Inflated Effect Sizes and Underpowered Tests: How the Severity Measure of Evidence is Affected by the Winner's Curse

Guillaume Rochefort-Maranda

December 21, 2017

Contents

1 Introduction

In philosophy of statistics, Deborah Mayo and Aris Spanos have championed the following epistemic principle, which applies to frequentist tests:

> **Severity Principle (full)**. Data x_0 (produced by process G) provides good evidence for hypothesis H (just) to the extent that test T severely passes H with x_0. (Mayo and Spanos 2011, p.162).

They have also devised a severity score that is meant to measure the strength of the evidence by quantifying the degree of severity with which H passes the test T (Mayo and Spanos 2006, 2011; Spanos 2013). That score is a real number defined on the interval [0,1].

My aim in this paper is to show how the problem of inflated effect sizes corrupts the severity measure of evidence. This has never been done. Since the severity score is the predominant measure of evidence for frequentist tests in the philosophical literature, it is important to underscore its flaws.

The problem is that when a significant result is obtained by using an underpowered test, the severity score becomes particularly high for large discrepancies from the null-hypothesis. This means that such discrepancies are very well supported by the evidence according to that measure.

However, it is now well documented that significant tests with low power display inflated effect sizes (this is also known as the winner's curse). They systematically show departures from the null hypothesis H0 that are much greater than they really are:"theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated"(Ioannidis 2008, p.640) This is problematic in research contexts where the differences be-

tween H0 and H1 are particularly small and where the sample sizes are also small. See(Button et al. 2013; Ioannidis 2008; Gelman and Carlin 2014) for examples).

From an epistemological point of view this means that a significant result produced by an underpowered test does not provide evidence for large discrepancies from H0. Therefore, the severity score is an inadequate measure of evidence.

Given that we are now aware of the phenomenon of inflated effect sizes, it would be irresponsible to rely on the severity score to measure the strength of the evidence against the null. Instead, one must take appropriate measures to try and avoid using underpowered tests by setting a threshold for the sample size or by replicating the results of the experiment.

Unfortunately, this solution is incompatible with with Spanos and Mayo's claims to the effect that there is a common fallacies "wherein an a level rejection is taken as more evidence against the null, the higher the power of the test" (Mayo and Spanos 2006, p.344).

This paper contains two main sections. In the first section, I explain the problem of inflated effect sizes generated by underpowered tests with more details. I also provide an example by using a A Student's t-Test. In the final section, I explain why the severity score is an inadequate measure of evidence.

2 The argument and the methodology

The main argument that I put forward in this paper is very simple.

- An observed test statistic will display a misleading departure (large effect size) from both H0 and H1 when an underpowered test is significant.
- The severity score justifies larger discrepancies from the null when the observed effect size is large.

- Therefore, the severity score is a measure that will be systematically wrong when evaluating the result of a underpowered test.

The premises of this argument are now established facts. The first premise more particularly is a well-known mathematical phenomenon:

> 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 winners curse (Button et al. 2013, p.366).

and it affects real scientific practice (it is not merely a theoretical problem):

> Our results indicate that the average statistical power of studies in the field of neuroscience is probably no more than between 8% and 31%, on the basis of evidence from diverse subfields within neuro-science. If the low average power we observed across these studies is typical of the neuroscience literature as a whole, this has profound implications for the field. A major implication is that the likelihood that any nominally significant finding actually reflects a true effect is small. (Button et al. 2013, p.371).

Now, the purpose of this paper is not to prove the first premise or to show that it is a real problem. As one can see, this as already been done. I will however illustrate the problem with a simulation study. In other words, the simulation is not meant to prove what has already been proven but merely to given an example.

The originality of this paper is to put the first and second premise together in order to dismiss the severity score as an adequate measure of evidence. The simulation study is also helpful because we can experiment with a full knowledge

of the real discrepancy between H0 and H1. We would not be able to do that with a real case-study.

2.1 Inflated Effect Sizes Generated by Underpowered Tests

The fact is that the lower the power of a test, the more H0 and H1 are similar. Consequently, the more extreme a test statistics must be under H1 in order to trigger a significant result. This implies that a significant result provided by a low powered test will necessarily display a departure from what we expect under both H0 and H1, such that we will have the illusion that H1 is much more different from H0 than it really is.

There are two necessary conditions to observe this phenomenon: significance and low power.

> Inflation is expected when, to claim success (discovery), an association has to pass a certain threshold of statistical significance, and the study that leads to the discovery has suboptimal power to make the discovery at the requested threshold of statistical significance. Both conditions are necessary to inflate effect sizes.
>
> (Ioannidis 2008, p.641).

This problem is fairly easy to illustrate. Imagine that a statistician S has obtained two different samples of 10 independent and identically distributed observations: $(X_1, X_2, ..., X_{10})$ and $(Y_1, Y_2, ..., Y_{10})$. Their respective distributions are defined as follows:

(i) $\boldsymbol{X_i} \sim \mathcal{N}(\mu_1 = 1.01, \sigma_1^2 = 36)$

(ii) $\boldsymbol{Y_j} \sim \mathcal{N}(\mu_2 = 1, \sigma_2^2 = 36)$

where μ represents the mean of a normal distribution and σ^2 its variance.

