Euro Jnl Phil Sci (2016) 6:281–295
DOI 10.1007/s13194-016-0138-0

ORIGINAL PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Significance testing, p-values and the principle of total evidence

Bengt Autzen[1]

Received: 5 June 2015 / Accepted: 15 February 2016 / Published online: 5 March 2016

Abstract The paper examines the claim that significance testing violates the Principle of Total Evidence (PTE). I argue that p-values violate PTE for two-sided tests but satisfy PTE for one-sided tests invoking a sufficient test statistic independent of the preferred theory of evidence. While the focus of the paper is to evaluate a particular claim about the relationship of significance testing and PTE, I clarify the reading of this methodological principle along the way.

Keywords Significance testing · p-values · Principle of total evidence · Law of likelihood · Evidence · Bayes · Error statistics

1 Introduction

Significance testing is widely used across the natural and social sciences. Given its popularity in scientific practice, it might come as a surprise that significance testing has attracted severe criticism in both the statistical and philosophical literature. For instance, the relationship between significance testing and Bayesian inference as illustrated by Lindley's paradox has led to an ongoing discussion (e.g., Sprenger 2013; Spanos 2013; Robert 2014). Further, the relationship between significance tests and effect size has been subject to criticism (McCloskey and Ziliak 1996; Ziliak and McCloskey 2008). In addition, significance testing has been criticised on the grounds that p-values depend on unobserved data (Wagenmakers 2007) and that their interpretation is problematic (Trafimow 2003). This paper is concerned with an objection

✉ Bengt Autzen
b.autzen@bristol.ac.uk

[1] Department of Philosophy, University of Bristol, Cotham House, Bristol BS6 6JL, UK

made by Sober (2008): the claim that significance testing violates the Principle of Total Evidence (PTE). If significance testing violates an independent and widely accepted methodological principle, then this would constitute a forceful criticism as it does not rely on the prior commitment to a particular statistical methodology.

I will offer a limited defence of significance testing against Sober's objection. My argument proceeds in two steps. First, I will show that the application of PTE requires the prior specification of a criterion for evidential assessment. Second, I will demonstrate that when a plausible criterion for evidential assessment is presupposed, using p-values for inductive inference does not violate PTE for a large and important class of significance tests. In particular, I will argue that p-values violate PTE for two-sided tests but satisfy PTE for one-sided tests with sufficient test statistic from likelihoodist, Bayesian and error-statistical perspectives. Along the way, I will also shed some light on the reading of PTE. Given the importance of significance testing in scientific practice, it should be emphasised that I do not aim to defend the use of p-values *tout court*. Every particular objection against significance testing merits a careful investigation. Here, the focus is on the relationship between significance testing and PTE.

Before turning to Sober's argument, some terminology has to be introduced. Suppose one is interested in the mean adult size of a certain fish species. In order to infer the mean size in this species, one takes measurements of a particular fish population in a pond. The size measurements constitute a random sample $X = (X_1, X_2, ..., X_n)$ of size n. The random variables X_i are assumed to be independent and normally distributed with unknown mean μ and known standard deviation $\sigma = 1$. Now, suppose one would like to test the hypothesis H_0 - referred to as the 'null hypothesis' by statisticians - asserting that the mean μ is equal to, say, 4cm (i.e., $H_0 : \mu = 4$). In order to measure the discrepancy between the parameter value of the mean postulated by the null hypothesis and the sample mean, a test statistic has to be specified. A canonical choice is to use the test statistic $\tau(X) = \sqrt{n}(\bar{X} - \mu_0)/\sigma$, where $\bar{X}$ is the sample mean and μ_0 equals 4. As a result the test statistic $\tau(X)$ follows the standard normal distribution under the null hypothesis. After observing a sample realisation x, a significance tester then calculates the 'p-value', formally defined as $P(\tau(X) \geq \tau(x); H_0 \text{ is true})$ for a one-sided test. That is, the p-value is the probability of observing a sample realisation that would have given rise to a value of the test statistic equal or larger than the one actually observed. While a one-sided test examines only deviations in one direction from the null hypothesis, a two-sided test takes deviations in both directions into account. In the two-sided case the p-value is therefore given by $P(|\tau(X)| \geq |\tau(x)|; H_0 \text{ is true})$.[1]

Having calculated the p-value, the question of what to do next arises. At this stage there are two different approaches to significance testing within the camp of frequentist statistics. One school of thought, tracing back to Fisher (1925), considers the p-value as a measure of the strength of evidence for or against the null hypothesis: the smaller the p-value, the less plausible the null hypothesis. Statisticians in this

[1] In a one-sided test the null hypothesis is sometimes given by $H_0 : \mu \leq \mu_0$. The p-value is then given by $\sup_{\mu<\mu_0} P(\tau(X) \geq \tau(x); H_0)$.

