A Paradoxical Feature of the Severity Measure of Evidence

Guillaume Rochefort-Maranda

March 26, 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, pp.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].

In this paper, I put forward a paradoxical feature of the severity score as a measure of evidence. To do this, I create a scenario where a frequentist statistician S is interested in finding out if there is a difference between the means of two normally distributed random variables. The null hypothesis (H0) states that there is no difference between the two means.

A Student's t-Test test yields a significant result and S uses the severity score to show her peers just how much a difference strictly greater than 0.1 is warranted by the data. The severity score for such a difference is quite high. Hence S believes that she has obtained good and strong evidence for such a difference.

However, I also show that when S repeats her experiment 100,000 times and performs a Kolmogorov-Smirnof test for the uniformity of the p-values, she does not find a significant result against H0. This is paradoxical. According to the severity score, the first test provides excellent evidence for a difference between the two means. Yet, the second test provides no evidence whatsoever. I argue that this paradox must lead to the rejection of the severity score as a measure of evidence.

The paradox illustrates the fact that the severity score will inevitably fail to adequately measure the evidence provided by a significant test with low power. Tests with very low power will be significant only if the observations are deviant under both H0 and the alternative hypothesis H1. Therefore, the results of those significant tests will generate misleadingly high severity scores for differences between H0 and H1 that are excessively overestimated.

Of course, significant tests with low power are relatively rare. They happen slightly more often than the significant level. If the severity measure of evidence gets things right most of the time, what is the problem? The problem is that we can do better.

A significant result will provide better evidence against the null if the test is more powerful or if we have succeeded in rejecting H0 after several repetitions of the test. This is partially incompatible 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, pp.334.

2 The Scenario

Here is the scenario. 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 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 4.249611; a test statistic $t_{obs} = 1.914$; and a p-value = 0.03583 (See Appendix to reproduce the results). Therefore, S rejects H0 (p-value< 0.05). The test is significant.

Now, 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:

$$t_s = \frac{(2.688654 + 1.560957) - (0.1)}{S_p \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$SEV(\mu_1 - \mu_2 > 0.1) = \boldsymbol{F(t_s)} = 0.9610043$$

where $\boldsymbol{F(t_s)}$ 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 we have 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, pp.169) for more details on how to compute such a severity score. If that probability 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.

In a nutshell, the important result to remember here is that S has found a significant result (p-value=0.03583). She thus rejects H0 and finds a high severity score for the claim $\mu_1 - \mu_2 > 0.1$ (severity score=0.9610043). Hence, S believes that she has good evidence for such a difference.

However, when she repeats her experiment 100,000 times, she is 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).

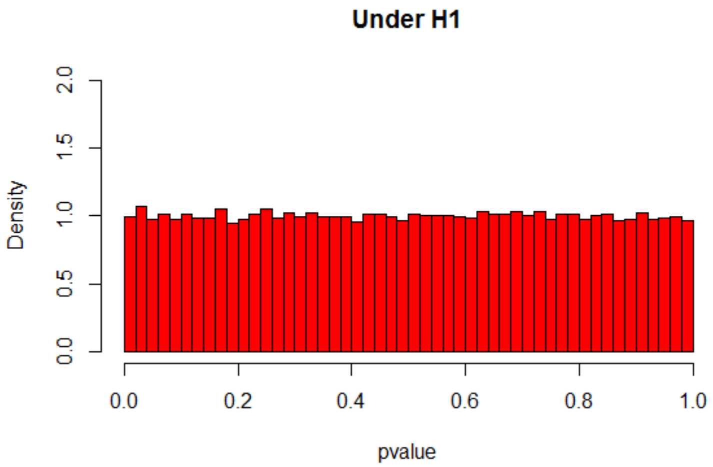


Figure 1: Histogram estimation of the density of the p-values, under the assumption that H1 is true, made with 100,000 simulations

Given that a p-value follows a uniform distribution under H0 but not under H1, S has done a Kolmogorov-Smirnov test for the uniformity of the p-values. Doing so, she obtained a test statistic of 0.0016108 and a p-value of 0.9576 (See Appendix to reproduce the results). This means that S cannot reject the hypothesis stating that those p-values follow a uniform distribution. This also means that she cannot reject the hypothesis stating that the two means are equal.