S only knows two things about the parameters of the two normal distributions:

(1) $\mu_1 > \mu_2$ or $\mu_1 = \mu_2$

(2) $\sigma_1 = \sigma_2$

She does not know their exact value. Consequently, in order to make an inference about the difference between μ_1 and μ_2, S uses a one-tailed Student's t-Test where H1: $\mu_1 > \mu_2$ and H0: $\mu_1 = \mu_2$. The variances are estimated with the samples.

The statistic used for such a test is defined as follows:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

where

$$S_p = \sqrt{\frac{9S_1^2 + 9S_2^2}{18}},$$

$$S_1^2 = \sum_{i=1}^{10} \frac{((x_i) - \bar{X})^2}{9},$$

$$\bar{X} = \sum_{i=1}^{10} \frac{x_i}{10},$$

$$S_2^2 = \sum_{i=1}^{10} \frac{((y_i) - \bar{Y})^2}{9},$$

and

$$\bar{Y} = \sum_{i=1}^{10} \frac{y_i}{10}.$$

It is called a Student's t-Test because the statistic t follows a Student distribution (with 18 degrees of freedom in this case).

For a significance level α of 0.05, S will reject H0 (accept H1) if she finds a test statistic t_{obs} such that the probability of obtaining a result at least as distant (on

the positive axis) from 0 as t_{obs} is smaller than or equal to 0.05 under H0. If not, then she will fail to reject H0.

The probability that will determine the rejection (or non-rejection) of H0 is called "the p-value". In this particular case, α is the probability of rejecting H0 when H0 is true. It is also called "the probability of making a Type-I error". The probability of rejecting H0 when H1 is true is called "the power of the test" (π) and the probability of not rejecting H0 when H1 is true is "the probability of making a Type-II error" ($\beta = 1 - \pi$). In this case, the power of the test is very low given the small difference between the populations, the high variances and the small sample size.

In short, S expects the statistic t to be close to 0 under H0 because there should not be any difference between the two distributions. If the test statistics is much bigger than 0, then she will reject H0 and accept H1 because that would be too improbable under H0. If it is relatively close to 0, then she will not reject H0 because that is not too improbable under H0.

After S proceeds with the t-test, she finds a difference of 4.250; a test statistic $t_{obs} = 1.914$; and a p-value = 0.036 (See Appendix to reproduce the results). Therefore, S rejects H0 (p-value< 0.05). The test is significant.

In fact, the result is quite remarkable. S has observed a difference between the two means of 4.250 when the true difference is only 0.01. This is because we have a significant result with an underpowered test such that the effect size incredibly bigger than reality (450 times greater). S would thus be wrong to believe that there is such a substantial difference between H0 and H1.

2.2 The Severity Score

Now, suppose that S would like to use the severity score for $\mu_1 - \mu_2 > 0.1$ in order to quantify the strength of the evidence attached to that claim. She computes that score as follows:

$$t_s = \frac{(4.250) - (0.1)}{S_p \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$SEV(\mu_1 - \mu_2 > 0.1) = \boldsymbol{F}(\boldsymbol{t_s}) = 0.961$$

where $\boldsymbol{F}(\boldsymbol{t_s})$ is the cumulative distribution function of a Student's distribution with 18 degrees of freedom evaluated at point t_s.

In English, this means that S has computed the probability of obtaining a less extreme result under the assumption that $\mu_1 - \mu_2 = 0.1$. This is the meaning of the severity score in this context. See (Mayo and Spanos 2011, p.169) for more details on how to compute such a severity score.

If the severity score is high, then we can infer that the data provides good evidence for $\mu_1 - \mu_2 > 0.1$ (see the first quote in the introduction). This is the case here and it should not come as a surprise given that S has observed such an inflated effect size. Notice that the severity score will be higher the greater the observed size effect (just look at the numerator of the fraction that generates t_s).

In a nutshell, S has found a significant result (p-value=0.036). She thus rejects H0 and finds a high severity score for the claim $\mu_1 - \mu_2 > 0.1$ (severity score=0.961). Hence, S believes that she has good evidence for such a difference that is at least ten times larger than the true difference.

However, S would be epistemically irresponsible to trust the severity score given what is now known about the problem of effect sizes and underpowered tests. If the severity score is high for $\mu_1 - \mu_2 > 0.1$, it is because the observed effect size is very big. Inflated effect sizes corrupt the severity measure of evidence.

Therefore, the severity score is an inadequate measure of evidence and should be rejected. That score is sensitive to the inflated effect sizes provided by underpowered tests. In order to assess the strength of the evidence, one must make sure that a departure from the null is not an artefact of an underpowered test. The severity score is useless for that purpose.

3 The More Power the Better

Naturally, in light of what has just been said, one must try to make sure that a test is powerful in order to generate good evidence against the null. "If the discovery studies were fully powered, inflation would not be an issue"(Ioannidis 2008, p.641). The more power the better.

