

1 Harold Jeffreys’s Default Bayes Factor Hypothesis Tests:
2 Explanation, Extension, and Application in Psychology

3 Alexander Ly, Josine Verhagen, & Eric-Jan Wagenmakers
4 University of Amsterdam

5 Abstract

6 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.

Keywords: Model Selection; Bayes factors; Harold Jeffreys.

7 Consider the common scenario where a researcher entertains two competing hypotheses.
8 One, the null hypothesis \mathcal{H}_0 , is implemented as a statistical model that stipulates the nullity of
9 a parameter of interest (i.e., $\mu = 0$); the other, the alternative hypothesis \mathcal{H}_1 , is implemented
10 as a statistical model that allows the parameter of interest to differ from zero. How should
11 one quantify the relative support that the observed data provide for \mathcal{H}_0 versus \mathcal{H}_1 ? Harold
12 Jeffreys argued that this is done by assigning prior mass to the point null hypothesis (or
13 “general law”) \mathcal{H}_0 , and then calculate the degree to which the data shift one’s prior beliefs
14 about the relative plausibility of \mathcal{H}_0 versus \mathcal{H}_1 . The factor by which the data shift one’s

This work was supported by the starting grant “Bayes or Bust” awarded by the European Research Council. Correspondence concerning this article may be addressed to Alexander Ly, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: alexander.ly.nl@gmail.com.

1 prior beliefs about the relative plausibility of two competing models is now widely known as
2 the Bayes factor, and it is arguably the gold standard for Bayesian model comparison and
3 hypothesis testing (e.g., Berger, 2006; Lee & Wagenmakers, 2013; Lewis & Raftery, 1997;
4 Myung & Pitt, 1997; O'Hagan & Forster, 2004).

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

14 In an attempt to make Jeffreys's default Bayes factor tests accessible to a wider audience
15 we explain the basic principles that drove their development and then focus on two popular
16 inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian *t*-test) and
17 testing the nullity of a correlation. We illustrate Jeffreys's methodology using data sets
18 from psychological studies. The paper is organized as follows: The first section provides
19 some historical background and outlines four of Jeffreys's convictions regarding scientific
20 learning. The second section shows how the Bayes factor is a natural consequence of these
21 four convictions. We decided to include Jeffreys's own words where appropriate, so as to give
22 the reader an accurate impression of Jeffreys's ideas as well as his compelling style of writing.
23 The third section presents the procedure from which so-called default Bayes factors can be
24 constructed. This procedure is illustrated with the redevelopment of the Bayesian counterpart
25 for the *t*-test and the Bayesian correlation test. For both the *t*-test and the correlation test,
26 we also derive one-sided versions of Jeffreys's original tests. We apply the resulting Bayes
27 factors to data sets from psychological studies. The last section concludes with a summary
28 and a discussion.

29 **Historical and Philosophical Background of the Bayes Factor**

30 **Life and work**

31 Sir Harold Jeffreys was born in 1891 in County Durham, United Kingdom, and died in
32 1989 in Cambridge. Jeffreys first earned broad academic recognition in geophysics when he
33 discovered the earth's internal structure (Bolt, 1982; Jeffreys, 1924). In 1946, Jeffreys was
34 awarded the Plumian Chair of Astronomy, a position he held until 1958. After his "retirement"
35 Jeffreys continued his research to complete a record-breaking 75 years of continuous academic
36 service at any Oxbridge college, during which he was awarded medals by the geological,
37 astronomical, meteorological, and statistical communities (Cook, 1990; Huzurbazar, 1991;

1 Lindley, 1991; Swirles, 1991). His mathematical ability is on display in the book “Methods
2 of Mathematical Physics”, which he wrote together with his wife (Jeffreys & Jeffreys, 1946).

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

15 For several decades, Jeffreys was one of only few scientists who actively developed,
16 used, and promoted Bayesian methods. In recognition of Jeffreys’s persistence in the face
17 of relative isolation, E. T. Jaynes’s dedication of his own book, “Probability theory: The
18 logic of science”, reads: “Dedicated to the memory of Sir Harold Jeffreys, who saw the truth
19 and preserved it” (Jaynes, 2003). In 1980, the seminal work of Jeffreys was celebrated in
20 the 29-chapter book “Bayesian Analysis in Econometrics and Statistics: Essays in Honor of
21 Harold Jeffreys” (e.g., Geisser, 1980; Good, 1980; Lindley, 1980; Zellner, 1980). In one of its
22 chapters, Dennis Lindley discusses ToP and argues that “The *Theory* is a wonderfully rich
23 book. Open it at almost any page, read carefully, and you will discover some pearl.” (Lindley,
24 1980, p. 37).

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

30 **Jeffreys’s view of scientific learning**

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

36 **Conviction i: Inference is inductive.** Jeffreys’s first conviction was that scientific
37 progress depends primarily on induction (i.e., learning from experience). For instance, he

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

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

23 **Conviction iii: The test of a general law requires it be given prior probabili-**
 24 **ty.** Jeffreys’s third conviction stems from his rejection of treating a testing issue as one of
 25 estimation. This is explained clearly and concisely by Jeffreys himself:

26 “My chief interest is in significance tests. This goes back to a remark in
 27 Pearson’s *Grammar of Science* and to a paper of 1918 by C. D. Broad. Broad
 28 used Laplace’s theory of sampling, which supposes that if we have a population
 29 of n members, r of which may have a property ϕ , and we do not know r , the
 30 prior probability of any particular value of r (0 to n) is $1/(n + 1)$. Broad showed
 31 that on this assessment, if we take a sample of number m and find them all with
 32 ϕ , the posterior probability that all n are ϕ s is $(m + 1)/(n + 1)$. A general rule
 33 would never acquire a high probability until nearly the whole of the class had been
 34 inspected. We could never be reasonably sure that apple trees would always bear
 35 apples (if anything). The result is preposterous, and started the work of Wrinch
 36 and myself in 1919–1923. Our point was that giving prior probability $1/(n + 1)$ to
 37 a general law is that for n large we are already expressing strong confidence that
 38 no general law is true. The way out is obvious. To make it possible to get a high

1 probability for a general law from a finite sample the prior probability must have
 2 at least some positive value independent of n ." Jeffreys (1980, p. 452)

3 The allocation of probability to the null hypothesis is known as the simplicity postulate
 4 (Wrinch & Jeffreys, 1921), that is, the notion that scientific hypotheses can be assigned
 5 prior plausibility in accordance with their complexity, such that "the simpler laws have the
 6 greater prior probabilities" (e.g., Jeffreys, 1961, p. 47; see also Jeffreys, 1973, p. 38). In
 7 the case of testing a point null hypothesis, the simplicity postulate expresses itself through
 8 the recognition that the point null hypothesis represents a general law and, hence, requires a
 9 separate, non-zero prior probability.

10 Jeffreys's view of the null hypothesis as a general law is influenced by his background
 11 in (geo)physics. For instance, Newton's law of gravity postulates the existence of a fixed
 12 universal gravitational constant G . Clearly, this law is more than just a statement about
 13 a constant; it provides a model of motion that relates data to parameters. In this context,
 14 the null hypothesis should be identified with its own separate null model \mathcal{M}_0 rather than be
 15 perceived as a simplified statement \mathcal{H}_0 within the model \mathcal{M}_1 .

16 Hence, Jeffreys's third conviction holds that in order to test the adequacy of a null
 17 hypothesis, the model that instantiates that hypothesis needs to be assigned a separate prior
 18 probability, which can be updated by the data to a posterior probability.

