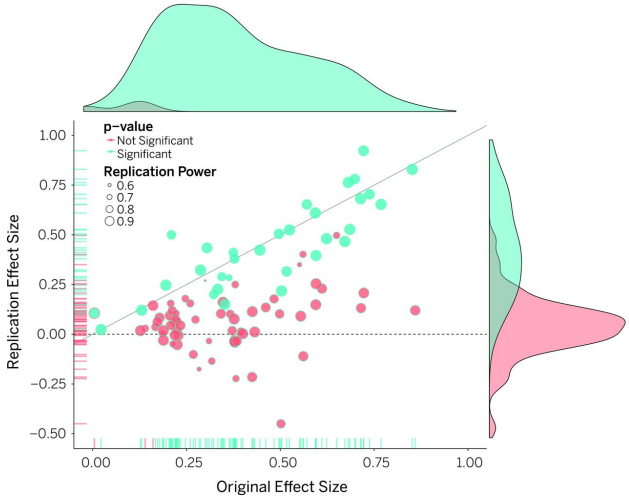

Figure 1: Original study effect size vs replication effect size

Figure 1 is the result of the above mentioned experiment. The blue points along the diagonal represent perfect replicability. However, as can be seen in the figure 1, there are far more red points than blue points and some experiments even resulted in negative correlation (replication study results directly contradict original study results).

4 Statistics Quiz

The answers and key take-aways of the quiz are as follows:

1. Treatment A is preferred over Treatment B. To understand this answer, we must first understand the definition of standard error and how it relates to the sample size:

$$\sigma_{se} = \frac{\sigma_{sd}}{\sqrt{n}} \tag{1}$$

In the equation above, σ_{sd} refers to the standard deviation in the population, n refers to the sample size and $sigma_{se}$ refers to the standard deviation of the sampling distribution. In simple words, this equation

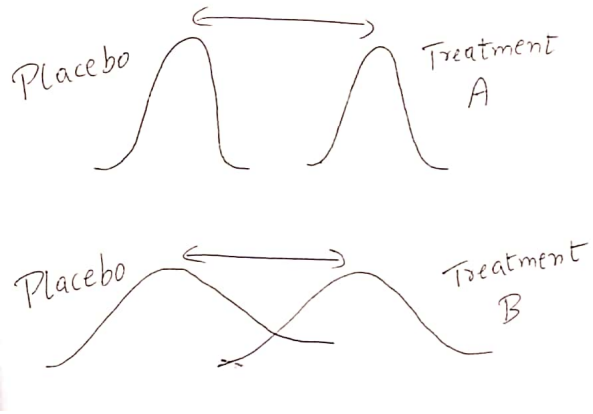

Figure 2: Treatment A vs Treatment B

tells us how uncertain we are about the mean. Therefore, though the mean separation is same in both the treatments, the population in Treatment A must have had lesser variance than that in Treatment B. If we plot out the effects of the placebo treatment and the effects of both Treatments A and B, the graph would look something like Figure 2. This graph clearly tells us that in some cases, the placebo performs better than Treatment B but this is never the case in Treatment A.

2. All the statements are false. The reasons should become clear after reviewing the definition of p-value.

3. The key take-away from this question is that data-visualisation can be used to mislead the readers and we should be wary of such attempts. The actual answer of the second part of this question is 57% but the professor says that most people are willing to bet good money that the special boulder would slide farther than the normal one with a probability of upwards of 65%.

5 Review of Statistics

- We cannot prove that theories are correct easily.
- Sometimes. we can find contradictions to prove things false.
- However, most of the time, we have to settle for ruling things out as "unlikely".
- We ask the question, "How unlikely are my results in a boring world?" (We will work with the assumption that when the null hypothesis H_0 is true, the world is boring)
- In practice, assume boring world (H_0 is true)

For example, let us consider the experiment wherein we toss a coin and try to find out if it is unbiased. For this experiment, we first take a fair coin (boring world) and toss it multiple times (say 100 times). We repeat this experiment (tossing a fair coin 100 times) multiple times and obtain the probabilities for each experiment. We can expect to see values like $p_0 = 0.52$, $p_0 = 0.48$ etc. However, when we plot these probabilities, we should obtain a normal distribution with the mean at around $p_0 = 0.5$. This normal distribution should look like the first figure in Figure 3 We now take the actual coin which we wanted to test and toss it 100 times. We now plot the resulting probability in the normal distribution obtained from the boring world. If the resulting p-value is small, like the second figure in Figure 3, we reject the null hypothesis and conclude that the coin is biased. This is because the result we observed did not conform to the result we should have observed if our world was boring (if the coin was fair). However, if the resulting p-value is large, like the third figure in Figure 3, then we conclude that our world is boring and declare that the coin is fair.
Therefore, the p-value tells us the probability of data we observed we given that we are in a boring world (null

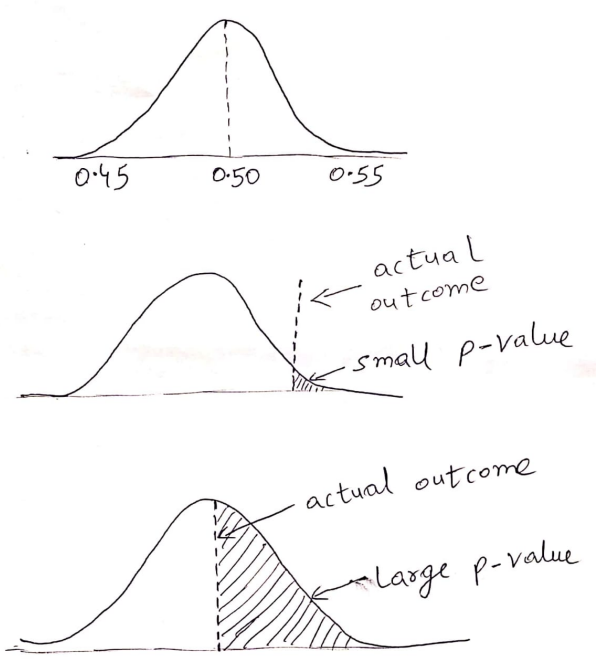

Figure 3: Coin Toss Experiment

hypothesis H_0 is true).
We also have to decide a threshold for the p-value below which we reject the boring world. This threshold is called the false positive rate α. The convention is to pick $\alpha = 0.05$.
Power is defined as the probability of rejecting the null hypothesis H_0 given that it is false. In order to estimate this probability, we conduct the opposite experiment, i.e we assume an exciting world and run the experiment multiple times. Then, we compute the number of times we reject the null hypothesis and this results in Power. In order to correctly detect even small changes, i.e to be able to reject a boring world even with $p = 0.52$, we need huge sample size.