ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys

Alexander Ly *, Josine Verhagen, Eric-Jan Wagenmakers

University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, The Netherlands

HIGHLIGHTS

- Reply to Robert (2016) The expected demise of the Bayes factor.
- Reply to Chandramouli and Shiffrin (2016) Extending Bayesian induction.
- Further elaboration on Jeffreys's Bayes factors.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Bayes factors
Induction
Model selection
Replication
Statistical evidence

ABSTRACT

Our original article provided a relatively detailed summary of Harold Jeffreys's philosophy on statistical hypothesis testing. In response, Robert (2016) maintains that Bayes factors have a number of serious shortcomings. These shortcomings, Robert argues, may be addressed by an alternative approach that conceptualizes model selection as parameter estimation in a mixture model. In a second comment, Chandramouli and Shiffrin (2016) seek to extend Jeffreys's framework by also taking into consideration data distributions that do not originate from either of the models under test. In this rejoinder we argue that Robert's (2016) alternative view on testing has more in common with Jeffreys's Bayes factor than he suggests, as they share the same "shortcomings". On the other hand, we show that the proposition of Chandramouli and Shiffrin (2016) to extend the Bayes factor is in fact further removed from Jeffreys's view on testing than the authors suggest. By elaborating on these points, we hope to clarify our case for Jeffreys's Bayes factors.

In our original article (Ly, Verhagen, & Wagenmakers, 2016) we outlined how Harold Jeffreys constructed his hypothesis tests. Jeffreys's tests contrast a precise, point-null hypothesis $\mathcal{M}_0$ versus a more general alternative hypothesis $\mathcal{M}_1$. Here the point-null hypothesis represents a general law, an invariance, or a categorical causal claim (e.g., "apple trees always bear apples"; "people cannot look into the future"; "Alzheimer's disease is caused by a fungal infection of the central nervous system"), whereas the alternative hypothesis relaxes that law. Jeffreys's tests require a thoughtful specification of the prior distribution for the parameter of interest, and much of Jeffreys's work was concerned with providing good default specifications—"good" in the sense that they adhere to general common-sense desiderata (e.g., Bayarri, Berger, Forte, & García-Donato, 2012). We are pleased that our summary attracted two comments by renowned researchers; below we respond to their ideas in a way that we hope is consistent with the overall philosophy of Harold Jeffreys himself.

1. Rejoinder to Robert

In general, Robert's (2016) comments highlight the inevitable subtleties in constructing a Bayes factor. His alternative mixture model procedure is practical and may be immensely valuable for specific situations (i.e., hierarchical models) that are common in psychological research. Nevertheless, we believe Robert's suggestion about the demise of the Bayes factor to be an overstatement.

1.1. Robert's critique on the Bayes factor

Our understanding of Jeffreys's method is partly based on the work by Robert and colleagues (2009), and it should, therefore, not come as a surprise that Robert's view and ours overlap to a considerable degree. Robert's arguments for dismissing the Bayes factor can be grouped in terms of (1) its usage in making decisions and (2) the care that needs to be taken in choosing the priors.

* Corresponding author.
E-mail address: a.ly@uva.nl (A. Ly).

http://dx.doi.org/10.1016/j.jmp.2016.01.003
0022-2496/

1.1.1. First critique: the distinction between inference and decision making

We share Robert's discontent with the statistical practice that emphasizes all-or-none decisions at some arbitrary threshold, and we agree that scientific learning should instead be guided by a continuous measure of evidence. In the process of eviscerating *p*-value null hypothesis tests, Rozeboom (1960, pp. 422–423) already expressed a similar sentiment:

> "The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true".

Our favorite continuous measure of evidence is of course a Bayes factor constructed from a pair of priors selected according to Jeffreys's desiderata, or a Jeffreys's Bayes factor in short. It is important to note that this measure provides only the first of three Bayesian ingredients needed for decision making. The other two ingredients are the prior model probabilities (which, combined with the Bayes factor, yield posterior model probabilities) and the specification of a loss function (or equivalently, a utility function; Berger, 1985, Lindley, 1977, and Robert, 2007).

For instance, consider a Bayes factor of $\mathrm{BF}_{10}(d) = 4.6$ for the observed data d. This Bayes factor can be converted to a posterior model probability of $P(\mathcal{M}_0 \mid d) = 0.17$ when we set $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$ (Ly et al., 2016). One possible subsequent decision rule is then to accept $P(\mathcal{M}_1 \mid d)$ because it has the highest posterior model probability. We did not intend to suggest such a procedure, as the decision is clearly sensitive to the prior model probabilities. Furthermore, we do not recommend uniform prior model probabilities regardless of scientific context. In fact, when decision making is desired, the assignment of prior model probabilities is left to the substantive researcher. Such flexibility in assignment introduces subjectivity, and this may be seen either as a disadvantage or as an advantage. At any rate, prior model probabilities can be used to formalize the adage that "extraordinary claims require extraordinary evidence" (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Moreover, the prior model probabilities can be used to address the problem of multiplicity (e.g., Jeffreys, 1961; Scott & Berger, 2010; Stephens & Balding, 2009). A similar argument applies to utility functions: these may be subjective and hard to elicit, but such difficulties do not sanction the practice of ignoring utility functions altogether, at least not when the purpose is to make decisions.

Thus, Robert worries that computation of Bayes factors may tempt users to make all-or-none decisions while disregarding prior model probabilities or loss functions. We agree with Robert that there is a considerable difference between inference and decision making, and that scientific learning should be guided by a continuous measure of evidence that incorporates what we have learned from the observed data. The Bayes factor is such a measure.