19 **Conviction iv: Classical tests are inadequate.** Jeffreys's fourth conviction was
 20 that classical "Fisherian" p -values are inadequate for the purpose of hypothesis testing. In the
 21 preface to the first edition of ToP, Jeffreys outlines the core problem: "Modern statisticians
 22 have developed extensive mathematical techniques, but for the most part have rejected the
 23 notion of the probability of a hypothesis, and thereby deprived themselves of any way of
 24 saying precisely what they mean when they decide between hypotheses" (Jeffreys, 1961, p.
 25 ix). Specifically, Jeffreys pointed out that the p -value significance test "(...) does not give the
 26 probability of the hypothesis; what it does give is a convenient, though rough, criterion of
 27 whether closer investigation is needed." (Jeffreys, 1973, p. 49). Thus, by selectively focusing
 28 on the adequacy of predictions under the null hypothesis —and by neglecting the adequacy
 29 of predictions under the alternative hypotheses— researchers may reach conclusions that are
 30 premature (see also the Gosset-Berkson critique, Berkson, 1938; Wagenmakers, Verhagen, Ly,
 31 Matzke, et al., in press):

32 "Is it of the slightest use to reject a hypothesis until we have some idea of what
 33 to put in its place? If there is no clearly stated alternative, and the null hypothesis
 34 is rejected, we are simply left without any rule at all, whereas the null hypothesis,
 35 though not satisfactory, may at any rate show some sort of correspondence with
 36 the facts." (Jeffreys, 1961, p. 390).

37 Jeffreys also argued against the logical validity of p -values, famously pointing out that
 38 they depend on more extreme events that have not been observed: "What the use of P implies,

1 therefore, is that a hypothesis that may be true may be rejected because it has not predicted
 2 observable results that have not occurred. This seems a remarkable procedure.” (Jeffreys,
 3 1961, p. 385). In a later paper, Jeffreys clarifies this statement: “I have always considered
 4 the arguments for the use of P absurd. They amount to saying that a hypothesis that may or
 5 may not be true is rejected because a greater departure from the trial value was improbable;
 6 that is, that it has not predicted something that has not happened.” (Jeffreys, 1980, p. 453).

7 In sum, Jeffreys was convinced that induction is an extended form of logic; that this
 8 “logic of partial beliefs” needs to treat degrees of belief as probabilities; that simple laws or
 9 hypotheses should be viewed as separate models that are allocated non-zero prior probabilities,
 10 and that a useful and logically consistent method of hypothesis testing need to be comparative,
 11 and needs to be based on the data at hand rather than on data that were never observed.
 12 These convictions coalesced in Jeffreys’s development of the Bayes factor, an attempt to
 13 provide a consistent method of model selection and hypothesis testing that remedies the
 14 weaknesses and limitations inherent to p -value statistical hypothesis testing.

15 **The Bayes factor hypothesis test**

16 In reverse order, we elaborate on the way in which each of Jeffreys’s convictions moti-
 17 vated the construction of his Bayes factor alternative to the classical hypothesis test.

18 **ad. Conviction iv: Classical tests are inadequate.** Jeffreys’s development of
 19 a Bayesian hypothesis test was motivated in part by his conviction that the use of classical
 20 p values is “absurd”. Nevertheless, Jeffreys reported that the use of Bayes factors generally
 21 yields conclusions similar to those reached by means of p values: “As a matter of fact I have
 22 applied my significance tests to numerous applications that have also been worked out by
 23 Fisher’s, and have not yet found a disagreement in the actual decisions reached” (Jeffreys,
 24 1961, p. 393); thus, “In spite of the difference in principle between my tests and those
 25 based on the P integrals (...) it appears that there is not much difference in the practical
 26 recommendations.” (Jeffreys, 1961). However, Jeffreys was acutely aware of the fact that
 27 disagreements can occur (see also Edwards, Lindman, & Savage, 1963; Lindley, 1957). In
 28 psychology, these disagreements appear to arise repeatedly, especially for cases in which the
 29 p value is in the interval from .01 to .05 (Johnson, 2013; Wetzels et al., 2011).

30 **ad. Conviction iii: The test of a general law requires it be given prior**
 31 **probability.** Jeffreys first identified the null hypothesis with a separate null model \mathcal{M}_0
 32 that represents a general law and pits it against the alternative model \mathcal{M}_1 which relaxes the
 33 restriction imposed by the law. For instance, for the t -test, \mathcal{M}_0 : normal data X with $\mu = 0$
 34 –the law says that the population mean is zero– and \mathcal{M}_1 : normal data X that allows μ to
 35 vary freely. As we do not know whether the data were generated according to \mathcal{M}_0 or \mathcal{M}_1 we
 36 consider the model choice a random variable such that $P(\mathcal{M}_1) + P(\mathcal{M}_0) = 1$.

37 **ad. Conviction ii: Induction requires a logic of partial belief.** As the un-
 38 knowns are considered to be random, we can apply Bayes’ rule to yield posterior model

1 probabilities given the observed data, as follows:

$$P(\mathcal{M}_1 | d) = \frac{p(d | \mathcal{M}_1)P(\mathcal{M}_1)}{P(d)}, \quad (1)$$

$$P(\mathcal{M}_0 | d) = \frac{p(d | \mathcal{M}_0)P(\mathcal{M}_0)}{P(d)}, \quad (2)$$

2 where $p(d | \mathcal{M}_i)$ is known as the marginal likelihood which represents the “likelihood of the
3 data being generated from model \mathcal{M}_i ”. By taking the ratio of the two expressions above, the
4 common term $P(d)$ drops out yielding the key expression:

$$\underbrace{\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_0 | d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}}. \quad (3)$$

5 This equation has three crucial ingredients. First, the prior odds quantifies the relative
6 plausibility of \mathcal{M}_1 over \mathcal{M}_0 before any datum is observed. Most researchers enter experiments
7 with prior knowledge, prior experiences, and prior expectations, and these can in principle
8 be used to determine the prior odds. Jeffreys preferred the assumption that both models are
9 equally likely a priori, such that $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$. This is consistent with the Wrinch-
10 Jeffreys simplicity postulate in the sense that prior mass $1/2$ is assigned to a parsimonious
11 model (e.g., $\mathcal{M}_0 : \mu = 0$, the general law), and the remaining $1/2$ is spread out over a larger
12 model \mathcal{M}_1 where μ is unrestricted. In general then, the prior odds quantify a researcher’s
13 initial skepticism about the hypotheses under test. The second ingredient is the posterior
14 odds, which quantifies the relative plausibility of \mathcal{M}_0 and \mathcal{M}_1 after having observed data d .
15 The third ingredient is the Bayes factor (Jeffreys, 1935): the extent to which data d update
16 the prior odds to the posterior odds. For instance, when $\text{BF}_{10}(d) = 9$, the observed data d
17 are 9 times more likely to have occurred under \mathcal{M}_1 than under \mathcal{M}_0 ; when $\text{BF}_{10}(d) = 0.2$, the
18 observed data d are 5 times more likely to have occurred under \mathcal{M}_0 than under \mathcal{M}_1 . The
19 Bayes factor, thus, quantifies the relative probability of the observed data under each of the
20 two competing hypotheses.

21 Typically, each model \mathcal{M}_i has unknown parameters θ_i that, in accordance to Jeffreys’s
22 second conviction, are considered as random with a density given by $\pi_i(\theta_i)$. By the law
23 of total probability the “likelihood of the data being generated from model \mathcal{M}_i ” is then
24 calculated by integrating out the unknown parameters within that model, that is, $p(d | \mathcal{M}_i) =$
25 $\int f(d | \theta_i, \mathcal{M}_i)\pi_i(\theta_i)d\theta_i$, where $f(d | \theta_i, \mathcal{M}_i)$ is the likelihood, that is, the function that relates
26 the observed data to the unknown parameters θ_i within model \mathcal{M}_i (e.g., Myung, 2003).
27 Hence, when we do not know which of two models ($\mathcal{M}_0, \mathcal{M}_1$) generated the observed data
28 and both models contain unknown parameters, we have to specify two prior densities (π_0, π_1)

1 which formalize our uncertainty before any datum has been observed.

