The Principle of Total Evidence and Classical Statistical Tests

Guillaume Rochefort-Maranda

December 14, 2017

Contents

1 Introduction

Classical statistical inferences have been criticised for various reasons. To assess the soundness of such criticisms is a very important task because they are widely used in everyday scientific research. This is one of the reasons why the philosophy of statistics is an exciting field of study.

In this paper, I focus on two such criticisms. The first one claims that the use of the p-value violates (or can violate) the principle of total evidence (PTE). It is a thesis that has been defended by Elliott Sober and Bengt Autzen. The second one says that the result of classical tests does not only depend on the data but on the sampling plan of the experimenter also. The underlying criticism of course is that the sampling plan is not part of the evidence and that classical tests therefore violate PTE. The intentions of the experimenter should not affect the result of an inference. See (Howson and Urbach 2006; Romeijn 2017).

My aim is to show that both criticisms are unsound. Doing so, I hope to clarify the concept of p-value and the nature of the evidence in classical statistical tests. The point of my paper is to show that the identification of the evidence on which those criticisms rest is inadequate.

This paper contains three main sections. In the first section, I define its focus and provide a non-technical/basic explanation of frequentist tests with the help of an analogy. In the second section, I dismiss Sober's and Autzen's views on PTE and p-values. In order to make my point, I derive absurd conclusions by using the notion of evidence on which they construct their criticisms. I shall conclude by proposing a more adequate characterisation of evidence for frequentist tests.

In the final section, I rebut the criticism based on the sampling plan also known as the "stopping rule criticism". I argue that the different conclusions that stem from different stopping rules is explained by a difference in the evidence that has

been used. The intentions of the experimenter have nothing to do with it.

2 Frequentist Tests: An Intuitive Guide

In this paper, I shall not discuss the PTE **per se**. My goal is to pinpoint three mistakes about frequentist inferences from which the two criticisms mentionned in the introduction stem:

- The aim of a significance test is to find out whether or not the observations are sufficiently improbable under the null hypothesis.
- It is preferable to use tests such that the p-value is a sufficient statistic.
- The fact that different conclusions stem from different stopping rules is explained by a difference in the intentions of the experimenters.

Those mistakes have led some to conclude that frequentist inferences violate (or can violate) PTE. By exposing those mistakes for what they are, I show that those arguments are unsound.

In order to reach that goal, I do not need to analyse PTE nor do I need to take issue with how it is used in the arguments I shall discuss in the following sections. I rely on a very general definition of this principle according to which one should use all evidence and only the evidence when making an inference (See (Neta 2008, p.90) for such a general defintion).

I will however attempt to clarifying the procedure involved in classical inferences and to identify the information required to make a frequentist inference. That information is what I call ”the evidence”.

Frequentist tests function a little bit like we would evaluate an archer after several trials. First, we will establish a region near the center of the target where

most of A's arrows should fall under the assumption that she is a good archer. If most of her arrows do not fall within that region, then we shall no longer think of her as a good archer. This kind of inference consists in making a prediction as to where the arrow will fall under the supposition that A is a good archer and a disposition to reject that assumption should A's arrows mostly fall outside that predicted region. The separation of the target into two regions is essential to the inference because we cannot predict where most of the arrows will land exactly. We can only predict the area where they will land.

Things are very similar with frequentist tests except that we do not test archers, but values of (or constraints on) parameters and we do not observe arrows, but test statistics. However, much like in the archer scenario, we predetermine a region where a test statistic will fall with a high probability under the assumption H0 that a certain parameter (it could be a vector) holds as opposed to some other. We shall reject that assumption H0 should the test statistic fall outside that predicted region.

Again, we need to separate the space that the statistic can occupy into two regions in order to be able to make a prediction. It is usually impossible to predict the exact value of a statistic. Just think of a probability density like the normal distribution. The probability of every state is 0. It that case it is impossible to predict a single state but only a range of states.

