The severity score is fundamentally flawed: A reply to Spanos' ”Severity and Trustworthy Evidence”

November 2, 2022

Contents

1 Introduction

Aris Spanos' recent critical appraisal of my paper on the severity measure of evidence and the Winner's Curse (Spanos 2022) has the merit of allowing me to clarify my arguments. I show that Spanos' paper contains mistakes, contradictions, and an example that ultimately supports my conclusions. The bottom line is that the severity score is fundamentally defective because it favors the rejection of the null hypothesis due to sampling error as opposed to the discovery of true discrepancies from the null. Since the severity score is a function of the test statistics, it is especially corrupted by the large sampling error of underpowered test.

This paper contains two main sections. In the first, I clarify my original arguments. This allows me to pinpoint mistakes and contradictions in Spanos' paper. In the second, I drive my point home with yet another example, one that can allow me to dismiss some more of Spanos' critical comments.

2 Clarifications

I would like to specify that I do not claim that there is a fundamental problem with classical theory-testing. I claim that there is a fundamental problem with the severity interpretation of classical theory-testing. In order to show this I published the following example in which I ask the reader to imagine a statistician named S who has obtained two different samples of 10 independent and identically distributed observations (IID): $(X_1, X_2, ..., X_{10})$ 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.

Perhaps this was not clear in my original paper but the IID property of the observations is a given. It is known to S. There is no issue about the misspecification of the model. Hence Spanos' discussion on this topic is, unfortunately, irrelevant and rather ironic. The way Spanos' paper presents a statistically significant downtrend in the differences between the x_i and the y_j (in the order that they are sampled) (Spanos 2022, p.7) in order to show that the observations are not ID does not only miss the target, it is nothing short of extraordinary when engaging in a discussion on the dangers of making inferences based on small samples. The data used in the simulation is only there for illustration purposes and to allow the readers to conduct the experiment themselves with R. As will be shown in the next section, such examples can easily be produced with not such downtrend.

Besides the IID property of the observations, S only knows two things about the parameters of the two normal distributions:

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

(2) $\sigma_1 = \sigma_2$

and she is interested in knowing if the data could support the idea that there is a difference between the two means that is strictly larger than 0.1. Why 0.1? This is a genuine and important question for S. Smaller differences are not important to her. S does not know that the true difference between the means is 0.01. Otherwise, she would not even see the point of the experiment. She truly believes that there is a discrepancy larger than 0.1 and thinks that even a small sample can generate a significant result. When Spanos writes "detecting a tiny discrepancy $\mu_1 - \mu_2 = 0.01$ will still be a hopeless task with $n = 10$. Intuitively , this amounts to attempting to use $n = 10$ to distinguish between two almost identical densities" (Spanos 2022, p.11). He is quite right. No one is disputing that. Poor S, little does she know.

As for me, the author of the original paper, I have an agenda when I choose 0.1. I want to show that the severity score will mislead S's into strongly believing that there is a discrepancy that is strictly larger than 0.1, which is at least at least ten times larger than the true difference. "Ten times larger" has a dramatic tone to it. Indeed, "Rochefort-Maranda (2020) cherry-picks a particular discrepancy $\gamma_1 = 0.1$" (Spanos 2022, p.10). It is cherry-picking, yet fair play. It is like cherry-picking one black swan to falsify the statement "All swans are white".

Now, 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

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

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 estimators here are both unbiased and consistent. The former means that the expected value of $\bar{X}$, $\bar{Y}$, $\bar{X} - \bar{Y}$, S_1^2, and S_2^2 are equal to the real parameters or interest. In other words, if S were to repeat the same experiment an infinite amount of times, the average of the results would be equal to the unknown parameters. In the Appendix, I repeat the experiment two million times and show that the mean of the estimates is practically identical to the real parameters.

Moreover, as the number of observations tends to infinity, the estimators converge in probability towards the real parameters. Thus, they are consistent. In Figure 1, I illustrate this property for $\bar{X} - \bar{Y}$ by increasing the number of observation by one, starting with ten observations per group, fifty thousand time (See Appendix to repeat the experiment). One can see that the observed difference converges towards 0.01 as the sample size increases. Therefore, when Spanos suggests that the examples uses ”inconsistent estimators” (Spanos 2022, p.11) he is provably wrong. I recommend that the readers try similar lines of code and test the result with the other estimators.