2 **ad. Conviction i: Inference is inductive.** In principle, every pair of priors π_0, π_1
 3 could be used, but Jeffreys specified these two priors based on two desiderata, namely: pre-
 4 dictive matching and information consistency. Predictive matching implies that the Bayes
 5 factor equals 1 when the data are completely uninformative; information consistency implies
 6 that the Bayes factor equals 0 or ∞ when the data are overwhelmingly informative. These
 7 desiderata ensure that the correct inference is reached in extreme cases, and in doing so they
 8 provide useful restrictions for the specification of the prior distributions.

To achieve the desired result that the Bayes factor equals $\text{BF}_{10}(d) = 1$ for completely
 uninformative data, π_0, π_1 need to be chosen such that the marginal likelihoods of \mathcal{M}_0 and
 \mathcal{M}_1 are predictively matched to each other, that is,

$$\int_{\Theta_0} f(d|\theta_0, \mathcal{M}_0)\pi_0(\theta_0)d\theta_0 = p(d|\mathcal{M}_0) = p(d|\mathcal{M}_1) = \int_{\Theta_1} f(d|\theta_1, \mathcal{M}_1)\pi_1(\theta_1)d\theta_1 \quad (4)$$

9 for every completely uninformative data set d .

10 On the other hand, when data d are overwhelmingly informative in favor of the al-
 11 ternative model the Bayes factor should yield $\text{BF}_{10}(d) = \infty$ or, equivalently, $\text{BF}_{01}(d) =$
 12 $1/\text{BF}_{10}(d) = 0$, as this then yields $P(\mathcal{M}_1 | d) = 1$ for any prior model probability $P(\mathcal{M}_1) > 0$.
 13 A Bayes factor with this property is known to be information consistent.

14 **Jeffreys's Procedure for Constructing a Default Bayes Factor**

15 We now elaborate on Jeffreys's general procedure in constructing default Bayes factors
 16 – the specification of the two priors π_0, π_1 – such that the procedure fulfills the desiderata
 17 discussed above.

18 **Step 1. Nest π_0 within π_1**

19 In null hypothesis tests the model \mathcal{M}_1 can be considered an extension of \mathcal{M}_0 by inclusion
 20 of a new parameter, that is, $\theta_1 = (\theta_0, \eta)$ where θ_0 denotes the common parameters and η
 21 denotes the test-relevant parameter. Equivalently, \mathcal{M}_0 is said to be nested within \mathcal{M}_1 due to
 22 the connection $f(d|\theta_0, \mathcal{M}_0) = f(d|\theta_0, \eta = 0, \mathcal{M}_1)$. Jeffreys exploited the connection between
 23 these two likelihood functions to induce a relationship between π_1 and π_0 . In general one has
 24 $\pi_1(\theta_0, \eta) = \pi_1(\eta|\theta_0)\pi_1(\theta_0)$, but due to the nesting Jeffreys treats the common parameters
 25 within \mathcal{M}_1 as in \mathcal{M}_0 , that is, $\pi_1(\theta_0) = \pi_0(\theta_0)$. Furthermore, when η can be sensibly related to
 26 θ_0 , Jeffreys redefines the test-relevant parameter as δ , and decomposes the prior as $\pi_1(\eta, \theta_0) =$
 27 $\pi_1(\delta)\pi_0(\theta_0)$. For instance, in the case of the t -test Jeffreys focuses on effect size $\delta = \frac{\mu}{\sigma}$.

This implies that once we have chosen π_0 , we have then completely specified the
 marginal likelihood $p(d|\mathcal{M}_0)$ and can, therefore, readily calculate the denominator of the
 Bayes factor $\text{BF}_{10}(d)$ given data d . Furthermore, due to the nesting of π_0 within π_1 we can

also calculate a large part of the marginal likelihood of \mathcal{M}_1 as

$$p(d|\mathcal{M}_1) = \int_{\Delta} \underbrace{\int_{\Theta} f(d|\theta_0, \delta, \mathcal{M}_1) \pi_0(\theta_0) d\theta_0}_{h(d|\delta)} \pi_1(\delta) d\delta, \quad (5)$$

1 where $h(d|\delta)$ is the test-relevant likelihood, a function that only depends on the data and the
 2 test-relevant parameter δ as the common parameters θ_0 are integrated out. The following two
 3 steps are concerned with choosing $\pi_1(\delta)$ such that the resulting Bayes factor is well-calibrated
 4 to extreme data.

5 **Step 2. Predictive matching**

6 Typically, a certain minimum number of samples n_{\min} is required before model \mathcal{M}_1 can
 7 be differentiated from \mathcal{M}_0 . All possible data sets with sample sizes less than n_{\min} are consid-
 8 ered uninformative. For example, at least $n_{\min} = 2$ observations are required to distinguish
 9 $\mathcal{M}_0 : \mu = 0$ from \mathcal{M}_1 in a t -test. Specifically, confronted with a single Gaussian observation
 10 unequal to zero, for instance, $x_1 = 5$, lack of knowledge about σ within \mathcal{M}_0 means that we
 11 cannot exclude \mathcal{M}_0 as a reasonable explanation for the data.

Indeed, a member of \mathcal{M}_0 , a zero-mean normal distribution with a standard deviation of seven, produces an observation less than five units away from zero with 53% chance. Similarly, lack of knowledge about σ also means that \mathcal{M}_1 cannot be excluded as a reasonable explanation of the data. To convey that – for the purpose of discriminating \mathcal{M}_0 from \mathcal{M}_1 – nothing is learned from any data set with a sample smaller than n_{\min} we choose $\pi_1(\delta)$ such that

$$p(d|\mathcal{M}_0) = p(d|\mathcal{M}_1) = \int_{\Delta} h(d|\delta) \pi_1(\delta) d\delta \quad (6)$$

12 for every data set d with a sample size less than n_{\min} . In sum, $\pi_1(\delta)$ is chosen such that when
 13 the data are completely uninformative, $\text{BF}_{10}(d) = 1$.

14 **Step 3. Information consistency**

15 Even a limited number of observations may provide overwhelming support for \mathcal{M}_1 . In
 16 the case of the t -test, for instance, the support that an observed non-zero mean provides
 17 for \mathcal{M}_1 should increase without bound when the observed variance, based on any sample
 18 size $n \geq n_{\min}$, goes to zero. Consequently, for data d with a sample size greater or equal
 19 to n_{\min} that point undoubtedly to \mathcal{M}_1 , Jeffreys chose $\pi_1(\delta)$ such that $p(d|\mathcal{M}_1)$ diverges to
 20 infinity. That is, in order to achieve information consistency $p(d|\mathcal{M}_0)$ needs to be bounded
 21 and $\pi_1(\delta)$ needs to be chosen such that $p(d|\mathcal{M}_1) = \int_{\Delta} h(d|\delta) \pi_1(\delta) d\delta$ diverges to infinity for
 22 overwhelmingly informative data of any size n greater or equal to n_{\min} .

1 Summary

2 Jeffreys's procedure to construct a Bayes factor begins with the nesting of π_0 within
 3 π_1 and the choice of π_0 is, therefore, the starting point of the method. The specification of
 4 π_0 yields $p(d|\mathcal{M}_0)$. Next, the test-relevant prior π_1 is chosen such that $p(d|\mathcal{M}_1)$ is well-
 5 calibrated to extreme data that are either completely uninformative or overwhelmingly in-
 6 formative. Together with π_0 , this calibrated test-relevant prior forms the basis for Jeffreys's
 7 construction of a Bayes factor.

8 As a default choice for π_0 , Jeffreys used his popular "Jeffreys's prior" on the common
 9 parameters θ_0 (Jeffreys, 1946). Derived from the likelihood function $f(d|\theta_0, \mathcal{M}_0)$, this default
 10 prior is translation invariant, meaning that the same posterior is obtained regardless of how the
 11 parameters are represented (e.g., Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2015).
 12 Jeffreys's translation-invariant priors are typically improper, that is, non-normalizable, even
 13 though they do lead to proper posteriors for the designs discussed below.

