ARTICLE IN PRESS
Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology

Alexander Ly *, Josine Verhagen, Eric-Jan Wagenmakers

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

HIGHLIGHTS

- The Bayes factor follows logically from Jeffreys's philosophy of model selection.
- The ideas are illustrated with two examples: the Bayesian t-test and correlation test.
- The Bayes factors are adapted to one-sided tests.
- The Bayes factors are illustrated with various applications in psychological research.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Model selection
Bayes factors
Harold Jeffreys

ABSTRACT

Harold Jeffreys pioneered the development of default Bayes factor hypothesis tests for standard statistical problems. Using Jeffreys's Bayes factor hypothesis tests, researchers can grade the decisiveness of the evidence that the data provide for a point null hypothesis $\mathcal{H}_0$ versus a composite alternative hypothesis $\mathcal{H}_1$. Consequently, Jeffreys's tests are of considerable theoretical and practical relevance for empirical researchers in general and for experimental psychologists in particular. To highlight this relevance and to facilitate the interpretation and use of Jeffreys's Bayes factor tests we focus on two common inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian equivalent of the t-test) and testing the nullity of a correlation. For both Bayes factor tests, we explain their development, we extend them to one-sided problems, and we apply them to concrete examples from experimental psychology.

Consider the common scenario where a researcher entertains two competing hypotheses. One, the null hypothesis $\mathcal{H}_0$, is implemented as a statistical model that stipulates the nullity of a parameter of interest (i.e., $\mu = 0$); the other, the alternative hypothesis $\mathcal{H}_1$, is implemented as a statistical model that allows the parameter of interest to differ from zero. How should one quantify the relative support that the observed data provide for $\mathcal{H}_0$ versus $\mathcal{H}_1$? Harold Jeffreys argued that this is done by assigning prior mass to the point null hypothesis (or "general law") $\mathcal{H}_0$, and then calculate the degree to which the data shift one's prior beliefs about the relative plausibility of $\mathcal{H}_0$ versus $\mathcal{H}_1$. The factor by which the data shift one's prior beliefs about the relative plausibility of two competing models is now widely known as the Bayes factor, and it is arguably the gold standard for Bayesian model comparison and hypothesis testing (e.g., Berger, 2006; Lee & Wagenmakers, 2013; Lewis & Raftery, 1997; Myung & Pitt, 1997; O'Hagan & Forster, 2004).

In his brilliant monograph "Theory of Probability", Jeffreys introduced a series of default Bayes factor tests for common statistical scenarios. Despite their considerable theoretical and practical appeal, however, these tests are hardly ever used in experimental psychology and other empirical disciplines. A notable exception concerns Jeffreys's equivalent of the t-test, which has recently been promoted by Jeffrey Rouder, Richard Morey, and colleagues (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009). One of the reasons for the relative obscurity of Jeffreys's default tests may be that a thorough understanding of "Theory of Probability" requires not only an affinity with mathematics but also a willingness to decipher Jeffreys's non-standard notation.

In an attempt to make Jeffreys's default Bayes factor tests accessible to a wider audience we explain the basic principles that drove their development and then focus on two popular inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian t-test) and testing the nullity of a correlation. We illustrate Jeffreys's methodology using data sets from psychological studies. The paper is organized as follows: The first section provides some historical background and outlines four of Jeffreys's convictions regarding scientific learning. The second section shows how the

* Corresponding author.
E-mail address: alexanderly.nl@gmail.com (A. Ly).

http://dx.doi.org/10.1016/j.jmp.2015.06.004

Bayes factor is a natural consequence of these four convictions. We decided to include Jeffreys's own words where appropriate, so as to give the reader an accurate impression of Jeffreys's ideas as well as his compelling style of writing. The third section presents the procedure from which so-called default Bayes factors can be constructed. This procedure is illustrated with the redevelopment of the Bayesian counterpart for the t-test and the Bayesian correlation test. For both the t-test and the correlation test, we also derive one-sided versions of Jeffreys's original tests. We apply the resulting Bayes factors to data sets from psychological studies. The last section concludes with a summary and a discussion.

1. Historical and philosophical background of the Bayes factor

1.1. Life and work

Sir Harold Jeffreys was born in 1891 in County Durham, United Kingdom, and died in 1989 in Cambridge. Jeffreys first earned broad academic recognition in geophysics when he discovered the earth's internal structure (Bolt, 1982; Jeffreys, 1924). In 1946, Jeffreys was awarded the Plumian Chair of Astronomy, a position he held until 1958. After his "retirement" Jeffreys continued his research to complete a record-breaking 75 years of continuous academic service at any Oxbridge college, during which he was awarded medals by the geological, astronomical, meteorological, and statistical communities (Cook, 1990; Huzurbazar, 1991; Lindley, 1991; Swirles, 1991). His mathematical ability is on display in the book "Methods of Mathematical Physics", which he wrote together with his wife (Jeffreys & Jeffreys, 1946).