3 The Paradox and its Solution

So, here lies the paradox: S has good and strong evidence that there is a difference strictly greater than 0.1 between the two means according to the severity score associated with a Student's t-Test. However, 100,000 repetitions of that very same experiment cannot provide enough evidence against H0 according Kolmogorov-Smirnof test for the uniformity of the p-values. This does not make sense.

On one hand, the severity measure tells us that the first test is enough to reject H0 and also sufficient to justify that there is a difference strictly greater than 0.1 between the two means. On the other hand, by claiming that S has good quality evidence for her claims, it looks like we are purposefully ignoring all the other possible p-values and severity scores that this experiment has produced (see Figure 1). This feels like cheating.

Fortunately, this kind of paradox is easy to solve. In light of the Kolmogorov-Smirnof test, we see that the result of the Student's t-Test is a misleading anomaly such that we know that it does not provide good evidence. However, it is not so easy to understand why this is a problem for the severity measure of evidence. Given that deviant results are inherent to statistical inferences, it is normal to expect the severity measure to fail every now and then. As long as it gets things right most of the time, there is no real problem with that measure of support.

But there is more here than meets the eye. The paradox that I have exposed here shows the results of two tests. Both provide different conclusions with respect to H0. However, the Kolmogorov-Smirnof is so much more powerful that the Student's t-Test (It is based on 100,000 observation!). If we have not been able to detect a difference with such a powerful test, then we must conclude that the result of the Student's t-Test is deviant and that is why S has been mislead into thinking that there might have been a difference as big as 0.1 when the true difference is 0.01.

Now, here is the crux of this paper. Tests with very low power will be significant only if the observations are deviant under both H0 and H1. Therefore, the results of those significant tests will generate misleadingly high severity scores for differences between H0 and H1 that are excessively overestimated. In other words, that measure of evidence is bound to fail in those cases. It will inevitably fail to adequately measure the strength of the evidence provided by tests with low power. This is what happened with the Student's t-Test and this is why we can generate the paradox.

The problem for the severity score is that we can do better. We can avoid its pitfalls by including the power of the test into the measure of evidence (the more the better) and/or to include the results of repetitions into the measure of evidence (the more the better). Repetitions can indicate (not prove) if we are dealing with deviant results or not.

Even if S had not repeated her experiment 100,000 times, she would not have been warranted into thinking that she had good evidence for her claims because she failed to check whether or not her results are deviant. This is a costly mistake especially when S is working with only 10 observations from each group. Only two or three repetitions of her test would have been enough to realise that she is probably making a mistake and that the severity score has mislead her. Thus, one must reject the severity score as a good measure of evidence. We cannot rely on that measure without first establishing that our test's power is high enough. High powered tests will not trigger significant results with deviant observations.

4 Conclusion

In sum, I have created a scenario where a statistician S finds a significant result (p-value=0.03583) to a one-tailed Student's t-Test and a high severity score for the claim $\mu_1 - \mu_2 > 0.1$ (severity score=0.9610053). This is supposed to show that S obtained good evidence for such a difference (see the first quote in the introduction). However, if we wish to maintain that S obtained good evidence for such a difference, then we encounter a paradox where one test provides good evidence against H0 but where 100,000 repetitions of that test do not.

Scenarios like this one are easy to construct because the severity score is computed independently of the power of a test (see how it is computed in Section 2). This is a major flaw. Here is the recipe that I have followed: take any statistical test such that if their power is low enough, then the distributions of the test statistic under H1 and H0 are almost identical. This will maximise the variance of the p-value under H1. Then, choose any significant result with a high severity score for a given hypothesis (with a significance level of 0.05, they happen a little more often than 5% of the time). You will then find a severity score paradox because several repetition of the test will not be sufficient to reject H0 with a Kolmogorov-Smirnov test for the uniformity of the p-values.

The main lesson here is that tests with very low power will be significant only if the observations are deviant under both H0 and H1. Therefore, the results of those significant tests will generate misleadingly high severity scores for differences between H0 and H1 that are excessively overestimated. In other words, that measure of evidence is bound to fail in those cases. It will inevitably fail to adequately measure the strength of the evidence provided by significant tests with low power.

Of course, proponents of the Error-Statistical philosophy acknowledge the importance of repeating experiments. But their measure of evidence is only defined in function of one test statistic, i.e., in function of the result of one test. That is why a very deviant test statistic can generate a high severity score. Moreover, the power of the test is irrelevant when we compute the severity score. A more appropriate measure of evidence would need to be able to encompass the results of multiple tests and take power as a safeguard against misleading deviant results.

