How reliable is published research?

Shobhit Singh

> There is no cost for getting thing wrong, the cost is not getting them published
>
> Brian Nosek

May 20, 2022

With hundreds of thousands of research papers published every year, one would natural expect few of them to be refuted using further evidence and argumentation. But studies have shown that number might not be as few as you might think. This claim that most of the published research findings might be false isn't a new idea. John P. A. Ioannidis, largely bought it to light with his now famous paper aptly titled "Why Most Published Research Findings Are False" in which he modelled a framework for finding the number of false studies using some variables. Here though we are going to take a less theoretical approach based more in empirical evidence.

1 Introduction

In 1996, when priming was all the hype in the scientific community, John Bargh published a study in the *Journal of Personality and Social Psychology* linking walking speeds to reading certain words. He asked asked two groups of students to scramble some sentences. The first group was exposed to words that were related to older people and the second group had normal words. After finishing the test, they had to walk to the office to submit the paper.
It was seen that the elderly group of people walked on average , one second slower (7.30 and 8.30 seconds) to the neutral ones. [1]

In another paper published by Daryl Bem, a well-respected psychologist, he attempted to rigorously study the possibility of seeing the future. In his study, a computer was set up to place an image behind one of the two curtains and the participants would guess which curtain the computer would place the image behind. The images were of two sorts. Neutral and erotic. The success rate after 100 trial sessions for the neutral pictures was 49.8% (as expected) for the neutral pictures but for the erotic/semi nude pics the success rate was 53%! [2]

Now you might think that 53% is not a not a significant deviation but to know how the data is interpreted in a research setting, we have to take a look at how these studies are done.

2 Hypothesis testing

Most of statistics based research what is known a Hypothesis testing. You have a hypothesis which by general consensus is agreed upon, and through your research you aim to check the validity of that hypothesis.

2.1 Terminology

→ **Null Hypothesis** : The statement which is being tested through the study.
→ **Alternate Hypothesis** : The hypothesis which is the direct contradiction of the null hypothesis and will be the conclusion of the study if the null is disproved.

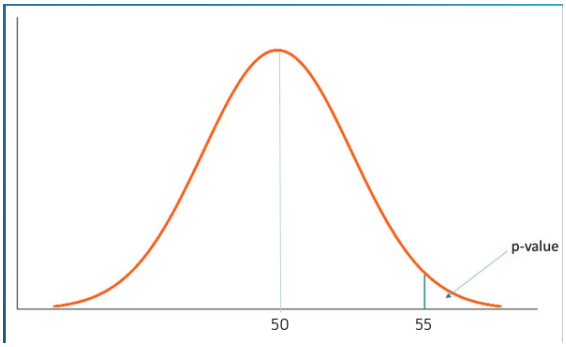
2.2 Sample study

To get a better understanding, let us try to simulate our own study.
Say you have a coin that you suspect of being biased. The only way to know is to run some tests. So you toss it 100 times and get head 55 times.

Null Hypothesis : P(Head) = 0.5
Alternate Hypothesis : P(Head) $\neq$ 0.5

Now we calculate P(Head $\geq$ 55). Since I have modeled the study on a discrete binomial distribution, it is just the sum of all probabilities from 55 to 100.
P(Head $\geq$ 55) = 0.1841
This value is called the **p-value** of the study.

$\rightarrow$ **p-value** : P value tells us how likely a result AT LEAST this extreme is if the Null hypothesis is true. i.e. If the p value is very very, we can say that it is very likely that the null hypothesis is not true and if its high, we can say that the null hypothesis is likely true.



In our case, p-value is 0.18 that means that there is a 18% chance that a fair coin gave us 55 heads in 100 tries.
This p-value is said to have reached **Statistical significance** once it goes below a particular **significance level**. When the p-value reaches below the significance level, we can discard the null hypothesis and adopt the alternate hypothesis.

But what is the magical p-value below which the coin becomes biased?
0.05

Over the years, a p-value of 0.05 has become a gold-standard of reliability across academia. So we will be using the same for our purposes. Using these standards, we can say that out coin is **Not biased** since 0.18 > 0.05.

2.3 Interpreting earlier studies

Now that we know what p-values and significance levels are, we can look at the studies mentioned earlier.

The "Slow walker" study had the p-value of 0.05
The "Feeling the future study" had the p-value of 0.01 when the images were erotic/semi-nude

And this is why they were considered true.