Notice here how there is an important distinction to be made between the probability of falling within a certain region and the probability of a particular test statistic. The first one is needed to make an inference. The second one is relatively uninformative by itself. Every possible test statistic can be very improbable. What matters is whether or not it falls within a certain region. This will determine the outcome of the inference.

Furthermore, notice how we do not define a probability density or mass on the hypotheses that we are testing. We define probability densities or masses on

statistics. This means that our inferences will not attribute probabilities on the hypothesis. For example, there is no such thing as the probability of H0 given the test statistic within a frequentist framework. One might want to analyse frequentist tests in those terms but this is a no a path I will attempt to follow in this paper.

The main focus of this paper is to shine a light on the evidence used in a frequentist inference. The information that we need to use in order to reach a conclusion in a classical context is quite rich. We need to determined the correct distribution of the test statistic. We also need to know how to establish the correct regions on which to define the test. This requires a certain knowledge of the alternative hypotheses. We also need to know where the test statistic lies in order to give a final verdict. This is all part of the evidence that we need in order to make a sound frequentist inference. The p-value is simply an indicator as to whether or not the test statistic has landed into the rejection region. Informally speaking, the p-value is the probability of obtaining a test statistic that is at least as extreme as the one that we have observed, i.e., at least as close to the rejection region or at least as deep inside the rejection region.

If we were to schematise a frequentist inference we could say that the conclusions is whether or not to reject H0. The premise is whether or not the test statistic belongs to the critical region, and the background knowledge consists of all the information necessary to establish the appropriate critical region. Both the premise and the background knowledge are part of the evidence.

Of course, this kind of inference is not without shortcomings and there are also other types of statistical inferences. For instance, there are likelihoodist and Bayesian inferences. But in this paper I will not compare frequentist inferences with other types of inferences. I only dismiss some arguments that aim to show that frequentist inferences are internally problematic (without any comparison with other types of inference). Many philosophers have been dismissive of such infer-

ences and yet they misunderstand the basics.

3 The Principle of Total Evidence and P-values

In this section, I show how not to criticise frequentist inferences. I present a paradox and underscore the mistake from which it stems. This groundwork allows for a fruitful assessment of Sober's and Autzen's critical appraisal of p-value based inferences.

3.1 How not to Criticise Frequentist Inferences

Consider two one-sided frequentist tests as defined in Table 1 (See (Wagenmakers 2007, p.782) for a similar example).

Table 1: Two Mass Functions Under H0

distrbution	x=1	x=2	**x=3**	x=4	x=5
f($x|H0$)	0.5	0.46	**0.03**	0.006	0.004
g($x|H0$)	0.5	0.44	**0.03**	0.02	0.01

The first test involves the mass function $f(x)$ under the assumption that a null hypothesis H0 is true, the observed outcome x=3, and a p-value of 0.04. The second test involves the mass function $g(x)$ under the assumption that a H0 is true, the observed outcome x=3, and a p-value of 0.06. Assuming that the significance level for both tests is 0.05, then we will reject H0 in the first case, but not in the second.

Now this looks paradoxical. Surely, if the probability of x=3 is not low enough to reject H0 when we conduct the second test, then it should not be low enough to reject H0 when we do the first one. After all, that probability under H0 is the same for both tests. Yet, we claim to have evidence against H0 in the first case but not in

the second. It certainly looks as if the frequentist approach to testing hypotheses is incoherent.

This kind of paradox, however, rests on the incorrect assumption that the evidence is the same in both scenarios because the probability of the observation under H0 is the same in both. To understand why this is a mistake, we must analyse both inferences more meticulously.

In order to make an inference concerning H0 in both cases, we must first be able to predict a precise and probable range of possible outcomes for a very small probability of error under H0. This allows us to make falsifiable predictions under the assumption that H0 is true.

For the first test, the most precise and probable range of possible outcomes for an upper bound probability of error of 0.05 under H0 can only be x=1 and x=2. When our observation falls outside that range (when it falls inside the so-called critical region), then we reject H0 and we can claim to have evidence that the distribution is inadequately described under H0. That is why we end up rejecting H0 in the first test.