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.

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

Appendix

The test with 10 observations per group

```
set.seed(31)
x<-rnorm(10, 1.01, 6)
y<-rnorm(10, 1, 6)
grp<-c(rep(1, 10), rep(2, 10))
z<-c(x, y)
dat<-as.data.frame(cbind(z, grp))
test<-t.test(z~grp, data=dat, var.equal=T, alternative = "greater")
test

##
##  Two Sample t-test
##
## data:  z by grp
## t = 1.914, df = 18, p-value = 0.03583
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3994825       Inf
## sample estimates:
## mean in group 1 mean in group 2
##        2.688654       -1.560957
```

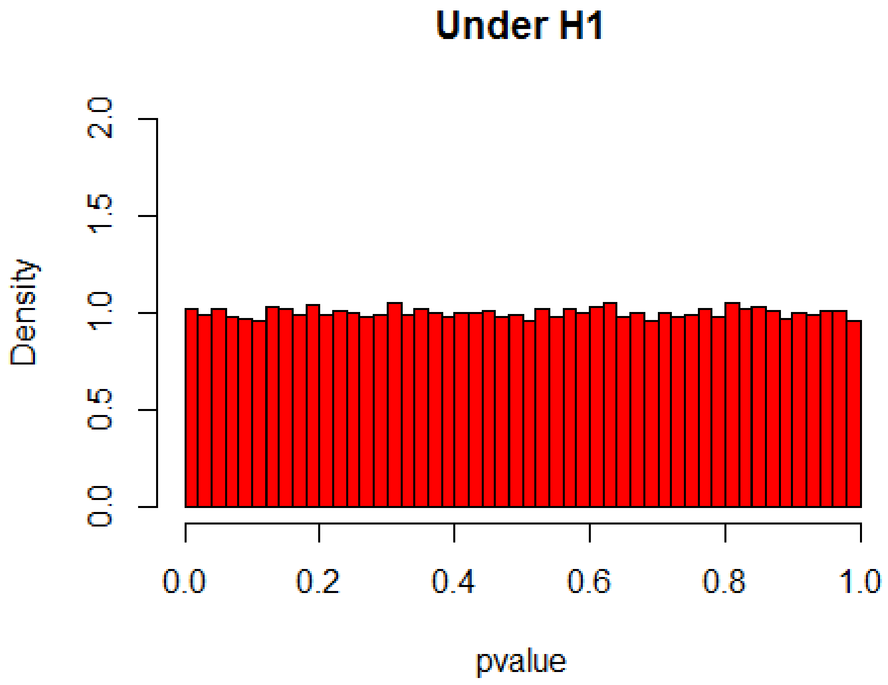
We find the severity score for a difference strictly larger than (0.1).

```
set.seed(31)
x<-rnorm(10, 1.01, 6)
y<-rnorm(10, 1, 6)
s1<-sum((x-mean(x))^2)/9
s2<-sum((y-mean(y))^2)/9
sp<-sqrt((9*s1+9*s2)/18)
a<-sqrt((1/10)+(1/10))
t<-((2.688654 + 1.560957)-(0.1))/(sp*a)
sev<-pt(t, df=18, lower.tail = T, log.p = FALSE)

sev

## [1] 0.9610043
```

We find the distributions of the p-value under the assumption that H1 is true for the test with 10 observations per group

```
pvh1<-rep(NA, 100000)
```

```r
for (i in 1:100000){
  x<-rnorm(10, 1.01, 6)
  y<-rnorm(10, 1, 6)
  grp<-c(rep(1, 10), rep(2, 10))
  z<-c(x, y)
  dat<-as.data.frame(cbind(z, grp))
  test<-t.test(z~grp, data=dat, var.equal=T, alternative = "greater")
  pvh1[i]<-test$p.value
#  print(i)
}
pvalue<-pvh1
hist(pvalue, freq=F, 50, ylim=c(0, 2), col=2, main="Under H1")
```



We perform a Kolmogorov-Smirnov test for the uniformity of the p-values under H1.

```r
ks.test(pvh1, "punif")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  pvh1
## D = 0.0016108, p-value = 0.9576
## alternative hypothesis: two-sided
```