tradition reluctantly specify particular thresholds according to which the data are evidence for (or against) the null hypothesis. Based on some early writings by Fisher, Spanos (1999, 690) offers the following rules of thumb, while maintaining that they can be criticised as ad hoc and unwarranted:[2]

- p-value > 0.1 indicates strong support for H_0
- $0.05 <$ p-value < 0.1 indicates some support for H_0
- $0.01 <$ p-value < 0.05 indicates lack of support for H_0
- p-value < 0.01 indicates strong lack of support for H_0

An alternative approach to significance testing is more closely related to the decision-theoretic framework associated with Neyman and Pearson (1933). Here, a significance test is specified such that the probability of rejecting a true null hypothesis, denoted by α, is fixed at some small number, usually 0.05 or 0.01, which is called the 'significance level' of the test. If the p-value is smaller than α, then the null hypothesis is rejected. Otherwise the null hypothesis is not rejected.[3]

Sober (2008) objects that using p-values for inductive inference violates PTE. When calculating p-values one considers a disjunction of events, in which the actual event is one of the disjuncts and, hence, uses a logically weaker description of the observed data. In Sober's own words:

> Fisher's test of significance [...] 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. 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. (Sober 2008, 53)

While both the evidentialist (or 'Fisherian') and the decision-theoretic approach to significance testing invoke the concept of a p-value, Sober's objection applies in different ways. In the case of the Fisherian approach, Sober's objection applies directly, as the notion of evidential support characterised by Spanos's scheme is based on the p-value. In contrast, Sober's objection applies to the decision-theoretic approach in an indirect way; it requires a principle connecting accept/reject decisions with the notion of evidence. One such principle is given by Sober:[4]

> If learning that e is true justifies you in *rejecting* (i.e., disbelieving) the proposition P, and you were not justified in rejecting P before you gained this

[2]Note that Spanos offers a more sophisticated evidential interpretation of hypothesis testing in his joint work with Mayo (Mayo and Spanos 2006).

[3]The probability of rejecting a true null hypothesis is referred to as a type I error in the Neyman-Pearson framework. Additionally, the Neyman-Pearson theory takes the type II error, that is, the probability of accepting a false null hypothesis, into account. A detailed description of the Neyman-Pearson theory, however, is not required for the purpose of this paper. I also set aside any interpretational issues resulting from hybrid forms of statistical testing combining Fisherian ideas with aspects of the Neyman-Pearson approach (e.g., Mayo 1996).

[4]Notation has been amended for consistency. Italics in original.

> information, then e must be evidence *against* P. If learning that e is true justifies you in *accepting* (i.e., believing) the proposition P, and you were not justified in rejecting P before you gained this information, then e must be evidence *for* P. (Sober 2008, 5)

The details of such a principle are not of concern here. What matters is that rejection needs to be understood as a form of 'evidential rejection' for Sober's objection to apply to the decision-theoretic approach to significance testing.[5]

2 Interpreting PTE

PTE is regularly invoked in philosophical discussions of scientific method. For instance, is has been argued that consensus methods in phylogenetic inference are in conflict with PTE (Barrett et al. 1991). Further, meta-analysis in medicine has been criticised on the grounds that it violates PTE (Stegenga 2011). In order to assess whether significance tests violate PTE, it has to be asked what this principle asserts in the first place. I will approach this question in an iterative manner by refining the interpretation of PTE in a number of steps. Sober (2008, 41) describes PTE as a 'pragmatic' principle, asserting that you should take account of everything you know. The roots of this principle can be traced back to Carnap's inductive logic. Inductive logic aims to assign an objective probability, called 'degree of confirmation', to a hypothesis based on the relationship between hypothesis and evidence. In this context Carnap introduces what he calls the 'requirement of total evidence':

> In the application of inductive logic to a given knowledge situation, the total evidence available must be taken as basis for determining the degree of confirmation. (Carnap 1962, 211)

Synthesizing Sober's and Carnap's remarks, a first interpretation of PTE, denoted as PTE_1, could then read like this: Take into account all available information when making inferences about a hypothesis of interest.

In order to assess the merits of PTE_1, let us return to the fish example introduced earlier. Following PTE_1, one should take into account all available information when making inferences regarding the mean adult fish size. One problem with PTE_1 is that in any real life situation it is unclear what the term 'all available information' amounts to. There is no such a thing as the logically strongest data set.[6] We can always add further attributes to the description of the data set. For instance, we can enrich the description of the data set containing the measurements of the fish population by

[5] For a further discussion, see Sober (2008, 5-7).