For the second test, the most precise and probable range of possible outcomes for an upper bound probability of error of 0.05 under H0 is x=1, x=2, and x=3. That is why we cannot claim to have evidence against H0 even if the probability of x=3 is also 0.03.

As we can see the evidence is very different in both scenarios. In one case we observe that the outcome of the experiment does not belong to the range of predicted outcomes, whereas it belongs to it in the second case. This is so, even if the probability of the observation is the same under both null hypotheses.

Of course, this difference between the two tests should not come as a surprise because the distributions are different in both cases. Our predictions cannot always be the same when we consider different distributions. That would be absurd.

The take away lesson here is to realise that the premise on which we rely in order to make a frequentist inference in this case is whether or not our observation falls within the most precise and probable range of possible outcomes for a small upper bound probability of error under H0. This is exactly the information that a p-value gives us and that is why it is such a central concept within the frequentist approach.

If we claim that the evidence of a frequentist test is the probability of the test statistic under H0, then we can derive absurd conclusions such as "the evidence is the same when we conduct test 1 and test 2, yet we reject H0 in one test but not the other". Unfortunately, we can still find criticisms of the frequentist approach that rest on the idea that the evidence for a frequentist test is the probability of the test statistic under H0, such as Elliot Sober's criticism of the p-value.

3.2 Sober on the P-value and the principle of Total Evidence

Sober's relatively recent criticism of the p-value can be presented in two simple steps. The first step consists in saying that the frequentist approach to theory testing (FA) dictates that we should reject the hypothesis that we are testing (or claim that we have evidence against it) if our observations are too improbable under that hypothesis.

> [A significance] test has the additional defect that **it violates the principle of total evidence**. In a significance test, the hypothesis you are testing is called the null hypothesis, and your question is whether *the observations are sufficiently improbable* according to the null hypothesis (Sober 2008, p.53, emphasis added).

The second step consists in pointing out that the p-value is not reporting the probability of the observations under the null-hypothesis, but the probability of

obtaining a result within a certain region.

> However, you don't consider the observations in all their detail but rather the fact that they fall in a certain region. You use a logically weaker rather than a logically stronger description of the data. Here's an example [...] that illustrates the point. You want to test the hypothesis that a coin is fair [...] by tossing the coin twenty times. Assume that the tosses are independent of each other. Suppose you obtain four heads. You then compute the probability of a disjunction in which "four heads" is one of the disjuncts.[...] The probability of this disjunction, conditional on the null hypothesis, is called the p-value for the test outcome (Sober 2008, p.53-54).

Therefore, we are not using all the information that is provided by the data when we use the p-value in order to make an inference about H0. Hence, frequentist inferences violate the principal of total evidence.

The problem with this argument obviously lies in the first step. The frequentist theory of statistical inference does not imply that we should reject the hypothesis that we are testing (or that we have evidence against it) if our observations are too improbable under that hypothesis. This is a misconception at the root of the paradox that I have presented in the first section.

The fact is that no matter how you define "improbable" (≤ 0.05, ≤ 0.001, or ≤ 0.0001), we can always construct a probability distribution such that every possible experimental result will be even less probable. It would thus be absurd to suggest that such distributions are inadequate given that our observations are always improbable under such distributions. If that kind of inference was sanctioned by FA, then we would have to reject (or claim that we have evidence against) every possible density function since the probability of an observation under any density

function is 0. This is preposterous.

To make a frequentist inference adequately, we must first be able to make a prediction with the probability distribution that we are testing. As I said before, we usually cannot predict any particular observations because each of them can be very improbable (their probability is actually 0 if we are dealing with a density). But what we can always do with a probability distribution is to establish a likely and an unlikely range of observations according to a predefined degree of probability that we judge to be "too improbable". This will allow us to make a prediction as to where our observations will lie.

For example, suppose that we do not have any knowledge about the nature of the alternative hypothesis and that we agree that 0.0027 is improbable. Then we can predict that an instantiation of random variable that follows a normal distribution will lie within 3 standard deviations from the mean of a normal distribution 99.73% of the time. Therefore, if our actual observation lies outside that range, then we have evidence that the probability distribution that we are testing does not correctly describe the random variable that we are studying.

