I will first identify the arguments and then discuss each one of them by providing evidence to support and disprove them, while also discussing where this evidence lacks.

**From:** Power failure: why small sample size undermined the reliability of neuroscience

Problem

* Low statistical power has a reduced chance of detecting a true effect.
* Does low power also reduce the reliability of a statistically significant finding?

Introduction

“It has been claimed and demonstrated that many (and possibly most) of the conclusions drawn from biomedical research are probably false1.”

“A simulation of genetic association studies showed that a typical dataset would generate at least one false positive result almost 97% of the time6, and two efforts to replicate promising findings in biomedicine reveal replication rates of 25% or less7,8 “

Arguments

Three main problems contribute to producing unreliable findings in studies with low power, even when all other research practices are ideal:

1. Low probability of finding true effects
2. Low Positive Predictive Value (PPV) when an effect is claimed
3. An exaggerated estimate of the magnitude of the effect when a true effect is discovered.

**Addressing point 1:**

Low power, by definition, means that the chance of discovering effects that are genuinely true is low. That is, low-powered studies produce more false negatives than high-powered studies.

When studies in a given field are designed with a power of 20%, it means that if there are 100 genuine non-null effects to be discovered in that field, these studies are expected to discover only 20 of them11.

**Addressing point 2:** *(this is what the other author argues against)*

The lower the power of a study, the lower the probability that an observed effect that passes the required threshold of claiming its discovery (that is, reaching nominal statistical significance, such as *p* < 0.05) actually reflects a true effect1,12. This probability is called the PPV of a claimed discovery.

The formula linking the PPV to power is:

PPV = ([1 – β] × R) ⁄ ([1− β] × R + α)

where (1−β) is the power, β is the type II error, α is the type I error and R is the pre-study odds (that is, the odds that a probed effect is indeed non-null among the effects being probed). The formula shows that, for studies with a given pre-study odds R, the lower the power and the higher the type I error, the lower the PPV. And for studies with a given pre-study odds R and a given type I error (for example, the traditional *p* = 0.05 threshold), the lower the power, the lower the PPV (see article for an example).

**Addressing point 3:**

Even when an underpowered study discovers a true effect, it is likely that the estimate of the magnitude of that effect provided by that study will be exaggerated. This effect inflation is often referred to as the ‘winner’s curse’13 and is likely to occur whenever claims of discov- ery are based on thresholds of statistical significance (for example, *p* < 0.05) or other selection filters (for example, a Bayes factor better than a given value or a false-discov- ery rate below a given value).

Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large. If, for example, the true effect is medium-sized, only those small studies that, by chance, overestimate the magnitude of the effect will pass the threshold for discovery.

Explanation:

To illustrate the winner’s curse, suppose that an association truly exists with an effect size that is equivalent to an odds ratio of 1.20, and we are trying to discover it by performing a small (that is, under- powered) study. Suppose also that our study only has the power to detect an odds ratio of 1.20 on average 20% of the time. The results of any study are subject to sampling variation and random error in the measurements of the variables and outcomes of interest. Therefore, on aver- age, our small study will find an odds ratio of 1.20 but, because of random errors, our study may in fact find an odds ratio smaller than 1.20 (for example, 1.00) or an odds ratio larger than 1.20 (for example, 1.60). Odds ratios of 1.00 or 1.20 will not reach statistical significance because of the small sample size. We can only claim the association as nominally significant in the third case, where random error creates an odds ration of 1.60. The winner’s curse means, therefore, that the ‘lucky’ scientist who makes the discovery in a small study is cursed by finding an inflated effect.

The winner’s curse can also affect the design and con- clusions of replication studies. If the original estimate of the effect is inflated (for example, an odds ratio of 1.60), then replication studies will tend to show smaller effect sizes (for example, 1.20), as findings converge on the true effect. By performing more replication studies, we should eventually arrive at the more accurate odds ratio of 1.20, but this may take time or may never happen if we only perform small studies. A common misconception is that a replication study will have sufficient power to replicate an initial finding if the sample size is similar to that in the original study14. However, a study that tries to replicate a significant effect that only barely achieved nominal statistical significance (that is, *p* ~ 0.05) and that uses the same sample size as the original study, will only achieve ~50% power, even if the original study accurately estimated the true effect size. This is illustrated in FIG. 1. Many published studies only barely achieve nominal sta- tistical significance15. This means that if researchers in a particular field determine their sample sizes by historical precedent rather than through formal power calculation, this will place an upper limit on average power within that field. As the true effect size is likely to be smaller than that indicated by the initial study — for example, because of the winner’s curse — the actual power is likely to be much lower. Furthermore, even if power calcula- tion is used to estimate the sample size that is necessary in a replication study, these calculations will be overly optimistic if they are based on estimates of the true effect size that are inflated owing to the winner’s curse phenomenon. This will further hamper the replication process.

Lower power in the presence of other biases…

Discussion

From arguments section: All other research practices are rarely ideal, how do they add to the problem of low power?

From point 3: What are the exact mechanisms behind the Winner’s curse?

What do they mean when saying that “those small studies that, **by chance**, overestimate the magnitude of the effect…”? (see point 3 explanation)

General note: are large sample sizes always better?