1.1.2. Second critique: the Jeffreys–Lindley–Bartlett paradox

We suspect that the Jeffreys–Lindley–Bartlett (henceforth JLB) paradox is central to Robert's (1993; 2014) dismissal of the Bayes factor and it is the main motivation for the development of the mixture model alternative. We take a closer look at the JLB paradox and discuss two consequences foreseen by Jeffreys, who was keenly aware of the "paradox" from the very beginning (Etz & Wagenmakers, 2015).

First, the JLB paradox implies that we cannot use improper priors to construct a Bayes factor. For instance, to *estimate* μ within the normal model $\mathcal{M}_1 : X \sim \mathcal{N}(\mu, 1)$, we typically employ Jeffreys's (1946) prior $\mu \propto 1$. The reason to do so stems from the fact that Jeffreys's prior is translation-invariant, leading to a posterior that is independent on how researchers parameterize the problem (Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2015). The JLB paradox implies that we cannot use this same (estimation) prior on the test-relevant parameter for a Bayesian test. More specifically, when we pit the aforementioned model $\mathcal{M}_1$ against the null model $\mathcal{M}_0 : X \sim \mathcal{N}(0, 1)$ the improper prior $\pi_1(\mu) \propto 1$ then becomes useless. To see this we consider the Jeffreys's prior as the limit of proper priors $\mu \sim \mathcal{N}(0, \tau^2)$ with τ tending to infinity. The Bayes factor for the observed data $d = (n, \bar{x})$ is then given by

$$\lim_{\tau\to\infty} \tilde{\mathrm{BF}}_{10\,;\,\tau}(d) = \lim_{\tau\to\infty} \frac{\int \exp\left[-\frac{n}{2}(\bar{x}-\mu)^2\right] \exp\left[-\frac{1}{2\tau^2}\mu^2\right] \mathrm{d}\mu}{\sqrt{2\pi}\,\tau\, \exp[-\frac{n}{2}\bar{x}^2]} = 0, \tag{1}$$

$$= \lim_{\tau\to\infty} \frac{1}{\sqrt{1+n\tau^2}} \exp\left[\frac{(\tau n \bar{x})^2}{2(1+n\tau^2)}\right] = 0, \tag{2}$$

regardless of the fixed sample size n and the observed sample mean $\bar{x}$. As such, the Bayes factor constructed from the improper Jeffreys's prior will always favor the null model and this also holds for other improper priors. Moreover, Eq. (2) shows that for fixed data $d = (n, \bar{x})$ and a Bayes factor constructed from a normal prior with hyperparameter τ we can obtain a Bayes factor in favor of the null hypothesis of arbitrary size (i.e., $\tilde{\mathrm{BF}}_{10\,;\,\tau}(d) < 1$) simply by taking τ large enough.

Hence, the JLB paradox effectively implies that a testing problem should be treated differently from one that is concerned with estimation. As such, when π_1 is interpreted as prior belief about the parameters θ_1, in the example above $\theta_1 = \mu$, one's belief about the parameter then changes depending on whether one is concerned with testing or estimating. More generally, this difference is due to the fact that estimation is typically a *within*-model affair. Recall that a model $\mathcal{M}_i$ specifies a relationship $f_i(d \mid \theta_i)$ that defines which parameters θ_i are relevant in the data generating process of the data d. Hence, the function f_i gives the (only) context in which the parameters θ_i can be perceived.

In essence, the f_i justifies that it is meaningful to calculate a posterior distribution for the parameter. To underline this point we add subscripts to the parameters indicating model membership in the next example, by taking $\theta_0 = \sigma_0$ and $\theta_1 = (\mu_1, \sigma_1)$ for f_0 and f_1 both normals. For example, when we assume that $\mathcal{M}_0 : X \sim \mathcal{N}(0, \sigma_0^2)$ only a posterior for the standard deviation σ_0 is worthwhile to be pursued, as the posterior for the population mean remains zero, regardless of the data. Within $\mathcal{M}_0$, the Jeffreys's prior for σ_0 is given by $\pi_0(\sigma_0) \propto \sigma_0^{-1}$, which can be updated to a posterior $\pi_0(\sigma_0 \mid d)$. On the other hand, under $\mathcal{M}_1 : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ we are dealing with two parameters of interest. Within $\mathcal{M}_1$, the Jeffreys's prior for μ_1 is $\pi(\mu_1) \propto 1$, for σ_1 is $\pi_1(\sigma_1) \propto 1/\sigma_1$ and we take $\pi_1(\mu_1, \sigma_1) = \pi_1(\mu_1)\pi_1(\sigma_1)$. These priors can be updated to posteriors $\pi_1(\mu_1 \mid d)$ and $\pi_1(\sigma_1 \mid d)$. Even though the two priors $\pi_0(\sigma_0)$ and $\pi_1(\sigma_1)$ have the same form, they do not lead to the same posterior. In fact, due to the presence of μ_1 as a parameter, the posterior mean of $\pi_1(\sigma_1 \mid d)$ within $\mathcal{M}_1$ will be smaller or equal to the posterior mean of $\pi_0(\sigma_0 \mid d)$ within $\mathcal{M}_0$. Thus, when we are interested in the standard error σ_i, it matters whether we believe that $\mathcal{M}_0$ holds true or whether the population mean μ_1 plays a role in the data generating process as specified by f_1. The Bayes factor helps us distinguish which of the two models is better suited to the data and which posterior for σ_i we should report. Hence, testing is a *between*-model matter. Jeffreys himself was very clear about the distinction between estimation and testing:

Please cite this article in press as: Ly, A., et al., An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. Journal of Mathematical Psychology (2016), http://dx.doi.org/10.1016/j.jmp.2016.01.003