Type M error can explain Weisburd’s Paradox*

Andrew Gelman

8 May 2016

Abstract

Simple calculations seem to show that larger studies should have higher statistical power, but empirical meta-analyses of published work in criminology have found zero or weak correlations between sample size and estimated statistical power. This is “Weisburd’s paradox” and has been attributed by Weisburd, Petrosino, and Mason (1993) to a difficulty in maintaining quality control as studies get larger, and attributed by Nelson, Wooditch, and Dario (2014) to a negative correlation between sample sizes and the underlying sizes of the effects being measured. We argue against the necessity of both these explanations, instead suggesting that the apparent Weisburd paradox is an artifact of systematic overestimation inherent in post-hoc power calculations, a bias that is large with small n . Furthermore, we recommend abandoning the use of statistical power as a measure of the strength of a study, because implicit in the definition of power is the bad idea of statistical significance as a research goal.

1. Introduction

Criminologists are heavily focused on policy evaluations, with the goal of increasing the use of evidence-based criminal justice policies and practices. Experiments and quasi-experiments play a big role in this research and policy agenda, and criminological research findings have considerable potential to influence (for better or worse) citizens’ lives, given the immense reach of the criminal justice system. About a third of all Americans are arrested by the age of 23 (Brame et al., 2012), and almost a fifth are convicted by that age (Bushway et al., 2014). Around 1% of American adults are incarcerated in jails or prisons (Pew, 2008).

The field of criminology is just now beginning to engage with the problems documented in other fields surrounding null hypothesis testing, forking paths, false positives, and publication bias; hence the special issue to which this article is being submitted. In the present paper we take a pair of meta-analyses by Weisburd, Petrosino, and Mason (1993) and Nelson, Wooditch, and Dario (2014) as a springboard to discuss the ways that our understanding of quantitative evidence can be distorted by biases in estimates of effect size and power.

2. Type M and Type S error

Statistical tests are typically understood based on Type 1 error (the probability of falsely rejecting a null hypothesis of zero effect, if it is in fact true) and Type 2 error (the probability of *not* rejecting a null hypothesis that is in fact false). But this paradigm does not match up well with much of social science, or science more generally.

Our first problem with Type 1 and Type 2 errors is that in many problems we do not think the null hypothesis can be true. For example, a change in sentencing guidelines will certainly produce *some* changes in behavior; the question is how these changes vary across persons and situations. The

*For the *Journal of Quantitative Criminology*. We thank Justin Pickett and Gary Sweeten for suggesting this topic and for helpful suggestions, and the National Science Foundation, Institute of Education Sciences, and Office of Naval Research for partial support of this work.

second issue is that, in practice, when hypothesis test is rejected, researchers and practitioners take this rejection as evidence about the magnitude and sign of the underlying effect. So, in evaluating a statistical test, we should be interested in the properties of the associated effect-size estimate, conditional on it being statistically significantly different from zero.

With these concerns in mind, Gelman and Tuerlinckx (2000) introduced the concept of Type S (“sign”) and Type M (“magnitude”) errors, both of which can occur when a researcher makes a *claim with confidence* (typically, a p-value of less than 0.05 or a confidence interval that excludes zero, but more generally any statement that is taken as strong evidence of a positive effect. A Type S error occurs when the sign of the estimated effect is of the opposite direction as the true effect. A Type M error occurs when the magnitude of the estimated effect is much different from the true effect. A statistical procedure can be characterized by its *Type S error rate*—the probability of an estimate being of the opposite sign of the true effect, conditional on the estimate being statistically significant—and its *expected exaggeration factor*—the expected ratio of the magnitude of the true effect divided by the magnitude of the underlying effect.

As discussed by Gelman and Carlin (2014), when a statistical procedure is noisy, the Type S error rate and the exaggeration factor can be large. They give an example where the Type S error rate is 24% (that is, any statistically significant estimate has a one-quarter chance of being in the wrong direction) and the expected exaggeration factor is 9.7. We discuss that example further in the section 4.

In this paper we are particularly concerned with Type M errors, or exaggeration factors, which can be understood in light of the “statistical significance filter.” Consider any statistical estimate. For it to be statistically significant, it has to be at least 2 standard errors from zero. Thus, the larger the standard error, the higher the estimate *must* be, if it is to be published and taken as serious evidence. No matter how large or small the underlying effect, the minimum statistically significant effect size *estimate* has this lower bound. This selection bias induces Type M error. And, as we shall see, it can explain a well-known paradox that has been reported in quantitative research.

3. Weisburd’s Paradox

Weisburd, Petrosino, and Mason (1993) document a surprising pattern in criminal justice experiments:

Using sample estimates as a guide, the very largest investigations are no more powerful than the very smallest. Indeed, we find little relationship in practice between sample size and statistical power.

As they note, this finding is counterintuitive because for any fixed design the standard error will decline with increasing sample size. If effect size is constant and standard error declines, statistical power will increase.