Considering that all the estimators in the experiment are consistent, meaning that they converge in probability towards the true value of the parameter, the following claim is very strange : ”For a particular M_θ; an optimal point estimator $\hat{\theta}(X)$ of θ does not entail the inferential claim $\hat{\theta}(x_0) \simeq \theta$ for a large enough n, wbere “$\simeq$” denotes “approximately equal to” (Spanos 2022, p.16). The estimators discussed here will converge as illustrated in Figure 1.

The problem with the estimators in S’s experiment is that they are quite variable. In other words, the sampling error is large. The problem does not come from some mysterious model misspecification at all (the sampling is very transparent and can be reproduced). Figure 1 shows this large variance (sampling error)

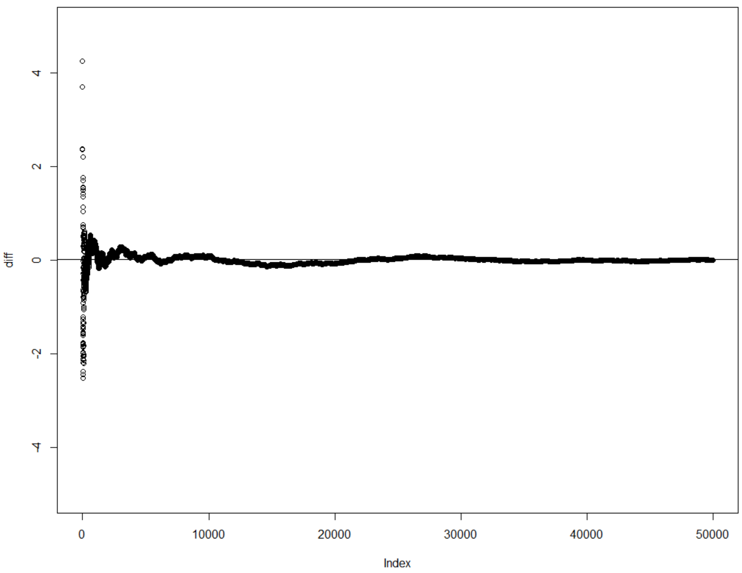

Figure 1: Observed differences (diff) between the sample means of each group showing the consistency of the estimator for the difference between the means of the two populations. Produced with 50,000 simulations. Each time, 1 observation is added. Seed equals 31. The horizontal line is at 0.01.

for small sample sizes and how it is reduced as we add more observations to the samples. Each repetition of the experiment with ten observations per group will generate quite variable results because the sampling error is large. In S's experiment, the lower the sample size, the lower the power of the test and the higher the sampling error. This means that the lower the power, the more significant results will be triggered because of the magnitude of the sampling error and not because of the magnitude of the actual difference between the two means of the populations under study. As we will see bellow, this also means that the high severity score for discrepancies strictly larger than 0.1 will be generated by the sampling error. The fact of the matter is that the severity score is corrupted by the sampling error.

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 (Rochefort-Maranda 2021)). Therefore, S rejects H0 (p-value$<$ 0.05). The test is significant.

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 a substantial difference between H_0 and H_1. Yet, this is what the severity score prescribes.

Going back to S and her experiment, she uses 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.

If the severity score is high, then we can infer that the data provides good evidence for $\mu_1 - \mu_2 > 0.1$.

> **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).

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 (see previous quotes) for such a difference that is at least ten times larger than the true difference.

Again, it is really peculiar when Spanos claims that I ”misinterpret the assignment of probability 0,961 meant for the inferential claim $\gamma > 0.1$ as an endorsement for $\gamma = 0.1$; it is not!” (Spanos 2022, p.10). I have never written such a statement in any of my publications!

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 due to the sampling error. Again, the sampling error corrupts the severity measure of evidence. Therefore, the severity score is an inadequate measure of evidence and should be rejected. In order to assess the strength of the evidence, one must make sure that a departure from the null is not an artifact of an underpowered test. The severity score is useless for that purpose.