14 The specification of the test-relevant prior requires special care, as priors that are too
 15 wide inevitably reduce the weighted likelihood, resulting in a preference for \mathcal{H}_0 regardless of
 16 the observed data (Jeffreys-Lindley-Bartlett paradox; Bartlett, 1957; Jeffreys, 1961; Lindley,
 17 1957; Marin & Robert, 2010). Consequently, Jeffreys's translation-invariant prior cannot be
 18 used for the test-relevant parameter.

19 Note that Jeffreys's methodical approach in choosing the two priors π_0, π_1 is fully based
 20 on the likelihood functions of the two models that are being compared; the priors do not rep-
 21 resent substantive knowledge of the parameters within the model and the resulting procedure
 22 can therefore be presented as a reference analysis that may be fine-tuned in the presence of
 23 additional information. In the following two sections we illustrate Jeffreys's procedure by
 24 discussing the development of the default Bayes factors for two scenarios that are particularly
 25 relevant for experimental psychology: testing the nullity of a normal mean and the testing
 26 the nullity of a correlation coefficient. Appendix A provides a list of additional Bayes factors
 27 that are presented in ToP.

28 **Jeffreys's Bayes Factor for the Test of the Nullity of a Normal Mean: The** 29 **Bayesian t -test**

30 To develop the Bayesian counterpart of the classical t -test we first characterize the data
 31 and discuss how they relate to the unknown parameters within each model in terms of the
 32 likelihood functions. By studying the likelihood functions we can justify the nesting of π_0
 33 within π_1 and identify data that are completely uninformative and data that are overwhelm-
 34 ingly informative. The test-relevant prior is then selected based on the desiderata discussed
 35 above. We then apply the resulting default Bayes factor to an example data set on cheating
 36 and creativity. In addition, we develop the one-sided extension of Jeffreys's t -test, after which
 37 we conclude with a short discussion.

1 Normal data

2 For the case at hand, experimental outcomes are assumed to follow a normal distribution
 3 characterized by the unknown population mean μ and standard deviation σ . Similarly, the
 4 observed data d from a normal distribution can be summarized by two numbers: the observed
 5 sample mean \bar{x} and the average sums of squares $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$; hence we write
 6 $d = (\bar{x}, s_n^2)$. The main difference between the null model $\mathcal{M}_0 : \mu = 0$ and its relaxation \mathcal{M}_1
 7 is reflected in the population effect size, which is defined as $\delta = \frac{\mu}{\sigma}$, as σ provides a scale
 8 to the problem. This population effect size cannot be observed directly, unlike its sampled
 9 scaled version the t -statistic, i.e., $t = \frac{\sqrt{n}\bar{x}}{s_\nu}$, where s_ν refers to the sample standard deviation
 10 based on $\nu = n - 1$ degrees of freedom. Extreme data can be characterized by $|t| \rightarrow \infty$ or
 11 equivalently by $s_n^2 \rightarrow 0$ and it is used in the calibration step of the Bayes factor to derive the
 12 test-relevant prior. To improve readability we remove the subscript n when we refer to the
 13 average sum of squares $s^2 = s_n^2$.

14 Step 1. Nesting of π_0 within π_1

Comparing the likelihood functions. A model defines a likelihood that struc-
 turally relates how the observed data are linked to the unknown parameters. The point null
 hypothesis \mathcal{M}_0 posits that $\mu = 0$, whereas the alternative hypothesis \mathcal{M}_1 relaxes the restric-
 tion on μ . Conditioned on the observations $d = (\bar{x}, s^2)$, the likelihood functions of \mathcal{M}_0 and
 \mathcal{M}_1 are given by

$$f(d | \sigma, \mathcal{M}_0) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2} [\bar{x}^2 + s^2]\right), \quad (7)$$

$$f(d | \mu, \sigma, \mathcal{M}_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2} [(\bar{x} - \mu)^2 + s^2]\right), \quad (8)$$

respectively. Note that $f(d | \sigma, \mathcal{M}_0)$ is a function of σ alone, whereas $f(d | \mu, \sigma, \mathcal{M}_1)$ depends
 on two parameters, σ and μ . By the nesting we can set $\pi_1(\mu, \sigma) = \pi_1(\mu | \sigma)\pi_0(\sigma)$. Jeffreys
 removed the scale from the problem by considering $\delta = \frac{\mu}{\sigma}$ as the test-relevant parameter
 which leads to $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$. The marginal likelihood of \mathcal{M}_1 is then given by

$$p(d | \mathcal{M}_1) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \int_{-\infty}^\infty \exp\left(-\frac{n}{2} \left[\left(\frac{\bar{x}}{\sigma} - \delta\right)^2 + \left(\frac{s}{\sigma}\right)^2\right]\right) \pi_1(\delta) d\delta \pi_0(\sigma) d\sigma, \quad (9)$$

15 which defines the numerator of the Bayes factor $\text{BF}_{10}(d)$ after $\pi_0(\sigma)$ and $\pi_1(\delta)$ are chosen.

16 **The denominator of $\text{BF}_{10}(d)$.** Jeffreys's default choice leads to $\pi_0(\sigma) \propto 1/\sigma$, the
 17 translation-invariant prior that Jeffreys's would use to arrive at a posterior for σ within either
 18 model. This prior specification leads to the following marginal likelihood of \mathcal{M}_0 :

$$p(d|\mathcal{M}_0) = \begin{cases} \frac{1}{2|\bar{x}|} & n = 1, & (10) \\ \frac{\Gamma(\frac{n}{2})}{2(\pi n \bar{x}^2)^{\frac{n}{2}}} & n > 1 \text{ and } s^2 = 0, & (11) \\ \frac{\Gamma(\frac{n}{2})}{2(\pi n s^2)^{\frac{n}{2}}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} & n > 1 \text{ and } s^2 > 0, & (12) \end{cases}$$

1 where t is the observed t -value and ν the degrees of freedom defined as before. Hence, Eqn. (10,
 2 11, 12) define the denominator of the Bayes factor $\text{BF}_{10}(d)$; Eq. (10) will be used to calibrate
 3 the Bayes factor $\text{BF}_{10}(d)$ to completely uninformative data, whereas Eq. (11) will be used for
 4 the calibration to overwhelmingly informative data. Some statisticians only report the right
 5 term $\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ of Eq. (12), as the first term also appears in the marginal likelihood of
 6 \mathcal{M}_1 and, thus, cancels out in the Bayes factor.

7 **Step 2. Predictive matching: Symmetric** $\pi_1(\delta)$

8 We now discuss how the test-relevant prior $\pi_1(\delta)$ can be chosen such that the resulting
 9 Bayes factor is well-calibrated. As elaborated above, we consider data sets with only one
 10 sample as completely uninformative in discriminating \mathcal{M}_0 from \mathcal{M}_1 . Jeffreys (1961, p. 269)
 11 studied Eq. (9) with $n = 1$, $\bar{x} > 0$, and, consequently, $s^2 = 0$, and concluded that $p(d|\mathcal{M}_1)$
 12 is matched to Eq. (10) whenever $\pi_1(\delta)$ is symmetric around zero.

13 **Step 3. Information consistency: Heavy-tailed** $\pi_1(\delta)$

14 On the other hand, observed data $\bar{x} > 0$, $s^2 = 0$ with $n > 1$ can be considered over-
 15 whelmingly informative as the t -value is then infinite. To obtain maximum evidence in favor
 16 of the alternative we require that $\text{BF}_{10}(d) = \infty$. This occurs whenever the marginal likelihood
 17 of \mathcal{M}_1 is infinite and $p(d|\mathcal{M}_0)$ finite, see Eq. (11). Jeffreys (1961, p. 269 – 270) showed that
 18 this is the case whenever the test-relevant prior $\pi_1(\delta)$ is heavy-tailed.