[6] Data d_1 are said to be logically stronger than data d_2 if and only if d_1 logically entail d_2. Further, data d_1 are said to be strictly logically stronger than data d_2 if and only if d_1 logically entail d_2 but d_2 do not logically entail d_1. Similarly, data d_1 are said to be logically weaker than data d_2 if and only if d_2 logically entail d_1. And, data d_1 are said to be strictly logically weaker than data d_2 if and only if d_2 logically entail d_1 but not vice versa.

noting whether, say, the fish were difficult to catch, whether it was raining, and whether Chelsea FC played that day.[7]

Given the problems with the notion of a logically strongest data set aiming to capture 'all available information' in an inference situation, an obvious remedy is to formulate PTE in terms of a contrastive principle. The second reading of PTE, denoted as PTE_2, therefore reads as follows: Suppose data d_1 are strictly logically stronger than data d_2, then one should use data d_1 when making inferences about the hypothesis of interest.

While PTE_2 is more satisfactory than PTE_1, it still has consequences that will strike many readers as counterintuitive. In particular, PTE_2 seems to give the false answer to the question of whether we are *always* doing something wrong if we use a logically weaker data set. It seems uncontroversial that PTE only requires using relevant information. So, using a strictly logically weaker data set is unproblematic if the additional information in the logically stronger data set is irrelevant. Sober writes:

> Although the principle of total evidence says that you must use all the *relevant* evidence you have, it does not require the spilling of needless ink. It does not require you to record irrelevant information. (Sober 2008, 44, my italics)

In a similar vein, Carnap (1962, 211) distinguishes between relevant and irrelevant evidence and demands either that an agent knows "nothing beyond [evidence] e or that the totality of his additional knowledge i be irrelevant for [hypothesis] H with respect to e".[8] Both Sober's and Carnap's refinements of PTE point to a third reading, asserting that one should take into account all relevant information when making inferences regarding a hypothesis. Again, it is preferable to phrase PTE in terms of a comparative claim (denoted as PTE_3): Suppose data d_1 are strictly logically stronger than d_2, then one should use data d_1 if the additional information contained in d_1 is relevant for the inference at hand.

PTE_3 naturally raises the question of how to establish whether the strictly logically stronger data are relevant for the inference at hand. Again, the existing literature offers some insights. Suppose data d_1 are strictly logically stronger than data d_2. Carnap's criterion for establishing that d_1 is relevant for hypothesis H given d_2 requires checking whether changing between d_1 and d_2 changes the degree of confirmation of H. Obviously, Carnap's relevance criterion is formulated in terms of his inductive logic. Abstracting from the details of Carnap's account, leads to the following, more general relevance criterion (denoted as RC): data d_1 are relevant for hypothesis H given data d_2 (with $d_1 \Rightarrow d_2$ and $d_2 \not\Rightarrow d_1$) if and only if using d_1 rather than d_2 changes the evidential assessment.[9]

[7] Note that the notion of logical entailment between data sets is agnostic about whether, speaking somewhat loosely, a data set is strengthened by adding more data (e.g., by taking measurements of more fish) or by providing a more fine grained description of the existing data (e.g., by taking more detailed measurements of the same fish).

[8] Notation has been amended.

[9] Note that based on RC relevance is a triadic relation; it defines the relevance of a strictly logically stronger data set for a hypothesis *given* a strictly logically weaker data set.

How can RC be put into practice? I will argue that applying RC presupposes what I will call a 'criterion for evidential assessment' (or 'theory of evidence' for short). Here, a criterion for evidential assessment refers to any account that specifies conditions under which some data d provide evidential support for a hypothesis H. As understood here, a criterion for evidential assessment is generic in character and supposed to capture a variety of philosophical and statistical accounts of evidence. According to the Bayesian theory of evidence, for instance, data d provide evidential support for hypothesis H if and only if the posterior probability of H exceeds the prior probability of H:[10]

Data d are evidence for hypothesis H if and only if $P(H|d) > P(H)$.

Similarly, the law of likelihood (LL) (Hacking 1965) qualifies as a criterion for evidential assessment even though it warrants only contrastive evidential claims. That is, LL establishes conditions under which some data d provide evidential support for one hypothesis H_1 over another hypothesis H_2:

Data d favour hypothesis H_1 over hypothesis H_2 if and only if $P(d|H_1) > P(d|H_2)$.[11]

A third prominent theory of evidence is provided by Mayo (1996). Mayo suggests that data d are evidence for hypothesis H just in case that H passes what she calls a 'severe test' with d. Hypothesis H passes a severe test with d if and only if a) d 'fits' or 'agrees with' H (with some suitable notion of 'fit') and b) there is a low probability that the test would have produced a result that fits H at least as well as d does, if H were false.

Having introduced the notion of a theory of evidence, I am in a position to state my preferred reading of PTE, denoted as PTE_4. The principle reads as follows:

Suppose data d_1 are strictly logically stronger than data d_2, then an inference about hypothesis H should be based on d_1 if changing between d_1 and d_2 changes the evidential assessment.

Alternatively, PTE_4 can be formulated in terms of the notion of relevance captured by RC: Suppose data d_1 are strictly logically stronger than data d_2 with d_1, then an inference about hypothesis H should be based on d_1 if data d_1 are relevant for H given d_2. As discussed, a theory of evidence has to be presupposed in order to apply PTE_4.