The idea that we have to be able to make predictions with the hypothesis that we are testing is at the heart of FA. By itself, the probability of the observations is totally meaningless in that context. It is simply not a sufficient piece of evidence for a frequentist inference. Only the probability of falling within a certain range of possible values is important. This is because it is the only type of prediction that we can always make with every possible probability distribution. Therefore, it is false to suggest that the probability of the observations that we make under the hypothesis that we are testing is a logically stronger description of the evidence than the p-value. The p-value determines if our test statistic (our observations) lies within the unlikely range of possible observations that we can make. That is what we need in order to make an inference and that is all we need to make that

inference.[1]

Hence, we can conclude that Sober's criticism is unsound. It rests on the mistaken idea that the probability of the observations that we make is (or should be) sufficient evidence when we are making a frequentist statistical test. It is not.

3.3 Autzen on the p-value and the principle of total evidence

Having thus put aside one of Sober's criticism of the p-value and clarified the nature of classical inferences in statistics, I believe it is also important to correct another mistake about frequentist tests and the nature of the p-value. The latter appears in Autzen's critical assessment the p-value.

Autzen claims that a p-value obtained by performing a one-sided test respects PTE because it is a sufficient statistic. This is not the case for a p-value obtained with a two-sided test. Therefore, we must prefer a one-sided test in order to comply with PTE. [2]

Summing up, I have established a one-to-one function between the

[1] Although the probability of the test statistic given H0 is null when it follows a density, the density evaluated on the test statistic is not. One might then argue that we can at least compare the densities evaluated on the test statistic under H0 and under H1 in order to evaluate the evidence for or against H0 and make an inference accordingly.

This would be another inferential paradigm (likelihoodist) in which the inference does not depend on the need to establish a critical region in order to make a prediction under H0. One simply reaches a conclusion by examining the likelihood function evaluated under H0 and H1. The outcome of the inference does not depend on our capacties to make predictions under H0 or H1 within this framework. In fact, the test statistics could land in an unexpected region under H0 and we will still be able to compare the likelihood ratios and make an inference in favor of H0. Since I am concerned here with defending the internal coherence of frequentist inferences, I shall not elaborate further on this topic.

[2] A sufficient statistic is a function of the observed sample such that the distribution of the sample is independent of the unknown parameter(s) of the distribution given that particular statistic.

> value of the sufficient statistic $\bar{X}$ and the p-value. This implies that the one-sided p-value constitutes a sufficient statistic for the mean of the normal distribution. While Sober (2008, 45) stresses the importance of sufficiency in the context of PTE, he does not mention that for a large class of significance tests the p-value constitutes a sufficient statistic. [...] I conclude that [the p-value] does not violate [PTE] (Autzen 2016, p.289).

But this is false. Inferences based on a two-sided test are not epistemically subpar. In fact, they can be preferable to one-sided tests. We can fail to have enough knowledge in order to use a one-sided test such that a two-sided test and a p-value that is not a sufficient statistic will provide the best inference.

Here is a simple example. Suppose that we have a sample of independent and identically distributed variables $(X_1, X_2, ..., X_n)$. They all follow a normal distribution with a variance equal to 1 and an unknown mean θ. Suppose also that we wish to test whether or not $\theta = 1$. This is our null-hypothesis H0. We decide to use $\bar{X}$ as our test statistic (it is sufficient).

Suppose further that we do not know anything about the alternative hypothesis. The real θ could be greater or less that 1. Yet, we decide to perform a one-sided test as if we knew that the real θ was greater than or equal to 1. Now let's assume that we obtain $\bar{X} = 0.001$. Naturally we will fail to reject H0. We will not consider that we have evidence against H0 because we have conducted our test as if we knew that H1: $\theta < 1$ is not a genuine alternative. But we do not have such knowledge. As such we have made an inadequate inference. We acted as if we had more evidence than what we actually have even if we used a sufficient statistic.