In order to abide by this principle, one can either make sure that the sample size is big enough or try to replicate the result of the experiment.

As Gelman and Carlin put it:

> The problem, though, is that if sample size is too small, in relation to the true effect size, then what appears to be a win (statistical significance) may really be a loss (in the form of a claim that does not replicate) (Gelman and Carlin 2014, p.642).

Thus, if one can manage to replicate the results of a significant test, then one can be more confident in the evidence. If not, then we should reject the evidence.

In the example presented above, even if S were to repeat her experiment 100,000 times, she would not be able to obtain enough evidence to reject H0. To see this, 100,000 p-values associated with 100,000 replications of the experiment are represented in Figure 1 (See Appendix to reproduce the results).

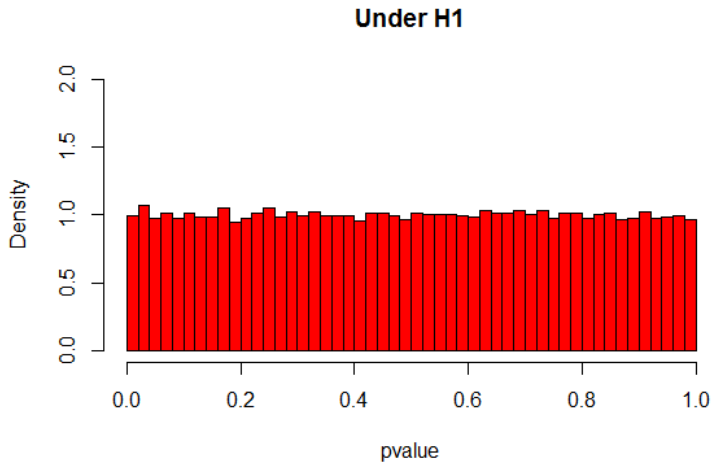

Figure 1: Histogram estimation of the density of the p-values, under the assumption that H1 is true, made with 100,000 simulations

Given that a p-value follows a uniform distribution under H0 but not under H1, S could perform a Kolmogorov-Smirnov test for the uniformity of the p-values. Doing so, she would obtain a test statistic of 0.002 and a p-value of 0.958 (See Appendix to reproduce the results). This means that S would not be able to reject the hypothesis stating that those p-values follow a uniform distribution. This also means that she would not be able to reject the hypothesis stating that the two means are equal.

The Source of Confusion

Unfortunately, proponents of the severity score do not believe that more powerful tests can provide better evidence against the null simply because we can detect minute differences from the null if our tests are powerful enough. Indeed, it is often said that we can always reject H0 with enough observations. Hence, it would

be wrong to conclude that there is an interesting discrepancy between H0 and H1 simply because we reject H0 with a powerful test.

But showing that a powerful test has only warranted the existence of a small discrepancy from H0 does not mean that we have little evidence against H0 and that H1 is not well supported by the evidence. The existence of a small difference from H0, if well justified, is enough evidence against H0. By analogy, a proof that a bone sprained is not worse evidence against the hypothesis that there is not bone damage than a proof that a bone is broken.

There is a clear distinction between (1) claiming that a significant test provides justification for an scientifically interesting difference between H0 and H1 and (2) claiming that it provides justification for a difference of λ between H0 and H1. A small difference between H0 and H1 can be extremely well-justified. What inflated effect sizes show is that if we want to justify the existence of a difference λ (whatever it may be), then we need a significant result obtained with a powerful test.

4 Conclusion

In a nutshell, the severity score is an inadequate measure of evidence and should be rejected. It is sensitive to the inflated effect sizes provided by underpowered significant tests. The point is that inflated effect sizes also inflate severity scores. This has not yet been pointed out in the philosophical literature.

I have illustrated this with an example. In order to assess the strength of the evidence, one must make sure that a departure from the null is not an artefact of an underpowered test. One can do so by taking reasonable precautions against low powered tests, such as trying to replicate the results of a test.

Like it was mentioned in the introduction, the problem of inflated effect sizes

provided by significant and underpowered tests is not merely a theoretical problem. The interested reader can consult (Gelman and Carlin 2014) who mention two specific examples taken from published work. This makes it all the more important to underscore the inadequacies of the severity score as a measure of evidence.

In sum, I have shown that the following quotes also applies to philosophy of science:

> it is not sufficiently well understood that "significant" findings from studies that are underpowered (with respect to the true effect size) are likely to produce wrong answers (Gelman and Carlin 2014, p.649).

Philosophers have overlooked the problem of inflated effect sizes. The winner's curse is crippling the severity measure.

References

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14*(5), 365–376.

Gelman, A. and J. Carlin (2014). Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology 19*(5), 640–648.

Mayo, D. G. and A. Spanos (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science 57*(2), 323–357.

Mayo, D. G. and A. Spanos (2011). Error statistics. *Philosophy of statistics 7*, 152–198.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science 80*(1), 73–93.

Ulrich Schimmack, M. H. and K. Kesavan. Reconstruction of a train wreck: How priming research went off the rails. Accessed: 2017-10-29.