Statistical Power and P-values: An Epistemic Interpretation Without Power Approach Paradoxes

Guillaume Rochefort-Maranda

December 16, 2017

Contents

1 Introduction **2**

2 The Paradox **3**

2.1 Technical Background 3

2.2 Epistemic Interpretations 4

2.3 A Paradox 6

3 The Consensus **8**

4 The Solution **10**

5 Conclusion **13**

1 Introduction

It has been claimed that if statistical power and p-values are both used to measure the strength of our evidence for the null-hypothesis when the results of our tests are not significant, then they can also be used to derive inconsistent epistemic judgements as we compare two different experiments. Those problematic derivations are known as power approach paradoxes. The consensus is that we can avoid them if we abandon the idea that statistical power can measure the strength of our evidence (Hoenig and Heisey 2001; Machery 2012). In this paper however, I put forward a different solution. I argue that every power approach paradox rests on an equivocation on ”strong evidence”.

The main idea is that we need to make a careful distinction between (i) the evidence provided by the quality of the test and (ii) the evidence provided by the outcome of the test. Both provide different types of evidence and their respective strength are to be evaluated differently.

Without loss of generality[1], I analyse only one power approach paradox in order to reach this conclusion. But first, I set-up the frequentist framework within which we can find such a paradox.

[1]My analysis is without loss of generality because every other formulation of the paradox rests on the same idea that I reject : power and p-values measure the same thing.

2 The Paradox

2.1 Technical Background

A statistical test contrasts two mutually exclusive propositions: H0 (the null hypothesis) and H1 (the alternative hypothesis). It also requires a decision rule. The decision rule states that if the probability of making observations as extreme or more extreme than the ones we have made is too low under the assumption that H0 is true, then we should reject H0 and thus accept H1 (significant result). If it is not too low, then we should not reject H0 (non-significant result).

That probability is called "the *p-value*". What "extreme" means depends on the probability distributions that are specified by H0 and H1. What "too low" means is relatively arbitrary. It is determined by the investigators and it is usually smaller than or equal to 0.01 or 0.05 depending on the field of study. That value is called "the significance level" ("α" for short) and it is also the probability (or the upper bound probability) of making a type I error if the null hypothesis is true. A type I error is the rejection of H0 when in fact H0 is true. A type II error, on the other hand, is the failure to reject H0 when H1 is true. If H1 is true, the probability of committing this kind of error is called "β".

Hypotheses are either simple or composite. They are simple only if every parameter involved is specified. If not, then the hypothesis is said to be composite. In fact, we can think of a composite hypothesis as a set of simple hypotheses. This distinction is important because it implies that we do not know the exact probability distribution of our observations un-

der a composite hypothesis. Therefore, we cannot always know the exact probability of making a type I or a type II error.

If H0 is composite, then we define our test in such a way that the probability of making a type I error is at most equal to the significance level[2]. In other words, we assume that the type I error is equal to the significance level even though it could in fact be smaller. On the other hand, if H1 is composite, then we only know that β is a function of the unspecified parameters. As we shall see, this has an important impact on the epistemic interpretation of a statistical test.

To give an example of a decision rule, here is one for a unilateral test with $\alpha = 0.05$ and where P_{H0}(test statistic $\geq$ observed test statistic) is the *p-value*:

- If P_{H0}(test statistic $\geq$ observed test statistic) ≤ 0.05, reject H0 and accept H1.
- If P_{H0}(test statistic $\geq$ observed test statistic) ≤ 0.05, do not reject H0.

2.2 Epistemic Interpretations

Now, the epistemic significance of α and β should be obvious. Since we wish to avoid making mistakes, then, *ceteris paribus*, we will prefer to perform a test such that α has the smallest value and $1 - \beta$ (the power of a test, "π" for short), the largest. But there is more. We can also use α and π to measure of the strength of our evidence.

[2]That is one of the reason why α can be an upper bound probability of making a type I error when H0 is true.