19 **The resulting Bayes factor**

20 Hence, a Bayes factor that meets Jeffreys's desiderata can be obtained by assigning
 21 $\pi_0(\sigma) \propto 1/\sigma$ and $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$, where $\pi_1(\delta)$ is symmetric around zero and heavy-
 22 tailed.

23 **Jeffreys's choice: The standard Cauchy distribution.** The Cauchy distribution
 24 with scale γ is the most well-known distribution which is both symmetric around zero and
 25 heavy-tailed:

$$\pi_1(\delta; \gamma) = \frac{1}{\pi\gamma \left(1 + \left(\frac{\delta}{\gamma}\right)^2\right)}. \quad (13)$$

1 As a default choice for $\pi_1(\delta)$, Jeffreys suggested to use the simplest version, the standard
2 Cauchy distribution with $\gamma = 1$.

3 **Jeffreys's Bayesian t -test.** Jeffreys's Bayes factor now follows from the integral in
4 Eq. (9) with respect to Cauchy distributions $\pi_1(\delta)$ divided by Eq. (12), whenever $n > 1$ and
5 $s^2 > 0$. Jeffreys knew that this integral is hard to compute and went to great lengths to
6 compute an approximation that makes his Bayesian t -test usable in practice. Fortunately,
7 we can now take advantage of computer software that can numerically solve the aforementioned
8 integral and we therefore omit Jeffreys's approximation from further discussion. By a
9 decomposition of a Cauchy distribution we obtain a Bayes factor of the following form:

$$BF_{10;\gamma}(n, t) = \frac{\gamma \int_0^\infty (1 + ng)^{-\frac{1}{2}} \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} (2\pi)^{-\frac{1}{2}} g^{-\frac{3}{2}} e^{-\frac{\gamma^2}{2g}} dg}{\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}}, \quad (14)$$

10 where g is an auxiliary variable that is integrated out numerically. Jeffreys's choice is obtained
11 when $\gamma = 1$. The Bayes factor $BF_{10;\gamma=1}(n, t)$ now awaits a user's observed t -value and the
12 associated n number of observations.

13 **Example: The Bayesian between-subject t -test**

14 To illustrate the default Bayesian t -test we extend Eq. (14) to a between-subjects design
15 and apply the test to a psychological data set. The development above is easily generalized to
16 a between-subject design in which observations are assumed to be drawn from two separate
17 normal populations. To do so, we replace: (i) the value of t by the two-sample (grouped)
18 t -value,¹ (ii) the effective sample size $n = \frac{n_1 n_2}{n_1 + n_2}$, and (iii) the degrees of freedom with $\nu =$
19 $n_1 + n_2 - 2$, see Rouder et al. (2009).

20 **Example 1** (Does cheating enhance creativity?). *Gino and Wiltermuth (2014, Experiment 2)*
21 *reported that the act of cheating enhances creativity. This conclusion was based on five ex-*
22 *periments. Here we analyze the results from Experiment 2 in which, having been assigned*
23 *either to a control condition or to a condition in which they were likely to cheat, participants*
24 *were rewarded for correctly solving each of 20 math and logic multiple-choice problems. Next,*

¹The two-sample t -value is defined as $t = \sqrt{n_1 n_2} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{n_2 s_{n_1-1}^2 + n_1 s_{n_2-1}^2}}$, where n_i , \bar{x}_i and s_{n_i-1} are the sample size, sample mean, and (unbiased) standard deviation of group i , respectively.

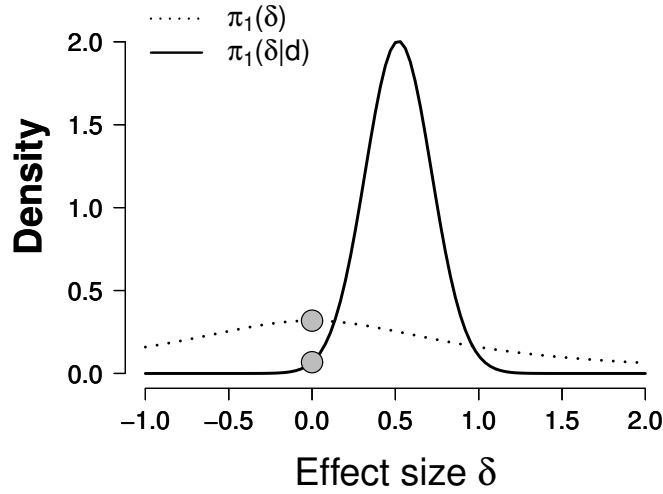


Figure 1. Posterior and prior distributions of the effect size for a two-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $BF_{10;\gamma=1} \approx 4.60$ equals the height ratio of the prior distribution $\pi_1(\delta)$ over the posterior distribution $\pi_1(\delta | d)$ at $\delta = 0$.

1 participants' creativity levels were measured by having them complete 12 problems from the
 2 Remote Association Task (RAT; Mednick, 1962).

3 The control group featured $n_1 = 48$ participants who scored an average of $\bar{x}_1 = 4.65$ RAT
 4 items correctly with a sample standard deviation of $s_{n_1-1} = 2.72$. The cheating group featured
 5 $n_2 = 51$ participants who scored $\bar{x}_2 = 6.20$ RAT items correctly with $s_{n_2-1} = 2.98$. These
 6 findings yield $t(97) = 2.73$ with $p = .008$. Jeffreys's default Bayes factor yields $BF_{10}(d) \approx 4.6$,
 7 indicating that the data are 4.6 times more likely under \mathcal{M}_1 than under \mathcal{M}_0 . With equal prior
 8 odds, the posterior probability for \mathcal{M}_0 remains an arguably non-negligible 17%.

9 For nested models, the Bayes factor can be obtained without explicit integration, using
 10 the Savage-Dickey density ratio test (e.g., Dickey & Lientz, 1970; Wagenmakers, Lodewyckx,
 11 Kuriyal, & Grasman, 2010; Marin & Robert, 2010). The Savage-Dickey test is based on the
 12 following identify:

$$13 \quad BF_{10}(d) = \frac{\pi_1(\delta = 0)}{\pi_1(\delta = 0 | d)}. \quad (15)$$

14 One of the additional advantages of the Savage-Dickey test is that it allows the result of the
 15 test to be displayed visually, as the height of the prior versus the posterior at the point of test
 16 (i.e., $\delta = 0$). Fig. 1 presents the results from Experiment 2 of Gino and Wiltermuth (2014).

17 \diamond

18 In this example, both the Bayesian and Fisherian analysis gave the same qualitative
 19 result. Nevertheless, the Bayes factor is more conservative, and some researchers may be
 20 surprised that, for the same data, $p = .008$ and posterior model probability $P(\mathcal{M}_0 | d) = .17$.

1 Indeed, for many cases the Bayesian and Fisherian analyses disagree qualitatively as well as
 2 quantitatively (e.g., Wetzels et al., 2011).