2.4 Significance of 0.05

There is no particular significance to 0.05. It was first introduced by R.A Fischer in his very famous book *Statistical Methods for Research Worker*. He writes

> " The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty."[3]

And that's it. It is just a convenient value. There is no reason all the researchers should stick to this hard and fast threshold. Their studies automatically do not become true as soon as they pass this barrier. Infact a p-value test was never supposed to be a definitive test but a informal quick check.[4]

2.5 What a p-value is not?

While this will be relevant later in the paper, it should be mentioned that while p-values are considered the only factor for rejecting null hypothesis, most researchers actually do not understand the true essence of it. So it should be convenient to clear up any misconceptions regarding it.

$\rightarrow$ A P value of 0.05 does not mean that there is a 95% chance that the alternate hypothesis is true. Instead, it signifies that if the null hypothesis is true, there is a 5% chance of obtaining data at least as extreme as the one recorded.[5]

3 What is a false positive

False Positive is a false correlation which accidentally has been proven true.

In the very ideal condition that a experiment has the **Statistical power** of 1 i.e. it can detect all the true correlations that are true, there still will be 5% false positives since we take a p-value of 0.05

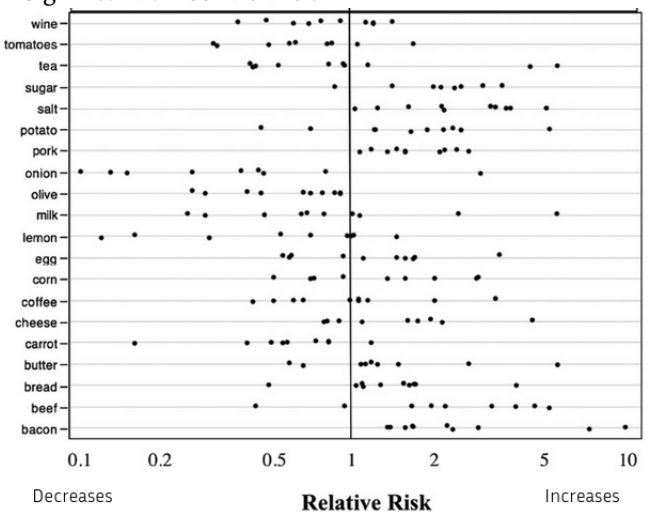
But thats not even close.

This is however like I said very ideal. For example average statistical power of studies in the field of neuroscience is probably no more than between 8% and 31%.[12]

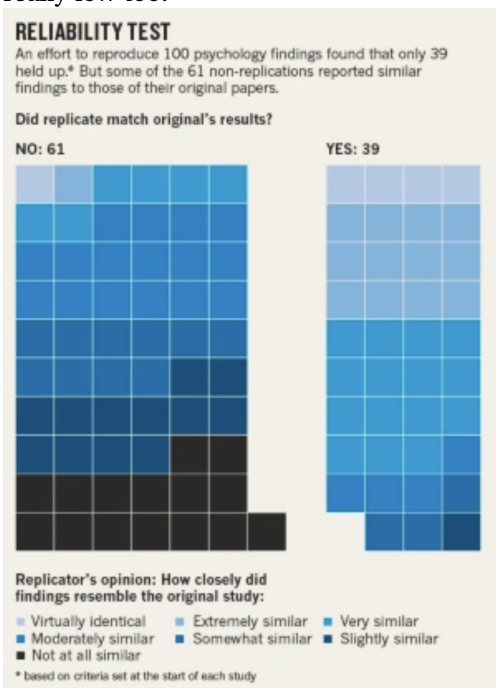
4 Empirical evidence for most studies being wrong

4.1 Case study 1

A study published by Dr John PA Ionnidis in 2013 looked at 50 most common cooking ingredients like bread, milk cheese , coffee etc and then looked at recent research linking those items to cancer. It was found that on the same items, there were several reputable studies contradicting each other. And in no insignificant amount either. [6]



really low too. [7]



→ **Cancer:**A similar study done for cancer research found out that out of 50 landmark cancer research papers, only 6 could be reproduced. That only 11%.[8]
→ **Economics:**A 2016 study in the journal Science found that one-third of 18 experimental studies from two top-tier economics journals (American Economic Review and the Quarterly Journal of Economics) failed to successfully replicate.[9]

I think it is fair to assume that this rate will be much much more in Nutritional Epidemiology since there is more corporate intervention and direct appeal to consumers involved in that field.