One can imagine three possibilities for a lack of relation between sample size and power:

1. Larger experiments are employed when studying smaller effects. For any given research topic, a larger sample would lead to more statistical power, but when studying different projects, this mathematical relation is confounded by selection.
2. Studies tend to decline in quality when they are made larger. As sample size increases, the bias and variance of individual measurements increases, enough to overwhelm the gain in precision achieved by larger samples.

3. Statistical power from small samples is not as large as researchers have believed. Crude calculations based on sample estimates will systematically overestimate the power attainable from small studies. The apparent lack of relation between sample size and power is actually just an artifact of a biased power estimate.

All three of these can be happening at once, but we shall consider them one at a time.

Explanation 1 is possible; indeed, very large effects such as a hypothetical study of the effectiveness of the guillotine can be ascertained with $n = 1$, whereas a savvy researcher will know that a large experiment is needed when studying a small effect. Nelson, Wooditch, and Dario (2014) perform a meta-analysis of experiments in criminology and report a strong negative correlation between sample size and effect size, but for reasons discussed below we suspect this pattern is merely an artifact of how these effect sizes have been reported.

Overall we doubt that explanation 1 could be behind Weisburd’s paradox. In general, we would expect the resources of large- n experiments to go toward studying effect that are believed to be large and important, with smaller sample sizes being allocated to more speculative topics.

Explanation 2 will depend on context. Weisburd, Petrosino, and Mason (1993) offer examples where increasing the size of a study makes it more difficult to exercise the control required to achieve low bias and variance, and Nelson, Wooditch, and Dario (2014) consider the possibility that large studies are more likely to involve heterogeneous populations and thus less stable effects. On the other hand, the opposite pattern can also hold, in which large experiments are more careful and controlled, with smaller studies being more amateurish and more subject to error. Thus, although we accept that some of the patterns identified by Weisburd et al. really happen in criminal justice experiments, we are skeptical of this being a general explanation for the apparent lack of relationship between sample size and power.

This brings us to explanation 3. Published estimates are typically at least 2 standard errors from zero, as that is what it takes to be statistically significant at the 5% level. And it is often not so difficult for researchers to find a statistically significant comparison in any dataset, whether via “p-hacking” using “multiple researcher degrees of freedom” (Simmons, Nelson, and Simonsohn, 2011) or through the more innocuous “garden of forking paths” (Gelman and Loken, 2015). In the criminal justice context, Nelson, Wooditch, and Dario (2014) consider the possibility of publication bias. The implicit rule that statistical significance is required for publication, coupled with the ease of finding statistically significant comparisons, ensures that published estimates are systematically biased upward—and, as noted above, this statistical significance filter is a larger problem for smaller, noisier studies and can lead to large Type M errors. And noisy overestimate of effect sizes lead directly to noisy overestimates of power (see, for example, Senn, 2002).

Now look at what Weisburd, Petrosino, and Mason (1993) did:

We developed an “average effect size” (AES) measure by taking the mean of all the effect coefficients at the follow-up period closest to one year... we developed an additional measure—maximum effect size (MES)—that provides an upper range of effect for the experiments.

Both these definitions are subject to the statistical significance filter, and using these as effect size measures will overestimate power, with this bias being larger for small- n studies.

Thus, without making any judgments about the first two explanations listed above, it seems that explanation 3 can explain Weisburd’s paradox all by itself. It is possible that explanations 1 and 2 also have validity, but the evidence provided by Weisburd et al. is not really enough to support such a claim, as their power calculations are based on seriously corrupted estimates of effect size.

One way this confusion can persist is that randomized experiments are celebrated for providing unbiased inference. But, even for randomized studies with unbiased measurements, the move to *statistically significant published estimates* leads to a bias that can be huge for small, noisy studies. So, clean design of individual studies does not guarantee freedom from bias for inferences such as those of Weisburd, Petrosino, and Mason (1993) which are based not on raw data but on published results.

In a more recent study, Nelson, Wooditch, and Dario (2014) replicated the meta-analysis of Weisburd et al. and again found no clear pattern between power and sample size. Nelson et al. report that “effect sizes decline as the sample size of the experiment increases,” but this pattern can again be explained by the statistical significance filter. Comparing effects in published vs. unpublished studies is not the same as significant vs. not significant, because the latter are often never even written up (Franco, Malhotra, and Simonovits, 2014). Nelson et al. report that “Publication bias, disclosure of fidelity issues, and year of publication do not account for a significant amount of variance in effect size,” but this finding does not contradict our story in which all these effect size estimates are polluted by selection bias.

4. The problem with statistical power

The conventional view is that you should avoid low-power studies because they are a waste of time. For example, in the paper discussed above, Weisburd et al. write:

Statistical power provides the most direct measure of whether a study has been designed to allow a fair test of its research hypothesis. When a study is underpowered it is unlikely to yield a statistically significant result even when a relatively large program or intervention effect is found.

This is correct but can mislead in that it too simply presents statistical significance as a goal.