If the ratio $\frac{\pi}{\alpha}$ is greater than 1, then we know that a significant result is more likely under H1 than H0. Therefore, we know that a significant result provides evidence for H1. Moreover, we can measure the strength of that evidence against H0 with that ratio. The greater the difference of that ratio with 1, the stronger is our evidence against H0.

Similarly, if the ratio $\frac{(1-\alpha)}{(1-\pi)}$ is greater than 1, then we know that a non-significant result is more likely under H0 than under H1. Thus, we know that a non-significant result provides evidence against H1. We can also measure the strength of our evidence against H1 with that ratio. The greater the difference of that ratio with 1, the stronger is our evidence against H1.

In many cases however, it is not possible to use such ratios because the power of our test can be too small. For example, when we use a *t-test* in order to tell if the theoretical means of two variables that can be measured in two different groups are significantly different (H0: $\mu_1 = \mu_2$ VS H1: $\mu_1 \neq \mu_2$), we simply cannot conclude that H1 is false given that our test is non-significant. The reason why we cannot do such a thing is that the difference between the two means can be infinitely close to zero such that the power of that test can be very small. As mentioned earlier, when we are facing a composite alternative, β (and π) is a function of the unspecified values of the parameters involved in H1.

But there is always a way around this kind of problem. In practice, we would not be interested in the exact equality between μ_1 and μ_2. We would be interested in trying to establish if there is a negligible or trivial difference between them. We could therefore specify the magnitude of

that negligible difference and ignore the fact that the power of our test could, in principle, be very small. In other words, we could interpret a non-significant result as evidence that the difference between μ_1 and μ_2 is negligible (*i.e.*, that H0 is approximately true) if we know that our test was powerful enough to detect a greater difference.

Thus, it seems quite reasonable to say that the strength of our evidence can be measured by the significance level of our test and its power. But *p-values* also seem to have a similar epistemic function. A *p-value* gives us the smallest α that would have yielded a significant result.

This means two things. Firstly, if the test is significant, then we know how much the measure of π/α can be increased by decreasing α in order to assess positively the strength of the evidence against H0. Secondly, if the test is not significant, then we can know how much the measure $\frac{(1-\alpha)}{(1-\pi)}$ can be increased by decreasing $1-\alpha$ in order to assess positively the strength of the evidence for H0.

In other words, it looks like our evidence against H0 can be stronger than we would have thought a priori the smaller the p-value. Likewise, if the idea of accepting H0 (rejecting H1) makes sense (*i.e.* if our test is powerful enough and if we are in a position to know that it is), it appears that our evidence against H1 can be stronger than we would have thought a priori the larger the p-value.

2.3 A Paradox

However, as soon as we reach this conclusion, we need to grapple with puzzling epistemic paradoxes. To see this, let us consider two different

t-tests.

As mentioned before, a t-test can be used when we wish to tell if the theoretical means of a variable that can be measured in two different groups are significantly different (H0: $\mu_1 = \mu_2$ VS H1: $\mu_1\mu_2$). For example, it could be used to tell if the expected diameter of a doughnut produced by shop is significantly different from the expected size of a doughnut produced by another shop of the same branch (a bilateral test). To conduct this experiment, we would need to take two samples of doughnuts from each shop and make a few assumptions in order to establish the probability density function of our t-statistic under H0. Then, we would need to estimate the mean and the variance of each group and obtain a t-statistic (our observation).

Now consider two such experiments: E1 and E2. Each focuses on a different pair of doughnut shops. Imagine also that we are ready to admit that H0 is basically correct if the difference is no greater than 1mm and that we can tell if our test is powerful enough to be able to detect a difference that is greater than 1mm.