3 **The one-sided extension of Jeffreys's Bayes factor**

4 Some reflection suggests that the authors' hypothesis from Example 1 is more specific
 5 – the authors argued that cheating leads to more creativity, not less. To take into account
 6 the directionality of the hypothesis we need a one-sided adaptation of Jeffreys's Bayes factor
 7 $\text{BF}_{10; \gamma=1}(n, t)$. The comparison that is made is then between the model of no effect \mathcal{M}_0
 8 and one denoted by \mathcal{M}_+ in which the effect size δ is assumed to be positive. We decompose
 9 $\text{BF}_{+0}(d)$ as follows:

$$\text{BF}_{+0}(d) = \frac{p(d | \mathcal{M}_+) p(d | \mathcal{M}_1)}{\underbrace{p(d | \mathcal{M}_1)}_{\text{BF}_{+1}(d)} \underbrace{p(d | \mathcal{M}_0)}_{\text{BF}_{10}(d)}}, \quad (16)$$

10 where $\text{BF}_{+1}(d)$ is the Bayes factor that compares the unconstrained model \mathcal{M}_1 to the pos-
 11 itively restricted model \mathcal{M}_+ (Morey & Wagenmakers, 2014; Mulder, Hoijtink, & Klugkist,
 12 2010; Pericchi, Liu, & Torres, 2008). The objective comparison between \mathcal{M}_+ and \mathcal{M}_1 is then
 13 to keep all aspects the same: $\pi_+(\sigma) = \pi_1(\sigma) = \pi_0(\sigma)$ except for the distinguishing factor of
 14 δ being restricted to positive values within \mathcal{M}_+ . For the test-relevant prior distribution we
 15 restrict $\pi_1(\delta)$ to positive values of δ , which by symmetry of the Cauchy distribution means
 16 that $\pi_+(\delta)$ accounts doubly for the likelihood when δ is positive and nullifies it when δ is
 17 negative (Klugkist, Laudy, & Hoijtink, 2005).

18 **Example 1** (One-Sided Test for the Gino and Wiltermuth Data, Continued). *For the data*
 19 *from Gino and Wiltermuth (2014, Experiment 2) the one-sided adaptation of Jeffreys's Bayes*
 20 *factor Eq. (14) yields $\text{BF}_{+0}(d) = 9.18$. Because almost all of the posterior mass is consistent*
 21 *with the authors' hypothesis, the one-sided Bayes factor is almost twice the two-sided Bayes*
 22 *factor. The result is visualized through the Savage-Dickey ratio in Fig. 2. \diamond*

23 **Discussion on the t -test**

24 In this section we showcased Jeffreys's procedure in selecting the instrumental priors
 25 π_0, π_1 that yield a Bayes factor for grading the support that the data provide for \mathcal{M}_0 versus
 26 \mathcal{M}_1 . The construction of this Bayes factor began by assigning Jeffreys's translation-invariant
 27 prior to the common parameters, that is, $\pi_0(\sigma) \propto 1/\sigma$. This is the same prior Jeffreys would
 28 use for estimating σ in either of the two models, when one of these models is assumed to
 29 be true. This prior on the common parameters then yields the denominator of the Bayes
 30 factor Eqn. (10, 11, 12). Jeffreys noted that when the test-relevant prior $\pi_1(\delta)$ is symmetric

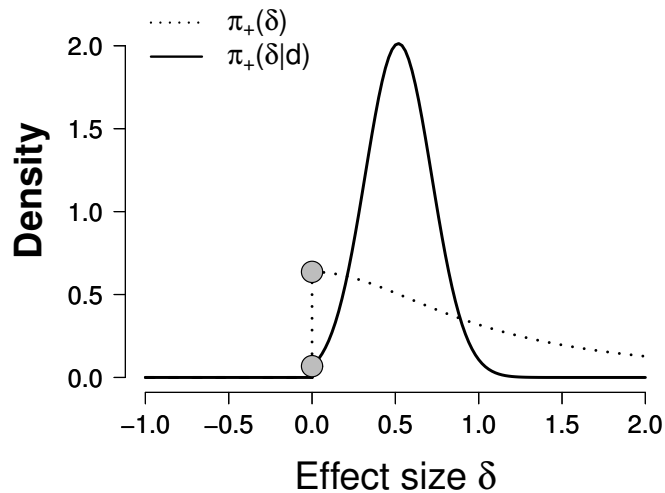


Figure 2. Posterior and prior distributions of effect size for a one-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $BF_{+0} = 9.18$ equals the height ratio of the prior distribution $\pi_1(\delta)$ over the posterior distribution $\pi_1(\delta | d)$ at $\delta = 0$. The prior distribution $\pi_+(\delta)$ is zero for negative values of δ . Furthermore, note that the prior distribution for $\delta \geq 0$ is twice as high compared to $\pi_1(\delta)$ in Fig. 1.

1 and heavy tailed, the resulting Bayes factor is guaranteed to yield the correct conclusion for
 2 completely uninformative data and for overwhelmingly informative data. Jeffreys (1961, p.
 3 272 – 273) noted that the standard Cauchy prior for δ yields a Bayes factor Eq. (14) (with
 4 $\gamma = 1$) that aligns with this calibration.

5 It took several decades before Jeffreys's Bayes factor for the t -test was adopted by
 6 Zellner and Siow (1980) who generalized it to the linear regression framework based on a
 7 multivariate Cauchy distribution. One practical drawback of their proposal was the fact that
 8 the numerical integration required to calculate the Bayes factor becomes computationally
 9 demanding as the number of covariates grows.

10 Liang, Paulo, Molina, Clyde, and Berger (2008) proposed a computationally efficient
 11 alternative to the Zellner and Siow (1980) setup by first decomposing the multivariate Cauchy
 12 distribution into a mixture of gamma and normal distributions followed by computational
 13 simplifications introduced by Zellner (1986). As a result, the Bayes factor can be obtained
 14 from only a single numerical integral, regardless of the number of covariates. The form of the
 15 numerator in Eq. (14) is in fact inspired by Liang et al. (2008) and introduced to psychology
 16 by Rouder et al. (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009). The
 17 combination $\pi_0(\sigma) \propto \sigma^{-1}$ and $\delta \sim \mathcal{C}(0, 1)$ was dubbed the JZS-prior in honor of Jeffreys,
 18 Zellner, and Siow; this is understandable in the framework of linear regression, although it
 19 should be noted that all ideas for the t -test were already present in the second edition of ToP

1 (Jeffreys, 1948, pp. 242–248).

2 **Model selection consistency.** In addition to predictive matching and information
 3 consistency, Liang et al. (2008) showed that Zellner and Siow's (1980) generalization of Jef-
 4 freys's work is also model selection consistent, which implies that as the sample size n increases
 5 indefinitely, the support that the data d provide for the correct data-generating model (i.e.,
 6 \mathcal{M}_0 or \mathcal{M}_1) grows without bound. Hence, Jeffreys's default Bayes factor Eq. (14) leads to
 7 the correct decision whenever the sample size is sufficiently large. Jeffreys's procedure of
 8 assigning default priors for Bayesian hypothesis testing was recently generalized by Bayarri,
 9 Berger, Forte, and García-Donato (2012). We now turn to Jeffreys's development of another
 10 default Bayes factor: the test for the presence of a correlation.

11 Jeffreys's Bayes Factor for the Test of the Nullity of a Correlation

12 To develop the Bayesian correlation test we first characterize the data and discuss how
 13 they relate to the unknown parameters within each model in terms of the likelihood functions.
 14 By studying the likelihood functions we can justify the nesting of π_0 within π_1 and identify
 15 data that are completely uninformative and data that are overwhelmingly informative. As
 16 was done for the Bayesian t -test, the test-relevant prior is selected based on a calibration
 17 argument. The derivations and calibrations given here cannot be found in Jeffreys (1961), as
 18 Jeffreys appears to have derived the priors intuitively. Hence, we divert from the narrative of
 19 Jeffreys (1961, Paragraph 5.5) and instead: (a) explain Jeffreys's reasoning with a structure
 20 analogous to that of the previous section; and (b) give the exact results instead, as Jeffreys
 21 used an approximation to simplify the calculations. In effect, we show that Jeffreys's intuitive
 22 choice is very close to our exact result. After presenting the correlation Bayes factor we relate
 23 it to Jeffreys's choice and apply the resulting default Bayes factor to an example data set
 24 that is concerned with presidential height and the popular vote. In addition, we develop
 25 the one-sided extension of Jeffreys's correlation test, after which we conclude with a short
 26 discussion.