To see the problem, suppose that a study is low power but can be performed for free, or for a cost that it is very low compared to the potential benefits that would arise from a research success. Then a researcher might rationally conclude that a lower-power study is still worth doing, that it is a gamble worth undertaking.

The traditional threshold for power is 80%; funding agencies are reluctant to approve studies that are not deemed to have at least an 80% chance of obtaining a statistically significant result. But under a simple cost-benefit calculation, there would be cases where 50% power, or even 10% power, would suffice, for simple studies such as psychology experiments where human and dollar costs are low. Why not roll the dice, if a successful finding could potentially bring large benefits (to society as well as to the researcher’s career)?

Why not, indeed?

The “why not” is that, in a low-power study, the seeming “win” of statistical significance can actually be a trap. Economists speak of a “winner’s curse” in which the highest bidder in an auction will, on average, be overpaying. Research studies—even randomized experiments—suffer from a similar winner’s curse, that by focusing on comparisons that are statistically significant, we (the scholarly community as well as individual researchers) get a systematically biased and over-optimistic picture of the world.

To put it in technical terms, statistically significant results are subject to Type S and Type M errors. For example, Gelman and Carlin (2014) discuss a study where the true effect could not realistically be more than 2 percentage points and it is estimated with a standard error

The winner's curse of the low-power study

Figure 1: When the standard error is small compared to the effect size, statistical power is low. In this diagram, the bell-shaped curve represents the distribution of possible estimates, and the red shaded zones correspond to estimates that are statistically significant (at least two standard errors away from zero). In this example, statistical significance is unlikely to be achieved, but in the rare cases where it does happen, it is highly misleading: there is a large chance the estimate has the wrong sign (a Type S error) and, in any case, the magnitude of the effect size will be vastly overstated (a Type M error) if it happens to be statistically significant. Thus, what would naively appear to be a “win” or a lucky draw—a statistically significant result from a low-power study—is, in the larger sense, a loss to science and to policy evaluation.

of 8.1 percentage points. Examining the statistical properties of the estimate using the normal distribution: conditional on it being statistically significant (that is, at least two standard errors from zero), it has at least a 24% probability of being in the wrong direction and is, by necessity, over 8 times larger than the true effect. Figure 1 demonstrates the winner’s curse in the context of this example.

A study with these characteristics is dead on arrival: it has a power of 6%, but it would be misleading to say it has even a 6% chance of success. In fact, arguably the 6% of the time that the study appears to succeed, represent the real failures, in that they correspond to ridiculous overestimates of treatment effects that are likely to be in the wrong direction as well. In such an experiment, to win is to lose.

Thus, the problem of a low-power study is not so much that it has a small chance of succeeding, but rather that an apparent success merely masks a larger failure. This point was made by Button et al. (2013) but does not seem to have yet been fully disseminated among research practitioners.

5. Discussion

In a meta-analysis of criminal justice experiments of various sizes and on various topics, Weisburd, Petrosino, and Mason (1993) found no empirical relation between sample size and estimated statistical power. They identified this as a paradox because, all else equal, power should increase with

sample size. We resolve this paradox by refusing to accept the simple post-hoc power estimates which in turn are based on systematic overestimates of effect sizes. Especially for small, noisy studies, statistically significant effects can have high Type M (“magnitude”) errors, which can obscure an underlying relation between sample size and true statistical power.

Beyond this, we recommend following Nelson, Wooditch, and Dario (2014) and thinking of studies in terms of effect size and variation rather than statistical power. Power is the probability of obtaining a statistically significant difference in a fixed design, and we resist the framing of statistical designs using statistical significance as a goal. Noisy studies are not just a problem because they have “low power,” thus unlikely to produce a successful result. No, the problem is worse: with sufficient noise, what seems like a success is actually a failure, in that statistically significant results are likely to be in the wrong direction and in any case will grossly overestimate effect sizes.

Hence we give the statistical advice to think less about power, both in evaluating patterns in past studies and in designing future experiments. Instead we recommend a greater acceptance of uncertainty, in particular, a willingness to publish non-statistically-significant results, and an understanding that results with p less than 0.05 can still fail to replicate in the general population.

That said, the mechanisms discussed by Weisburd et al. and Nelson et al.—explaining a hypothesized decline of effect sizes as studies get larger—are interesting and worthy of study in their own right. However, given the biases discussed above—selection in what studies are published and in what results are presented within any published study—will make it extremely difficult to study such patterns using the published record. Instead it could make more sense to study the relation between study size and effect size in a domain in which raw data were available.

References

- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). *Nature Reviews Neuroscience* **14**, 365–376.
- Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502–1505.
- Gelman, A.. (2015) The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* **41**, 632–643.
- Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.
- Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465.
- Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390.
- Nelson, M. S., Wooditch, A., and Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd’s paradox. *Journal of Experimental Criminology* **11**, 141–163.
- Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *British Medical Journal* **325**, 1304.
- Weisburd, D., Petrosino, A., and Mason, G. (1993). Designs sensitivity in criminal justice experiments. *Crime and Justice* **17**, 337–379.