The point is that our knowledge (or lack of knowledge) about the alternative hypothesis is an integral part of the evidence as we make a frequentist test. This is

because it allows us to determine a likely range of possible outcomes under H0 that we can predict for a small probabiliy of error. Look back at the archer's analogy. We need to have an idea of what a good and a bad archer would do in order to identify the relevant regions on the target.

The only way we can make an adequate inference in this case is to conduct a two-sided test and reject H0. Unfortunately, our knowledge about the alternative hypothesis is not considered as evidence by Autzen. As soon as we consider adequately the evidence used in a frequentist test, we realise the importance of two-sided tests.

Autzen misrepresents the epistemological foundations of a frequentist test and fails to recognise the importance of two-sided tests when he claims that "using one-sided tests with a sufficient test statistic is in accordance with PTE" (Autzen 2016, p.292) and that "this supports the view of choosing a one-sided test over a two-sided test"(Autzen 2016, p.292). Two-sided tests are not only used in a wide variety of contexts (just think of the tests being made on the parameters of a regression model), but they are also essential to make sound inferences.

In sum, Autzen's criticism of the p-value and of two-sided tests must be rejected because it fails to adequately define the evidence that is used when we perform a frequentist test. The crucial piece of information that we are looking for is whether or not our test statistic lies within a predicted range of possible outcomes. This piece of evidence rests on our knowledge (or lack of knowledge) of the alternative hypothesis.

Over all, both Sober's and Autzen's criticisms do not provide good reasons to believe that the p-value violates PTE. The reason for this does not reside in their definition of PTE. In the case of Sober's criticism, I have shown that the probability of the test statistic is not and should not be the information on which we rely in order to make a classical test. In the case of Autzen's criticism, I have shown why

we should not always use sufficient statistics when we perform a classical test.

4 The Principle of Total Evidence and the Stopping Rule

In this section, I discuss a similar problem in the sense that it implies the result of classical tests depends on information that should not be a part of the evidence. It is another way of saying that such inferences violate PTE. It is a criticism that we can find in many publications. The interested reader and consult (Howson and Urbach 2006; Romeijn 2017; Wagenmakers 2007; Kadane et al. 1996; Mayo 1996; Robbins 1985; Roberts 1967) for examples.

The structure of this criticism is always the same. Choose two different experiments that yield the exact same observations and where the same parameter is under test. Both experiments will yield different test statistics such that one produces a significant result but not the other. Since the difference between the two experiments is not the evidence but the experimental set up, which is determined by the personal preferences of the experimenter, then we must conclude that classical inferences do not totally rest on evidence.

Here is how Colin Howson and Peter Urbach put it:

> We suggest that such information about experimenters' subjective intentions, their physical strengths and their personal qualities has no inductive relevance whatever in this context, and that in practice it is never sought or even contemplated (Howson and Urbach 2006, p.158).

To understand where this is coming from, consider the following experiments. Both aim at testing whether or not a coin is fair, i.e., whether the probability of landing head is 0.5 or greater. In other words, the parameters that are being tested are the same under both experiments.

In the first experiment, statistician S decides that he will throw the coin six times and record how many heads he obtains. He obtains five heads in a row and then a tail. The fact that S obtained five heads in 6 trial is a statistic that follows a binomial distribution and under the assumption that the probability of success is 0.5 (H0), S cannot reject H0 with a significance level of 0.05 because the p-value is 0.109.

In the second experiment, statistician S* decides that he shall throw the coin six times at most until he obtains a tail. His observations are exactly the same as S but the test statistic this times follows a geometric distribution such that the p-value is 0.016. Thus, S* can reject H0 with a significance level of 0.05. How is this possible? The observations are the same and if the only difference between the two scenarios is how the experimenters subjectively decided to conduct their experiments (choose their sampling plan), then something is seriously wrong with this kind of inference.

Of course, the difference between the sampling plan is not the only difference between the two scenarios. The fact is that both experiments generated two test statistics and both S and S* were in a position to realise this. Both knew about the order of the sequence of heads and tails and both knew about the number of heads and tails.