4.2 Reproductiblity projects

The Reproductiblity project was started in 2011 by Brian Nosek. 250 researchers came together to verify and reproduce the key findings from articles published in leading psychology journals. This project later spread in other fields.

→ **Psychology :**Out of 100 very famed studies that were selected for reproduction, they could find statistical significance in only 39. And strength was

5 How?

The data presented by reproductiblity project is pretty grim compared to our original hypothesis of 5% false studies.
Maybe the next case study can answer why or how.

5.1 Case Study 2

A study published by International Archives of Medicine aimed at finding a link between chocolate and. weight loss. It had 3 groups, 5 people in each group. One group goes on a low carb diet. Another group goes on a low carb diet but eats a bar of chocolate everyday and the third group was the control.
Both of the diet groups lost about 5 pounds over the course of the study, while the control group's mean bidy weight remained unchanged. But the group on the low-carb diet plus chocolate lost weight 10 percent faster than the non chocolate group! and this result was statistically significant. [10]

This story was picked up by all the big media outlets looking for a eye-grabbing headline and what better to put on the front page than "Chocolate and help you lose weight"

Later John Bahhanmon who conducted this research just to expose how flawed academic publishing is, said this in an interview,

> "If you measure a large number of things about a small number of people, you are almost guaranteed to get a "statistically significant" result. Our study included 18 different measurements—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.—from 15 people. That study design is a recipe for false positives. With our 18 measurements, we had a 60% chance of getting some "significant" result with $p < 0.05$. (The measurements weren't independent, so it could be even higher.) The game was stacked in our favor."

6 p-hacking

This idea of getting a very low p-value by manipulating the sample space and data is called p-hacking. A small sample space and too many enough data qualifiers are almost gonna give you some significant result.
The practice of not releasing all the recorded data quantifiers or dropping some of them midway are all too common. In fact a recent survey revealed that more than half of Dutch scientists had been involved in misconduct (hiding flaws in research) and one in 12 had been involved in fabrication or falsification of research results.[11]

6.1 Why p-hacking

→ Overrealiablity on p-values : The p-value in current academic publishing environment are considered the gold-standard for reliability. You will have a very hard time trying to get a research published which doesn't have a p-value less than 0.05. And any study with p-value less than 0.05 will make the credibility of your research skyrocket. This however is pretty irrational and not how things works in reality.

→ Journals also don't publish negative articles. Usually only 5% to 10% of the published findings are negative.

→ Researcher bias : If you are a medical researcher getting grant from a company you will probably have a bias in their favour and run multiple tests to get the desired data without disclosing.

→ Media loves a new pop science research. And researchers try to give them this. The research who conducted the "If you can see the future" study got a lot of media attention for something that can be easily disproved but saying that you can't see in the future doesn't make money.

→ Journals don't publish reproduction studies. So you can't easily disprove something.

7 References

1. Bargh, J. A., Chen, M., Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology, 71(2), 230–244. doi:10.1037/0022-3514.71.2.230
2. https://www.wired.com/2010/11/feeling-the-future-is-precognition-possible/
3. (http://www.jerrydallal.com/LHSP/p05.htm)
4. Nuzzo, R. Scientific method: Statistical errors. Nature 506, 150–152 (2014). https://doi.org/10.1038/506150a
5. Baker, M. Statisticians issue warning over misuse of P values. Nature 531, 151 (2016). https://doi.org/10.1038/nature.2016.19503
6. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. Am J Clin Nutr. 2013 Jan;97(1):127-34. doi: 10.3945/ajcn.112.047142. Epub 2012 Nov 28. PMID: 23193004.
7. Baker, M. First results from psychology's largest reproducibility test. Nature (2015). https://doi.org/10.1038/nature.2015.17433
8. Begley, C., Ellis, L. Raise standards for preclinical cancer research. Nature 483, 531–533 (2012). https://doi.org/10.1038/483531a
9. https://science.sciencemag.org/content/351/6280/1433
10. https://gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800
11. https://www.sciencemag.org/news/2021/07/landmark-research-integrity-survey-finds-questionable-practices-are-surprisingly-common
12. Button, K., Ioannidis, J., Mokrysz, C. et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14, 365–376 (2013). https://doi.org/10.1038/nrn3475