The function of PTE_4 can be best illustrated by means of an example. Suppose we evaluate evidential claims within a likelihoodist framework. We observe ten coin tosses. It is assumed that the tosses are independent and each toss follows a Bernoulli

[10] This is sometimes referred to as the 'relative' Bayesian notion of evidence (e.g., Hartmann and Sprenger 2010).

[11] This is typically referred to as the qualitative part of the law of likelihood. The quantitative part asserts that the likelihood ratio $\frac{P(d|H_1)}{P(d|H_2)}$ measures the strength of the evidence. For the purpose of this paper I will focus exclusively on the qualitative part of the law of likelihood. While it is in principle possible that the choice between logically stronger and weaker data does not matter for qualitative questions but does matter for quantitative questions, nothing of import depends on this distinction in the context of the paper.

distribution with parameter p denoting the probability of 'heads'. The hypotheses under consideration are $H_1 : p = 0.5$ and $H_2 : p = 0.6$. We are given the following three description of the observational data:

- $d_1 = (H, H, T, H, T, T, T, H, H, H)$
- $d_2 = 6 \times H, 4 \times T$
- $d_3 = (H, H, T, H, T, T, T, (H \vee T), (H \vee T), (H \vee T))$

That is, data d_1 contain the outcomes of the ten coin tosses in its temporal order, data d_2 only note the frequency of the events 'heads' or 'tails' and data d_3 note the outcomes of the first seven tosses but only tell us that the last three tosses have occurred but not the outcomes of these last three tosses. As a result d_1 strictly logically entails both d_2 and d_3. Since both hypotheses assign probabilities to all three data sets, we do not need to invoke any further assumptions in order to specify the probability measure required for applying LL. Suppose we start with data d_3. According to LL, the data favour hypothesis H_1 over hypothesis H_2 since $P(d_3|H_1) > P(d_3|H_2)$. Does PTE_4 prescribe using the strictly logically stronger data d_1 when making inferences regarding the two hypotheses of interest? The answer is yes, since data d_1 favour hypothesis H_2 over hypothesis H_1 and, hence, change the (qualitative) evidential assessment. Now, suppose we start with data d_2. Data d_2 favour hypothesis H_2 over hypothesis H_1. Hence, the evidential assessment remains unchanged if we move from data d_2 to data d_1. Both data sets favour the same hypothesis. As a result, PTE_4 does not force us to operate on the logically stronger data set in this case.[12]

At this stage one might think about further aspects that should be taken into account when formulating PTE. For instance, I have presumed that the data d_1 and d_2 are freely available and that they can be analysed without any difference in computational cost. These assumptions might not be warranted in a more general discussion of PTE. However, for the purpose of examining Sober's argument against significance testing I set these issues aside.

3 Sober's objection revisited

Having made the case for PTE_4 as an adequate interpretation of PTE, I will now turn to the question of whether significance testing violates PTE as Sober suggests. In order to assess what data set should be used for inductive inference in any particular application, PTE_4 requires the prior specification of a theory of evidence. Without such a specification PTE_4 cannot be applied and, hence, neither be satisfied nor violated. The statistical framework that determines what counts as evidence is therefore primary to PTE. Sober, however, does not explicitly endorse a theory of evidence in his argument. In order to proceed, I will first adopt LL as the theory of evidence, given the central role of LL in Sober's writings (e.g., Sober 2009). PTE_4, however, does not force us to make this choice as the principle is neutral regarding the question of what theory of evidence to adopt in the first place.

[12] See also Sober (2008, 45).

As PTE_4 is concerned with prescribing the choice of data for inductive inference, the question is how this principle can be used to evaluate a statistical technique such as significance testing. A first answer might suggest comparing the data set used by the significance tester with the data set used by the likelihoodist. This suggestion, however, is problematic as both approaches start with the same data set, that is, a realisation of a random sample. So, there is no difference between the significance tester and the likelihoodist in this respect. In order to get Sober's argument off the ground, we have to compare a different pair of data sets. Since Sober's objection is concerned with the use of p-values for inductive inference, we will compare the realisation of the random sample used by the likelihoodist with a 'data' set containing only information about the p-value. In that case it is an open question whether changing between these two data sets affects the evidential assessment by means of LL. I will show that there is no universal conflict between PTE and the use of p-values for inductive inference. While violations do occur, there exists a large and important class of significance tests for which no conflict arises.

As an illustration, let us return to the test of the mean of a normal distribution with known variance (i.e., the 'fish example') introduced earlier. In that case the data are given by the realisation x of the random sample $X = (X_1, X_2, ..., X_n)$, denoted as d_1, and the p-value resulting from this sample realisation, denoted as d_2. My argument proceeds in two steps. In a first step, I will show that the data can be weakened in accordance with PTE_4 by moving from data d_1 to data $\tilde{d_1}$ consisting of a realisation of the sample mean $\bar{x}$. In a second step, I will examine whether the data can be further weakened from data $\tilde{d_1}$ to data d_2. As it will turn out, the second step requires distinguishing between one-sided and two-sided tests.