Unfortunately, both S and S* arbitrarily chose one over the other in order to make their respective inference. They both neglected crucial information concerning the different observations generated by the experiment and their respective probability distribution.

When we are in a position to realise multiple tests in order to make an inference, we cannot simply chose one of those tests without any reason. That would be tantamount to making an arbitrary inference based on arbitrary evidence and I believe that an epistemic principle such as PTE has been invoked in the philosophical

literature in order to rule out such arbitrariness from adequate epistemic practices. In other words, the kind of example mentioned above does not show that classical tests are defective, they stress the importance of an epistemic principle such as PTE.

Now, which of the two inferences mentioned above is the best is a question that we can only answer with a confirmation theory. By determining which inference provides the most justified conclusion, one can settle on one test over the other.

One such theory that is prevalent in the philosophical literature and that applies to frequentist test is the severity measure which asserts that data x_0 provide good evidence for hypothesis H to the extent that test T has severely passed H with x_0 (Mayo and Spanos 2010). Given all the information available in order to make frequentist tests one can then determine which is the best one with a measure of support such as the severity measure. But a full analyses of such theories is beyond the scope of this paper.

In sum, the stopping rule criticism is unsound since the different results involved in that argument can be explained by the fact that both experimenters do not consider all the evidence that should be used to make an adequate inference. In every variation of this criticism there is always an experimenter that ignores the fact that her experiment generated more than one test statistic, i.e., that ignores part of the evidence.

5 Conclusion

The p-value is an essential component of every frequentist inference. In this paper, I have focused on a common criticism of this concept. It is meant to show that its use violates (or can violate) the principle of total evidence. This claim has been made by Sober and Autzen.

Both versions of this criticism rest on a notion of evidence that leads to absurd conclusions. Sober's criticism assumes that the evidence used in a frequentist test is the probability of the test statistic under H0. This is false and it is at the root of the paradox presented in the first section.

Autzen's criticism assumes that the evidence used in a frequentist test should be a sufficient statistic. I have shown that we would make awful inferences if this was a requirement. No competent statistician would make such mistakes. Under minimal information about the alternative hypothesis, we need to use a two-sided test and thus a p-value that is not a sufficient statistic.

The purpose of a frequentist inference is to be able to make a falsifiable prediction by defining a critical region. In order to define such a critical region, we need to rely on our knowledge (or lack of knowledge) about the alternative hypothesis. Ultimately, the p-value informs us about whether or not our observation fell into the critical region.

Once we make those steps clear, both Sober and Autzen's criticisms can be dismissed. The crucial point to remember is that a density or a mass function usually only allows us to predict ranges of possible outcome, as opposed to one single outcome. Therefore, any criticism of a frequentist inference that defines the evidence without any reference to a range of outcomes to which the observation belongs misses the point completely.

In the last part of this paper, I have also debunked the stopping rule criticism, which also implies that frequentist inferences fail PTE. It claims to show that two experimenters can arrive at different conclusions with the same evidence. That criticism is unsound because in all of its variations there is always an experimenter that ignores a test statistic that has been generated by the experiment, i.e., that ignores part of the evidence.

References

Autzen, B. (2016). Significance testing, p-values and the principle of total evidence. *European Journal for Philosophy of Science 6*(2), 281–295.

Howson, C. and P. Urbach (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Kadane, J. B., M. J. Schervish, and T. Seidenfeld (1996). When several bayesians agree that there will be no reasoning to a foregone conclusion. *Philosophy of Science 63*, S281–S289.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. and A. Spanos (2010). *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge University Press.

Neta, R. (2008). What evidence do you have? *The British Journal for the Philosophy of Science 59*(1), 89–119.

Robbins, H. (1985). Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pp. 169–177. Springer.

Roberts, H. V. (1967). Informative stopping rules and inferences about population size. *Journal of the American Statistical Association 62*(319), 763–775.

Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University.

Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review 14*(5), 779–804.