Suppose that the resulting tests yield non-significant results, yet the p-value in E1 is smaller ($p1 < p2$) and the power of E1 is greater ($\pi 2 < \pi 1$). Therefore, if our evidence against H1 is stronger the greater the power (see previous ratios) and if our evidence against H1 is weaker the smaller the p-value, then we are led to admit that E1 provides the strongest evidence against H1 (for H0) and the weakest evidence against H1 (for H0). Hence, we have on our hands what is now known as a power approach paradox.

The recipe for such paradoxes is quite simple. First, we define two sta-

tistical tests, T1 and T2, such that the former is more powerful than the latter. Then we assume that they both yield non-significant results and stipulate that the p-value associated with T1 is smaller than the p-value associated with T2. Consequently, we can use the epistemic interpretation of the statistical power and the p-value in order to reach paradoxical conclusions such as "The result of T1 provides more evidence for H0 (against H1) and less evidence for H0 (against H1) than the result of T2". The paradox presented in this section and the ones we can find in (Hoenig Heisey 2001) and in (Machery 2012) follow exactly this recipe.

3 The Consensus

The main solution to this problem is to abandon the idea that the power of a test can measure the strength of the evidence. We can abandon this idea either because we believe that the power of a test cannot justify the acceptance of H0, or because we believe that it is an inappropriate measure even though it can be used to justify the acceptance of H0. John Hoenig and Dennis Heisey endorse the former belief, whereas Edouard Machery endorses the latter.

In their article, Hoenig and Heisey do not voice any objection against the epistemic interpretation of the p-value as a measure of the strength of the evidence against H0. However, they maintain that a non-significant result does not allow us to infer H0 and that the strength of the evidence cannot be measured with the power of the test. If we use the power of a test to make such an inference, we end up with a power approach paradox

end of discussion.

Edouard Machery, on the other hand, believes that the power of a test can only determine if we should accept H0 or not:

> one can assume that the power of an experiment is a property of a test that allows for the application of a decision rule specifying when the null hypothesis is to be accepted and the alternative hypothesis rejected: accept the null hypothesis from a negative result when and only when power is above some threshold (Machery 2012, p.816).

However, he claims that the power of a test does not measure anything. This obviously solves the paradox presented in section 2, because such an interpretation would prevent us from saying that E1 provides the strongest evidence against H1.

Machery would also claim that it is a mistake to say that E1 provides the weakest evidence against H1 because he believes that a p-value, just like the power of a test, does not measure anything. It merely determines if we should reject H0 or not:

> a p-value is a property of the data that allows for the application of a decision rule specifying when the null hypothesis is to be rejected and the alternative hypothesis accepted [...]. Under this interpretation, p-values are not taken to measure the strength of evidence against the null hypothesis (Machery 2012, p.816).

In the following section, I put forward a very different solution to the

power approach paradox. I maintain that the power of a test and the p-value can measure the strength of our evidence and argue that power approach paradoxes rest on an equivocation on ”strong evidence” (the equivocation is on the notion of evidence). I show that the paradoxes dissolve when we cash-out the meaning of ”strong evidence” adequately and interpret p-values, significance levels, and statistical power accordingly.

4 The Solution

As a matter of fact, it is relatively easy to create paradoxes that are very similar to power approach paradoxes. Here is an example. A police officer stops a driver on the side of the road. She suspects that the individual is driving under the influence of alcohol. Luckily, she has at her disposition two different instruments to measure blood alcohol concentration. Lets call them ”instrument A” and ”instrument B”. Instrument A is fairly reliable but instrument B is not. Now suppose that we know that our reflexes are more likely to be dangerously impaired if our blood alcohol content is above 0.05 mg/ml. On that occasion, instrument A produces a reading of 0.052 and instrument B produces a reading of 0.07.

Because instrument A is more reliable, the officer ought to believe that it provides the strongest evidence for the claim that the drivers faculties are dangerously impaired. Yet, because As reading is closer to the threshold, she also ought to believe that it provides the weakest evidence for the same claim. Clearly, something does not sound right. We are facing yet another paradox.