What would be better than the severity score to assess the strength of the evidence against the null would either be to repeat the experiment in order to see if the results are robust or, to obtain more observations in order to increase the power of the test and realise that many more observations are needed in order to convincingly reject H0 and track the truth (the difference is 0.01).

The first option is oddly dismissed by Spanos: ”Replicating n=10 many times, say N = 10000; will not address the inherent problem of untrustworthiness” (Spanos 2022, p.11). Yet, Spanos also claims that ” [by] replacing “31” with the other two seeds and simple variations on “7356581” by adding a digit, all the t-tests reverse the author’s result of rejecting H_0”(Spanos 2022, p.8). Now this is equivalent to saying that the repetition of the experiment will show that the evidence against the null is poor and that S’s results are not robust. There is a flagrant contradiction here in Spanos’ paper.

The second option is incompatible with Spanos and Mayo’s view to the effect that a more powerful test (not of the pre-data kind, the real kind, given the true unknown difference between the means) can provide better evidence against the null: ”wherein an α level rejection is taken as more evidence against the null, the higher the power of the test” (Mayo and Spanos 2006, p.344), which is obviously wrong in this case. I’m not sure how this can be debatable. The rejection of the null by S is obviously the result of the sampling error and therefore constitute poor evidence against the null. Even Spanos acknowledges that ”[a] huge sample size (n= 5312800) [is] called for” (Spanos 2022, p.10) in order to reject the null more convincingly.

More power yields a better estimate of the true difference (see previous comments on consistent estimators) and therefore better evidence against the null. In fact, one would need more than one hundred thousand repetitions of the experiment in order to find out that the distribution of the p-values is not uniform and that H_0 should be rejected. The Appendix shows such a result with three million repetitions, starting with a seed of 31 in order to meet Spanos’s challenge: ”It’s not obvious why the author did not use the seed “31” when simulating other NIID data in the same paper”(Spanos 2022, p.10). (more on the choice of seeds in the next section).

Given that a p-value follows a uniform distribution under H0 but not under H1, S could perform a Kolmogorov-Smirnov test with n=3,000,000 for the uniformity of the p-values. Doing so, she would obtain a test statistic of 0.0017245 and a p-value of 0.00000003561 (See Appendix to reproduce the results). This means that S would be able to convincingly reject the hypothesis H_0 with the help of a test that relies on a huge sample size. This clashes with the rejection of H_0 with only ten observations per group. The more powerful test clearly provides better evidence against H_0 here.

3 A New Example

Now let us assume that the observations were problematic in the previous example, that the model was somehow misspecified even though the result comes from a very transparent simulation. Then let us conjure up a similar example, but one that is not misspecified in the way that Spanos would object to.

Imagine the same scenario with S. This time around S makes 50 observations for each group and she is now interested in a difference larger than 2.5 between the two means. The pre-data power to detect a difference like that is pretty decent: 0.7989 and the true difference is still of 0.01.

She then observes a difference of 4.354739, a test statistics of $t = 3.2884$, and she rejects H_0 with a p-value of 0.0007. (See Appendix to repeat the experiment). Again, she computes a severity score for a difference strictly bigger than 2.5 of 0.9177493. She is thus mislead once again by the severity score.

Now this example can be generated with a seed of 947878140 (See Appendix). Contrary to what Spanos suggests: ”The seed used by Rochefort-Maranda (2020) for the data in figure 2 is “31,” which is an unfortunate choice due to its smallness” (Spanos 2022, p.8), the size of the seed does not matter at all. I can find such example at will no matter the size of the seed. Moreover, Spanos will find it impossible to find a trend in the differences of the observations in the order they are sampled (See Figure 2).

In other words, examples like the ones I gave are easy to produce. For very low powered tests, they occur slightly more often than five percent of the time and the best way to weed them out is to repeat the experiment or to increase the power of the test in order to find genuine evidence evidence against the null.