27 Bivariate normal data

28 For the case at hand, experimental outcome pairs (X, Y) are assumed to follow a bi-
 29 variate normal distribution characterized by the unknown population means μ_x, μ_y , stan-
 30 dard deviations σ, ν of X and Y respectively. Within \mathcal{M}_1 the parameter ρ characterizes
 31 the linear association between X and Y . To test the nullity of the population correla-
 32 tion it is helpful to summarize the data for X and Y separately in terms of their respec-
 33 tive sample means and average sums of squares: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ and
 34 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $u^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$, respectively. The sample correlation coefficient r then
 35 defines an observable measure of the linear relationship between X and Y :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{-n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i}{nsu}. \quad (17)$$

1 This sample correlation coefficient r is an imperfect reflection of the unobservable pop-
 2 ulation correlation coefficient ρ . The data can be summarized by the five quantities
 3 $d = (\bar{x}, s^2, \bar{y}, u^2, r)$.

4 The main difference between the null model \mathcal{M}_0 and \mathcal{M}_1 is reflected in the population
 5 correlation coefficient ρ , which cannot be observed directly, unlike its sampled version known
 6 as Pearson's r , Eq. (17). Extreme data can be characterized by $|r| = 1$ and this is used in the
 7 calibration step of the Bayes factor to derive the form of the test-relevant prior.

8 Step 1. Nesting of π_0 within π_1

9 **Comparing the likelihood functions.** The point null hypothesis \mathcal{M}_0 assumes that
 10 the data follow a bivariate normal distribution with ρ known and fixed at zero. Hence, \mathcal{M}_0
 11 depends on four parameters which we abbreviate as $\theta_0 = (\mu_x, \mu_y, \sigma, \nu)$, while the alternative
 12 model \mathcal{M}_1 can be considered an extension of \mathcal{M}_0 with an additional parameter ρ , i.e., $\theta_1 =$
 13 (θ_0, ρ) . These two bivariate normal models relate the observed data to the parameters using
 14 the following two likelihood functions:

$$f(d | \theta_0, \mathcal{M}_0) = (2\pi\sigma\nu)^{-n} \exp\left(-\frac{n}{2}\left[\left(\frac{\bar{x}-\mu_x}{\sigma}\right)^2 + \left(\frac{\bar{y}-\mu_y}{\nu}\right)^2\right]\right) \\ \times \exp\left(-\frac{n}{2}\left[\left(\frac{s}{\sigma}\right)^2 + \left(\frac{u}{\nu}\right)^2\right]\right). \quad (18)$$

$$f(d | \theta_1, \mathcal{M}_1) = (2\pi\sigma\nu\sqrt{1-\rho^2})^{-n} \exp\left(-\frac{n}{2(1-\rho^2)}\left[\frac{(\bar{x}-\mu_x)^2}{\sigma^2} - 2\rho\frac{(\bar{x}-\mu_x)(\bar{y}-\mu_y)}{\sigma\nu} + \frac{(\bar{y}-\mu_y)^2}{\nu^2}\right]\right) \\ \times \exp\left(-\frac{n}{2(1-\rho^2)}\left[\left(\frac{s}{\sigma}\right)^2 - 2\rho\left(\frac{rsu}{\sigma\nu}\right) + \left(\frac{u}{\nu}\right)^2\right]\right). \quad (19)$$

Note that $f(d | \theta_0, \mathcal{M}_0) = f(d | \theta_0, \rho = 0, \mathcal{M}_1)$ and because the population correlation ρ is defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{E(XY) - \mu_x\mu_y}{\sigma\nu}, \quad (20)$$

15 we know that ρ remains the same under data transformations of the form $\tilde{X} = aX - b$, $\tilde{Y} =$
 16 $cY - d$. In particular, we can take $b = \mu_x$, $d = \mu_y$, $a = 1/\sigma$, $c = 1/\nu$ and conclude that ρ does
 17 not depend on these common parameters θ_0 . Hence, we nest π_0 within π_1 orthogonally, that
 18 is, $\pi_1(\theta_0, \rho) = \pi_1(\rho)\pi_0(\theta_0)$.

19 **The denominator of $\text{BF}_{10}(d)$.** Jeffreys's default choice leads to assigning $\pi_0(\theta_0)$
 20 the joint prior $\pi_0(\mu_x, \mu_y, \sigma, \nu) = 1 \cdot 1 \cdot \frac{1}{\sigma} \frac{1}{\nu}$; this is the translation-invariant prior that Jeffreys

1 would use to update to the posterior for θ_0 within either model. When the averaged sum of
 2 squares are both non-zero, this yields the following marginal likelihood of \mathcal{M}_0 :

$$p(d | \mathcal{M}_0) = 2^{-2} n^{-n} \pi^{1-n} (su)^{1-n} \left[\Gamma \left(\frac{n-1}{2} \right) \right]^2. \quad (21)$$

3 Eq. (21) defines the denominator of the correlation Bayes factor $\text{BF}_{10}(d)$. Observe that this
 4 marginal likelihood does not depend on the sample correlation coefficient r .

5 **Step 2. Predictive matching: Symmetric $\pi_1(\rho)$**

6 **Deriving the test-relevant likelihood function.** We now discuss how the test-
 7 relevant prior $\pi_1(\rho)$ can be defined such that the resulting Bayes factor is well-calibrated.
 8 The conclusion is as before: we require $\pi_1(\rho)$ to be symmetric around zero. We discuss the
 9 result more extensively as it cannot be found in Jeffreys (1961). Furthermore, the test-relevant
 10 likelihood function reported by Jeffreys (1961, p. 291, Eq 8) is in fact an approximation of
 11 the result given below.

12 Before we can discuss the calibration we first derive the test-relevant likelihood function
 13 by integrating out the common parameters θ_0 from Eq. (19) with respect to the translation-
 14 invariant priors $\pi_0(\theta_0)$ as outlined by Eq. (5). This leads to the following simplification:

$$p(d | \mathcal{M}_1) = p(d | \mathcal{M}_0) \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho, \quad (22)$$

15 where h is the test-relevant likelihood function that depends on n, r, ρ alone and is given by
 16 Eqn. (24, 25). The Bayes factor, therefore, reduces to

$$\text{BF}_{10}(d) = \frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)} = \frac{p(d | \mathcal{M}_0) \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho}{p(d | \mathcal{M}_0)} = \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho. \quad (23)$$

17 Note that whereas $p(d | \mathcal{M}_0)$ does not depend on ρ or the statistic r (see Eq. (21)), the
 18 function h does not depend on the statistics $\bar{x}, s^2, \bar{y}, u^2$ that are associated with the common
 19 parameters. Thus, the evidence for \mathcal{M}_1 over \mathcal{M}_0 resides within n and r alone.

20 The test-relevant likelihood function $h(n, r | \rho)$ possess more regularities. In particular,
 21 it can be decomposed into an even and an odd function, that is, $h = A + B$, with A defined
 22 as

$$A(n, r | \rho) = (1 - \rho^2)^{\frac{n-1}{2}} {}_2F_1 \left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{1}{2}; (r\rho)^2 \right), \quad (24)$$

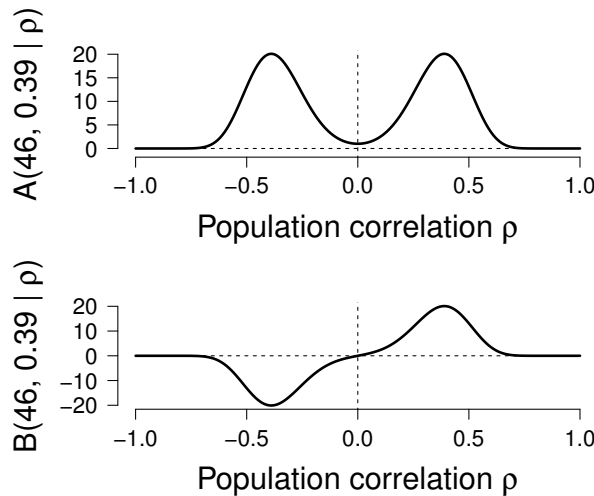


Figure 3. $A(n, r | \rho)$ is an even function of ρ , and $B(n, r | \rho)$ is an odd function of ρ . Together, A and B determine the function h from Eq. (23): $h(n, r | \rho) = A(n, r | \rho) + B(n, r | \rho)$. For this illustration, we used $n = 46$ and $r = 0.39$ based on the example data discussed below.