Our first focus is on the general philosophical framework for induction and statistical inference put forward by Jeffreys in his monographs "Scientific Inference" (Jeffreys, 1931, second edition 1955, third edition 1973) and "Theory of Probability" (henceforth ToP; first edition 1939, second edition 1948, third edition 1961). An extended modern summary of ToP is provided by (Robert, Chopin, & Rousseau, 2009). Jeffreys's ToP rests on a principled philosophy of scientific learning (ToP, Chapter I). In ToP, Jeffreys distinguishes sharply between problems of parameter estimation and problems of hypothesis testing. For estimation problems, Jeffreys outlines his famous transformation-invariant "Jeffreys's priors" (ToP, Chapter III); for testing problems, Jeffreys proposes a series of default Bayes factor tests to grade the support that observed data provide for a point null hypothesis $\mathcal{H}_0$ versus a composite $\mathcal{H}_1$ (ToP, Chapter V). A detailed summary of Jeffreys's contributions to statistics is available online at www.economics.soton.ac.uk/staff/aldrich/jeffreysweb.htm.

For several decades, Jeffreys was one of only few scientists who actively developed, used, and promoted Bayesian methods. In recognition of Jeffreys's persistence in the face of relative isolation, E. T. Jaynes's dedication of his own book, "Probability theory: The logic of science", reads: "Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it" (Jaynes, 2003). In 1980, the seminal work of Jeffreys was celebrated in the 29-chapter book "Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys" (e.g., Geisser, 1980; Good, 1980; Lindley, 1980; Zellner, 1980). In one of its chapters, Dennis Lindley discusses ToP and argues that "The *Theory* is a wonderfully rich book. Open it at almost any page, read carefully, and you will discover some pearl" (Lindley, 1980, p. 37).

Despite discovering the internal structure of the earth and proposing a famous rule for developing transformation-invariant prior distributions, Jeffreys himself considered his greatest scientific achievement to be the development of the Bayesian hypothesis test by means of default Bayes factors (Senn, 2009). In what follows, we explain the rationale behind Jeffreys's Bayes factors and demonstrate their use for two concrete tests.

1.2. Jeffreys's view of scientific learning

Jeffreys developed his Bayes factor hypothesis tests as a natural consequence of his perspective on statistical inference, a philosophy guided by principles and convictions inspired by Karl Pearson's classic book *The Grammar of Science* and by the work of W. E. Johnson and Dorothy Wrinch. Without any claim to completeness or objectivity, here we outline four of Jeffreys's principles and convictions that we find particularly informative and relevant.

1.2.1. Conviction i: Inference is inductive

Jeffreys's first conviction was that scientific progress depends primarily on induction (i.e., learning from experience). For instance, he states "There is a solid mass of belief reached inductively, ranging from common experience and the meanings of words, to some of the most advanced laws of physics, on which there is general agreement among people that have studied the data" (Jeffreys, 1955, p. 276) and, similarly: "When I taste the contents of a jar labelled 'raspberry jam' I expect a definite sensation, inferred from previous instances. When a musical composer scores a bar he expects a definite set of sounds to follow when an orchestra plays it. Such inferences are not deductive, nor indeed are they made with certainty at all, though they are still widely supposed to be" (Jeffreys, 1973, p. 1). The same sentiment is stated more forcefully in ToP: "(...) the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements" (Jeffreys, 1961, p. 5). Hence, inference is inductive and should be guided by the data we observe.

1.2.2. Conviction ii: Induction requires a logic of partial belief

Jeffreys's second conviction is that in order to formalize induction one requires a logic of partial belief: "The idea of a reasonable degree of belief intermediate between proof and disproof is fundamental. It is an extension of ordinary logic, which deals only with the extreme cases" (Jeffreys, 1955, p. 275). This logic of partial belief, Jeffreys showed, needs to obey the rules of probability calculus in order to fulfill general desiderata of consistent reasoning—thus, degrees of belief can be thought of as probabilities (cf. Ramsey, 1926). Hence, all the unknowns should be instantiated as random variables by specifying so-called prior distributions before any datum is collected. Using Bayes' theorem, these priors can then be updated to posteriors conditioned on the data that were actually observed.

1.2.3. Conviction iii: The test of a general law requires it be given prior probability

Jeffreys's third conviction stems from his rejection of treating a testing issue as one of estimation. This is explained clearly and concisely by Jeffreys himself:

> "My chief interest is in significance tests. This goes back to a remark in Pearson's *Grammar of Science* and to a paper of 1918 by C. D. Broad. Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property ϕ, and we do not know r, the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find them all with ϕ, the posterior probability that all n are ϕs is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been inspected. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919–1923. Our point was that giving prior probability $1/(n+$