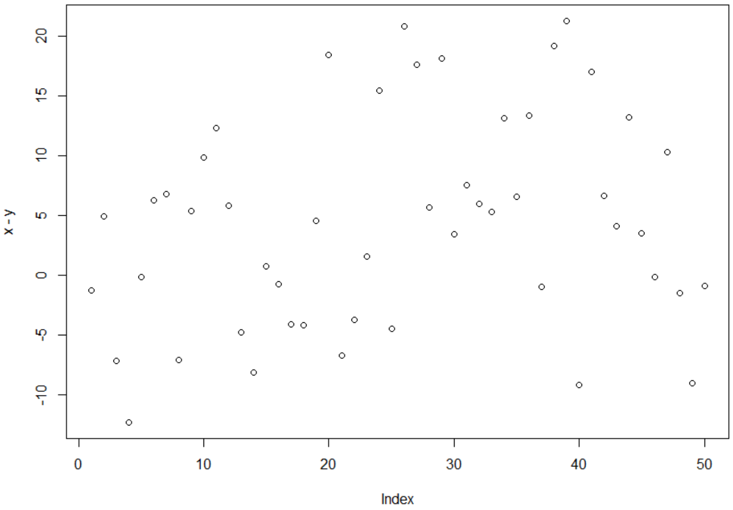

Figure 2: Plot of the differences for the observations between each group in the order they were sampled.

3.1 The More Power the Better

The slogan ”The more power the better” is not only good for pre-data power analysis. Powerful experiments can detect smaller discrepancies, but that is not an issue at all unless one gives the false impression that a significant test necessarily means that the discrepancy is large when it is not. Moreover, small discrepancies do not suggest that H_0 is true or even plausible.

misconception: It is not because we have strong evidence for a small(er) discrepancy from the null that we have less evidence against the null. Whether I find a thousand dollars in my pocket or ten cents, both are equally good evidence against the statement ”I have no money in my pocket”. The size of the discrepancy cuts no epistemic ice.

fact: (Ceteris Paribus) Sampling more observations will never be detrimental to a statistical inference. More observations means less sampling error. Just think of how informative a census of the population can be as opposed to a probabilistic survey.

There is a clear distinction between (1) claiming that a significant test provides justification for an scientifically interesting difference between H_0 and H_1 and (2) claiming that it provides justification for a difference of λ between H_0 and H_1. A small difference between H_0 and H_1 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.

In Spanos' paper, he is asking us to imagine two tests that are barely significant. One (test A) uses more observation than the other (test B). He also correctly points out that ”according to Rochefort-Maranda (2020), test (A) provides stronger evidence against H0: $\mu_0 = 0.5$ because $n1 > n2$” (Spanos 2022, p.16). Then he proceeds to claim that this is false because the lower bound confidence interval of test B is further away from 0.5. In reaction, I would simply say that if test A is barely significant and that we can find a larger estimate for μ with less observation with test B, then test B is corrupted by the higher sampling error and the Winner's Curse. So, yes, test A provides better evidence against H_0 even if there is only a small discrepancy at play. To claim that B provides better evidence against the null is tantamount to saying that we should favor the rejection of the null hypothesis due to sampling error as opposed to the discovery of true discrepancies from the null. John Pratt (1961) is simply mistaken (See Spanos 2022, p.16).

4 Conclusion

In sum, I still stand by my original conclusion: the severity score is fundamentally flawed. It favors the rejection of the null hypothesis due to sampling error as opposed to the discovery of true discrepancies from the null. At the root of the problem is the mistaken view according to which the detection of a large discrepancy is better evidence against H_0 than a smaller one. When holding on to such a view, powerful test are deemed to provide less evidence against the null when in fact they produce the most reliable conclusions one can obtain. In statistics, obtaining more observations (i.e., a larger sample size) can be costly or operationally impossible, but never detrimental to the quality of an inference because it reduces the variance associated with the sampling procedure (the sampling error). Of course, one can be mislead into thinking that there is a large discrepancy from the null when a powerful test is significant. But this does not mean that powerful test provides poor evidence against the null.

References

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.

Pratt, J. W. (1961). Review of lehmann's testing statistical hypotheses. *Journal of the American Statistical Association 56*(293), 163–167.

Rochefort-Maranda, G. (2021). Inflated effect sizes and underpowered tests: how the severity measure of evidence is affected by the winner's curse. *Philosophical Studies 178*(1), 133–145.

Spanos, A. (2022). Severity and trustworthy evidence: Foundational problems vs. misuses of frequentist testing. *Philosophy of Science*, 1–31.