The first step of modifying the problem by considering the logically weaker data $\tilde{d_1}$ rather than data d_1 is warranted since the sample mean $T(X) = \bar{X}$ is a sufficient statistic for the mean of the normal distribution. Formally, any real-valued function $T = r(X_1, X_2, ..., X_n)$ of the observations in the random sample is called a statistic. A statistic T is a sufficient statistic for parameter θ if for each t, the conditional distribution of $X_1, X_2, ..., X_n$ given $T = t$ and θ does not depend on θ. Speaking informally, a sufficient statistic summarizes all the information in a random sample that is relevant for estimating the parameter of interest. In particular, summarizing the data by means of a sufficient statistic $T(X)$ rather than the random sample X leaves the likelihood ratio within a class of hypotheses - here, hypotheses regarding the mean of the normal distribution - constant (Hacking 1965, 110). Hence, PTE_4 does not demand using the strictly logically stronger data d_1 rather than data $\tilde{d_1}$ when the theory of evidence is provided by LL.[13]

[13]Note that while the statistic $T(X) = \bar{X}$ constitutes a sufficient statistic for the mean of the normal distribution, one cannot assume that a single sufficient statistic exists for any parameter of interest. For some parameters, such as the centre of the Cauchy distribution, no single statistic is sufficient. More systematically, the Pitman-Koopman-Darmois theorem states that under certain regularity assumptions on the probability density, a necessary and sufficient condition for the existence of a sufficient statistic is that the probability density belongs to the exponential family. The exponential family (or Koopman-Darmois family) includes many of the most common probability distributions including the normal, exponential, gamma, beta, Bernoulli and Dirichlet distributions.

Next, we have to evaluate whether using data d_2 rather than data $\tilde{d_1}$ violates PTE$_4$. I will show that there exists a one-to-one function between the p-value and the value of the sufficient statistic $T(X) = \bar{X}$ in the case of the one-sided test but not in the case of the two-sided test. As the one-to-one function of a sufficient statistic is itself sufficient, the one-sided p-value is therefore a sufficient statistic for the mean of the normal distribution.

Let us consider the one-sided test first. Needless to say, there exists a mapping from the value of the sample mean $\bar{X}$ to the p-value $P(\tau(X) \geq \tau(x); H_0 \text{ is true})$ since the test statistic is defined as $\tau(X) = \sqrt{n}(\bar{X} - \mu_0)$. What about the opposite direction? Suppose we are given the p-value $P(\tau(X) \geq \tau(x); H_0 \text{ is true})$ resulting from the realisation of the sample mean $\bar{x}$. As the test statistic $\tau(X)$ follows a standard normal distribution under hypothesis H_0 we can use a standard normal table to infer $\tau(x)$ from the p-value. Based on the definition of the test statistic as $\tau(X) = \sqrt{n}(\bar{X} - \mu_0)$, we can then infer the realisation of the sample mean $\bar{x}$ by simple algebraic transformations. So, there exists a function from the p-value to the value of the sufficient statistic $\bar{X}$.

Summing up, I have established a one-to-one function between the 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. By applying the same reasoning that warranted the use of data $\tilde{d_1}$ rather than data d_1, I conclude that using data d_2 instead of data $\tilde{d_1}$ does not not violate PTE$_4$.

It is worth pointing out that the argument developed here sits well with the result that one-sided p-values can be interpreted as likelihood ratios (DeGroot 1973). DeGroot shows that for a given null hypothesis H_0, a set of alternative hypotheses H_1 can be constructed such that the p-value of a one-sided test is numerically identical with the likelihood ratio of the null hypothesis and the family of alternative hypotheses.[14] At the same time my argument differs from DeGroot's result. I have made no specific assumptions about the alternative hypothesis (or the family of alternative hypotheses) considered in a likelihood evaluation that would warrant drawing conclusions regarding the numerical equivalence between p-values and likelihood ratios. My argument holds for any alternative hypothesis about the mean of the normal distribution. This does not mean, however, that using p-values for inductive inference will yield the same conclusions as inferences by means of LL. In particular, I do not claim that p-values serve as a proxy for likelihood based inferences. Rather, I argue that there is no loss of relevant information when using the information contained in p-values as opposed to the original data set from a likelihoodist perspective.

Returning to the discussion of Sober's objection, matters are different in the case of the two-sided test. Here, the p-value is given by $P(|\tau(X)| \geq |\tau(x)|; H_0 \text{ is true})$.

[14] More formally, suppose that random variable X has probability density function (pdf) f when the null hypothesis H_0 is true. The set of alternative hypotheses H_1 described in terms of the pdf f_θ indexed by the real-valued parameter $\theta \in \Theta$ is then constructed as follows: If f is the pdf of random variable X, then f_θ is the conditional pdf of X given that $X \geq 0$ (DeGroot 1973, 967).