In this case, this is because we are not in fact comparing the same things. On the one hand, when we claim that instrument A provides the strongest evidence because it is more reliable, we are comparing the credibility of the support provided by the output of the instruments. On the other hand, when we claim that instrument A provides the weakest evidence, we are in fact trying to compare the degree to which the instruments outputs support the hypothesis. To do this adequately, we would in fact need to make a counterfactual statement such as "Were instrument B as reliable or more reliable than instrument A, it would have provided more support for the hypothesis".

As we can see, there is nothing paradoxical in making both comparative judgments. The paradox rests on an equivocation on "strong evidence". There is an important distinction to make between the credibility of the support given by the output of an instrument and the degree to which this output supports the hypothesis. One way to distinguish both concepts is to realise that the credibility of the support does not depend on the actual output of the instrument whereas the degree of support does. In what follows, I will argue that the power approach paradox described in section 2 also rests on the same kind of equivocation and thoroughly deconstruct the misleading chain of reasoning that was presented there.

The degree of support given by the quality of a statistical test is different from the degree of support given by the output of that test. Both can be evaluated for their strength as I have explained in section 2.2. When we are considering the result of a statistical test as evidence that a hypothesis is true, we must evaluate the quality of the test with the ratios presented

in section 2.2.

π/α and $\frac{(1-\alpha)}{(1-\pi)}$ can be interpreted as providing measures of the quality of the test because they are features of the decision procedure. Those values do not depend on the actual outcome of the experiment. They determine how the evidence is produced. As such, when we say that a test with more power provides stronger evidence against H1, we ought to mean that it provides more credible support against H1.

The magnitude of a p-value on the other hand is determined by the experimental outcome and it can be interpreted as providing a measure of the degree to which the output supports a hypothesis a posteriori. Hence, when we say that a test with a smaller p-value would have provided stronger evidence for H1, we ought to mean that the evidence would have provided more support for H1. The credibility of the support is not under scrutiny in this context. The degree of a posteriori support is.

Now we can understand why the chain of reasoning that led to the power approach paradox in section 2 was misleading. Firstly, the claim that the evidence against H1 was stronger in E1 because $\pi 2 < \pi 1$, meant that the the support against H1 was more credible because the quality of the test was stronger.

Secondly, the claim that the evidence against H1 was weaker in E1 because $p1 < p2$, meant that the information given by p1 was less supportive of H0 than the information given by p2. It was the degree of support given by the p-values that was being compared. The idea that was being conveyed was that ratio $\frac{(1-\alpha)}{(1-\pi)}$ could not be increased as much within the context of E1 as it could be within the context of E2. As such, both claims

are not in conflict. Thus, the paradox rests on an equivocation. Once we make the appropriate distinctions, similar paradoxes dissolve.

5 Conclusion

In this paper I have argued that power approach paradoxes dissolve when we cash-out the meaning of "strong evidence" adequately and interpret p-values, significance levels, and statistical power accordingly. Clearly, Hoenig, Heisey and Machery failed to do so.

The quality of the evidence is provided by i) **the quality of the test** and ii) **the output of the test**. I have claimed that the quality of the test can be measured by the following ratios: π/α and $\frac{(1-\alpha)}{(1-\pi)}$. I have also claimed that the support provided by the p-value (the output of the test) can be measure by how much we can improve the measures provided by those ratios a posteriori. What we have here is thus a coherent epistemic interpretation of a statistical test that is free of power approach paradoxes.

I would now like to conclude by saying that the power approach paradox is a problem within the frequentist framework. As such, I have put forward a solution within that framework. Whether or not the frequentist approach can withstand other criticisms is a different story. However, if the frequentist approach turns out to be epistemically unsound, it is not because of the power approach paradox.

References

Cox, D. R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology 14*, 325–331.

Hoenig, J. M. and D. M. Heisey (2001). The abuse of power. *The American Statistician 55*(1).

Machery, E. (2012). Power and negative results. *Philosophy of Science 79*(5), 808–820.