As a result the p-value does not stand in a one-to-one correspondence with the value of the sample mean $\bar{X}$. Speaking graphically, learning about the p-value does not tell us in which of the two tails of the normal distribution the realisation of the sample mean is to be found. Hence, there is no mapping from the two-sided p-value to the value of the sufficient statistic $T(X) = \bar{X}$. Now, it can then be shown that changing between data d_2 and data $\tilde{d}_1$ can lead to conflicting evidential assessments given LL (see Appendix). As a result the use of p-values violates PTE in the two-sided case.

Given that the choice between the one-sided and the two-sided test has implications for the question of whether the use of p-values violates PTE_4, it is natural to ask which of these two is to be employed by statisticians. The two-sided test is typically used to assess whether there is "some effect" in the data if the null hypothesis denotes, say, the absence of a difference between two treatments. However, Casella and Berger (1987, 106) critically remark that given their experience few experimenters are actually interested in the question of whether there is "some difference". Rather, there is a direction of interest in many experiments, such as establishing that "the new treatment is better", which renders the use of a two-sided test inappropriate. While the statistical issue of one-sided versus two-sided testing cannot be resolved in the current paper, it is clear that a one-sided p-value contains information about the direction of the effect, which is lost in the two-sided p-value.[15] So, if the direction of the effect matters to the investigator, there is a prima facie reason for employing a one-sided test. One-sided tests therefore constitute an important class of significance tests.

4 Other theories of evidence

So far, the discussion in this section presupposed LL as the theory of evidence needed to apply PTE_4. In order to complete the discussion of Sober's argument, I will also consider the Bayesian and the error-statistical accounts of evidence. As it turns out, the conclusion will be the same: for the class of one-sided significance tests with sufficient test statistic there is no conflict with PTE while the use of two-sided tests violates PTE.

In order to relate the previous discussion to the analysis of the Bayesian account, the following observation is helpful:[16] Suppose $T = T(X)$ is a sufficient statistic for parameter θ with parameter space Θ equal to an interval of real numbers. Then, for every possible prior prior probability density for θ the posterior probability density of θ given $X = x$ depends on x only through $T(x)$. No matter what prior one uses, one only has to consider the sufficient statistic for Bayesian inference, because the posterior distribution given $T = T(x)$ is the same as the posterior given the data $X = x$. As the p-value of a one-sided test invoking a sufficient statistic can itself be considered as a sufficient statistic, conditioning on a data set containing information about the p-value is the same as conditioning on the data $X = x$. Hence, there is no

[15] For proponents of one-sided testing see, for instance, Kaiser (1960) and Rice and Gaines (1994); for two-sided testing see, for instance, Dubey (1991) and Lombardi and Hurlbert (2009).

[16] See DeGroot and Schervish (2002, 377), exercise 16.

conflict between the use of p-values and PTE_4 for this class of significance tests from a Bayesian perspective.

Again, it is important to stress that this argument differs from DeGroot's (1973) and Casella and Berger's (1987) results that under certain assumptions p-values can be interpreted as posterior probabilities. Analogous to the observation that p-values are numerically identical to likelihood ratios, DeGroot identifies improper priors for which a one-sided p-value and posterior probability match. Similarly, Casella and Berger demonstrate that for many classes of priors there is a close numerical relationship between the posterior probability of the null hypothesis and a one-sided p-value. In contrast, showing that from a Bayesian perspective the use of a one-sided p-value is not in conflict with PTE does not allow any inferences with regard to the numerical equality of p-values and posterior probabilities.

Turning to Mayo's error-statistical account, an important difference to Bayesian and likelihood theories of evidence has to be noted right from the start. As the error statistician does not see a general problem in invoking tail probabilities for inductive inference, the relevant question is what kind of tail probability is suitable for evidential assessment. At the heart of the error-statistical theory is the quantitative measure of severity. In order to illustrate this tail probability, consider the following test scenario. Suppose a random variable is normally distributed with known variance and unknown mean μ_0. Further, suppose one wants to assess the severity with which the hypothesis $H_0 : \mu \leq \mu_0$ passes a test with the realisation of random sample $X = x$ against the alternative $H_1 : \mu > \mu_0$. Again, the test statistic $\tau(X) = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ is employed to measure deviations from H_0 in the direction of the alternative hypothesis H_1. The severity with which H_0 passes the test with data x is then defined as the probability that the test statistic would have taken a larger value if the alternative hypothesis H_1 had been true:

$$SEV(\mu \leq \mu_0)(x, H_1) = P(\tau(X) > \tau(x); \mu > \mu_0).$$

Since the alternative hypothesis H_1 consists of a continuum of point hypotheses it is unclear, however, how to evaluate this probability from a frequentist perspective. Mayo and Spanos (2006) observe that $SEV(\mu \leq \mu_0)(x, H_1)$ is bounded from below by the probability $P(\tau(X) > \tau(x); \mu = \mu_0)$, which is the one-sided p-value of the point null hypothesis $\mu = \mu_0$. As a result there is a close mathematical relationship between severity and one-sided p-values.

In order to assess whether the use of p-values violates PTE from an error-statistical perspective, one has to ask whether changing from data $d_1 = x$ to data d_2 containing information only about the p-value changes the evidential assessment. Again, the difference between one-sided and two-sided p-values is crucial. As the one-sided p-value stands in a one-to-one correspondence with the value of the test statistic $\tau(X)$ (and, hence, test statistic $T(X) = \bar{X}$), using data d_2 rather than d_1 is sufficient for establishing the severity of the test. Once the value of $T(X) = \bar{X}$ is known, one can calculate the severity of this test. Using a one-sided p-value does therefore not violate PTE from an error-statistical perspective. In contrast, the two-sided p-value does not allow to establish the severity of a test, as information about the direction of the effect is lost and the value of test statistic $T(X) = \bar{X}$ cannot be established based on knowledge of the two-sided p-value.

Table 1 Summary of results

	One-sided test (with sufficient test statistic)	Two-sided test
Likelihoodism	PTE satisfied	PTE violated
Bayes	PTE satisfied	PTE violated
Error statistics	PTE satisfied	PTE violated

By highlighting a difference between one-sided and two-sided tests, the error-statistical position mirrors the likelihoodist and Bayesian views on the relationship between PTE and significance testing. All three accounts agree that the use of one-sided p-values with sufficient test statistic is in accordance with PTE while the use of two-sided p-values violates this principle (see Table 1).

This result should not be too surprising since all three accounts of evidence subscribe to the Sufficiency Principle (SP). In order to state SP, the notion of the evidential meaning of an experimental outcome has to be introduced. The 'evidential meaning' of outcome x of experiment E, denoted as $Ev(E, x)$, is supposed to capture the "essential properties" of the statistical evidence provided by the observed outcome x of experiment E (Birnbaum, 1962, 270). The two experiments E (with outcome x) and E' (with outcome y) being 'evidentially equivalent' is denoted by $Ev(E, x) = Ev(E', y)$. SP then reads as follows (Birnbaum 1962):[17]

> If E is a specified experiment, with outcome x; if $T = T(X)$ is any sufficient statistic; and if E' is the experiment derived from E, in which any outcome x of E is represented only by the corresponding value $T(x)$ of the sufficient statistic; then for each x, $Ev(E, x) = Ev(E', T(x))$.

In essence, SP states that the evidential meaning of an observation depends only on the observed value of a sufficient statistic. Since the p-value of a one-sided test with a sufficient test statistic is itself sufficient, all three accounts of evidence agree that this quantity captures the evidential meaning of the observed data. SP is therefore to be seen as a statistical explication of PTE by specifying the conditions under which an evidential assessment should be unaffected when moving to a strictly logically weaker description of the data.

A final word on the question of whether to use a one-sided or a two-sided test. The present discussion suggests a further argument for the use of one-sided p-values. As using one-sided tests with a sufficient test statistic is in accordance with PTE from a variety of perspective of what counts as evidence - including likelihoodist, Bayesian and error-statistical positions - this supports the view of choosing a one-sided over a two-sided test.

[17] Notation has been amended.

5 Conclusion

The paper proposed PTE_4 as an adequate interpretation of PTE. According to PTE_4, strictly logically stronger data should be used if they affect the evidential assessment. Adopting this interpretation of PTE has consequences for assessing the claim that significance testing violates PTE. First, there is no theory-independent assessment of whether significance testing violates PTE. Second, when prominent theories of evidence are presupposed there is no conflict between the use of p-values and PTE for a large and important class of significance tests. Whatever the flaws of p-values and significance tests, violating PTE is not one of them under the premise that a one-sided test with a sufficient test statistic is employed.

Acknowledgments I would like to thank Greg Gandenberger, James Ladyman, Deborah Mayo, Samir Okasha, Jonathan Rougier, Elliott Sober and the anonymous reviewers of the journal for helpful comments on earlier versions of the manuscript. An award from the British Academy Postdoctoral Fellowship Scheme is gratefully acknowledged.

Appendix

Consider the example of the normal distribution with unknown mean and known variance. Suppose that a likelihoodist is, for whatever reason, interested in the two hypotheses $H_0 : \mu = 4$ and $H_1 : \mu = 5$. (Remember that in order to apply LL two candidate hypotheses need to be specified.) Further, suppose that data $\tilde{d}_1$ contain the information that the realisation of the sample mean is equal to 4.51cm (i.e., $\tilde{d}_1 : \bar{x} = 4.51$) based on a sample of, say, $n = 10$ measurements. Data d_2 contain numerical information about the two-sided p-value $P(|\tau(X)| \geq |\tau(x)|; H_0 \text{ is true})$ and, hence, only tell us that the realisation of the sample mean is equal to 3.49cm or to 4.51cm (i.e., $d_2 : \bar{x} = 3.49 \vee \bar{x} = 4.51$). I will show that changing between data $\tilde{d}_1$ and data d_2 leads to conflicting evidential assessments given LL.

Before applying LL, a difficulty has to be dealt with. Since the normal distribution is a continuous probability distribution, the probability of observing the event $\bar{X} = 4.51$ is equal to zero for both candidate hypotheses. Hence, by comparing the probability of the data under the candidate hypotheses, LL does not favour one hypothesis over the other. Instead of comparing the probability of the observation, likelihoodists therefore compare the value of the probability density of the sample mean $\bar{X}$ for each candidate hypothesis (here denoted as $f_{H_0}(t)$ and $f_{H_1}(t)$) at $t = 4.51$. Following this procedure, the data favour H_1 over H_0 based on $\tilde{d}_1$ since $f_{H_1}(4.51) > f_{H_0}(4.51)$.

The next step is to evaluate and compare the likelihoods $P(\bar{X} = 3.49 \vee \bar{X} = 4.51|H_0)$ and $P(\bar{X} = 3.49 \vee \bar{X} = 4.51|H_1)$. Let us focus on $P(\bar{X} = 3.49 \vee \bar{X} = 4.51|H_0)$. This is a somewhat unusual likelihood to evaluate, as likelihoodists typically consider a realisation of the random sample as the data set. So, how to proceed? I assume that the events $\bar{X} = 3.49$ and $\bar{X} = 4.51$ are mutually exclusive (under H_0). Hence, $P(\bar{X} = 3.49 \vee \bar{X} = 4.51|H_0)$ equals $P(\bar{X} = 3.49|H_0) + P(\bar{X} = 4.51|H_0)$. By applying the reasoning from the previous paragraph, the sum of probabilities $P(\bar{X} = 3.49|H_0) + P(\bar{X} = 4.51|H_0)$ is then evaluated by means of the sum $f_{H_0}(3.49) + f_{H_0}(4.51)$. As $f_{H_0}(3.49) + f_{H_0}(4.51) > f_{H_1}(3.49) + f_{H_1}(4.51)$ the data favour H_0 over H_1 based on the strictly logically weaker data d_2. As a result changing between data $\tilde{d_1}$ and d_2 changes the evidential assessment. PTE_4 therefore demands that inferences regarding the mean of the normal distribution are based on the strictly logically stronger data $\tilde{d_1}$. Using only information about the p-value as embodied in d_2 violates PTE_4 in the two-sided case.

References

Barrett, M., Donoghue, M.J., & Sober, E. (1991). Against consensus. *Systematic Zoology*, *40*, 486–493.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*, 269–306.

Carnap, R. (1962). *The logical foundations of probability*. Chicago: Chicago University Press.

Casella, G., & Berger, R. (1987). Reconcilling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, *82*, 106–111.

DeGroot, M.H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, *68*, 966–969.

DeGroot, M.H., & Schervish, M.J. (2002). *Probability and statistics*. Boston: Addison-Wesley.

Dubey, S.D. (1991). Some thoughts on the one-sided and two-sided tests. *Journal of Biopharmaceutical Statistics*, *1*, 139–50.

Fisher, R.A. (1925). *Statistical methods for research workers*. Oliver and Boyd: Edinburgh.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

Hartmann, S., & Sprenger, J. (2010). Bayesian epistemology. In Bernecker, S., & Pritchard, D. (Eds.), *Routledge companion to epistemology*. London: Routledge.

Kaiser, H.F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160–67.

Lombardi, C.M., & Hurlbert, S.H. (2009). Misprecription and misuse of one-tailed tests. *Austral Ecology*, *34*, 447–68.

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

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

McCloskey, D.N., & Ziliak, S.T. (1996). The standard error of regressions. *Journal of Economic Literature*, *34*, 97–114.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, *231*, 289–337.

Rice, W.R., & Gaines, S.D. (1994). 'Heads I win, tails you lose': testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*, *9*, 235–37.

Robert, C. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, *81*, 216–232.

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

Sober, E. (2009). *Did Darwin write the origin backwards? In: Proceedings of the national academy of sciences of the United States of America (Vol. 106, pp. 10048–55)*.

Spanos, A. (1999). *Probability theory and statistical inference*. Cambridge: Cambridge University Press.

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

Sprenger, J. (2013). Testing a precise null hypothesis: the case of Lindley's paradox. *Philosophy of Science*, *80*, 733–744.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*, 497–507.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychological Review*, *110*, 526–535.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Ziliak, S.T., & McCloskey, D.N. (2008). *The cult of statistical significance: how the standard error costs us jobs, justice and lives*. Ann Arbor: University of Michigan Press.