1 where ${}_2F_1$ is Gauss' hypergeometric function (see Appendix B for details). Observe that A is
 2 a symmetric function of ρ when n and r are given. The second function B is relevant for the
 3 one-sided test and is given by

$$B(n, r | \rho) = 2r\rho(1 - \rho^2)^{\frac{n-1}{2}} \left[\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \right]^2 {}_2F_1\left(\frac{n}{2}, \frac{n}{2}; \frac{3}{2}; (r\rho)^2\right), \quad (25)$$

4 which is an odd function of ρ when n and r are given. Thus, the test-relevant likelihood
 5 function h that mediates inference about the presence of ρ from n and r is given by $h(n, r | \rho) =$
 6 $A(n, r | \rho) + B(n, r | \rho)$. Examples of the functions A and B are shown in Fig. 3.

7 **Predictive matching and the minimal sample size of $n_{\min} = 3$.** Interestingly,
 8 the predictive matching principle implies the use of a symmetric test-relevant prior as in the
 9 previous case. Note that we cannot infer the correlation of a bivariate normal distribution
 10 whenever we have only a single data pair (x, y) ; r is undefined when $n = 1$. Furthermore,
 11 when $n = 2$ we automatically get $r = 1$ or $r = -1$ or $r = -1$ regardless whether or not $\rho = 0$
 12 holds true. As such, nothing is learned up to $n_{\min} = 3$ when testing the nullity of ρ . Hence,
 13 we have to choose $\pi_1(\rho)$ such that the resulting Bayes factor Eq. (23) equals one for $n = 1$
 14 and $n = 2$ regardless of the actually observed r .

15 Using $n = 1$ in Eq. (24) and Eq. (25) we see that $h(1, r | \rho) = A(1, r | \rho) + B(1, r | \rho) = 1$
 16 for every ρ and r . From a consideration of Eq. (23) it follows that for a Bayes factor of
 17 one with $n = 1$, we require $\pi_1(\rho)$ to integrate to one (i.e., $\text{BF}_{10}(d) = \int_{-1}^1 \pi_1(\rho) d\rho = 1$),
 18 underscoring Jeffreys's claim that test-relevant priors should be proper.² Similarly, for $n = 2$

²Jeffreys rejected the translation-invariant prior $\rho \propto (1 - \rho^2)^{-1}$ because it leads to unwelcome results when

1 we automatically obtain $|r| = 1$ and plugging this into Eq. (24) yields $A(2, |r| = 1 | \rho) = 1$.
 2 Thus, with $\pi_1(\rho)$ a proper prior this yields a Bayes factor of $\text{BF}_{10}(d) = 1 + \int_{-1}^1 B(2, |r| =$
 3 $1 | \rho) \pi_1(\rho) d\rho$. To ensure that the Bayes factor equals one for data with a sample size of $n = 2$
 4 we have to nullify the contribution of the function $B(2, |r| = 1 | \rho)$. This occurs when $\pi_1(\rho)$
 5 is symmetric around zero, since $B(2, r | \rho)$ is an odd function of ρ , see Fig. 3.

6 Step 3. Information consistency

7 On the other hand, a sample correlation $r = 1$ or $r = -1$ with $n \geq n_{\min} = 3$ can
 8 be considered overwhelmingly informative data in favor of the alternative model \mathcal{M}_1 . In
 9 our quest to find the right test-relevant prior that yields a Bayes factor that is information
 10 consistent, we consider the so-called stretched symmetric beta distributions given by

$$\pi_1(\rho; \kappa) = \frac{2^{\frac{\kappa-2}{\kappa}}}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})} (1 - \rho^2)^{\frac{1-\kappa}{\kappa}}, \quad (26)$$

11 where $\mathcal{B}(1/\kappa, 1/\kappa)$ is a beta function, see Appendix C for details. Each $\kappa > 0$ yields a
 12 candidate test-relevant prior. Jeffreys's intuitive choice is represented by Eq. (26) with $\kappa = 1$,
 13 as this choice corresponds to the uniform distribution of ρ on $(-1, 1)$. Furthermore, κ can
 14 be thought of as a scale parameter of the prior as in Eq. (13). We claim that a Bayes factor
 15 based on a test-relevant prior Eq. (26) with $\kappa \geq 2$ is information consistent.

16 The resulting Bayes factor

17 To prove the information consistency claim, ρ is integrated out of the test-relevant
 18 likelihood with $h = A + B$ as discussed above (Eq. (23)). This results in the following
 19 analytic Bayes factor:

$$\begin{aligned} \text{BF}_{10; \kappa}(n, r) &= \int_{-1}^1 h(n, r | \rho) \pi_1(\rho; \kappa) d\rho \\ &= \int_{-1}^1 A(n, r | \rho) \pi(\rho; \kappa) d\rho + \underbrace{\int_{-1}^1 B(n, r | \rho) \pi(\rho; \kappa) d\rho}_0 \\ &= \frac{2^{\frac{\kappa-2}{\kappa}} \sqrt{\pi} \Gamma\left(\frac{2+(n-1)\kappa}{2\kappa}\right)}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa}) \Gamma\left(\frac{2+n\kappa}{2\kappa}\right)} {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{2+n\kappa}{2\kappa}; r^2\right), \end{aligned} \quad (27)$$

20 where the contribution of the B -function is nullified due to symmetry of the prior. We
 21 call Eq. (27) Jeffreys's exact correlation test, as we believe that Jeffreys would have derived

testing the null hypothesis $\rho = 1$. However, Robert et al. (2009) noted that such a test is rather uncommon as interest typically centers on the point null hypothesis $\mathcal{M}_0 : \rho = 0$.

1 this Bayes factor $\text{BF}_{10;\kappa}(n, r)$, if he had deemed it necessary to calculate it exactly.

2 Table 1 lists the Bayes factors for a selection of values for κ and n with $r = 1$ fixed; the
 3 results confirm that the Bayes factor is indeed information consistent when $\kappa \geq 2$. Note that
 4 Jeffreys's choice of $\kappa = 1$ does not lead to a Bayes factor which provides extreme support
 5 for \mathcal{M}_1 when confronted with data that are overwhelmingly informative (i.e., $r = 1$ and
 6 $n_{\min} = 3$). However, this Bayes factor does diverge when $n \geq 4$. Thus, Jeffreys's intuitive
 7 choice for κ misses the information consistency criterion by one data pair. The resulting
 8 Bayes factor $\text{BF}_{10;\kappa}(n, r)$ now awaits a user's observed r -value and the associated n number
 9 of observations. In what follows, we honor Jeffreys's intuition and showcase the correlation
 10 Bayes factor using Jeffreys's choice $\kappa = 1$.

Table 1

The Bayes factor $\text{BF}_{10;\kappa=2}$ is information consistent as it diverges to infinity when $r = 1$ and $n \geq 3$, while Jeffreys's intuitive choice $\text{BF}_{10;\kappa=1}$ does not do so until $n \geq 4$. Hence, Jeffreys's intuitive choice $\kappa = 1$ misses the information consistency criterion by one observation. Furthermore, note the role of κ ; the smaller it is, the stronger the associated Bayes factors violate the criterion of information consistency.

n	$\text{BF}_{10;\kappa=5}$	$\text{BF}_{10;\kappa=2}$	$\text{BF}_{10;\kappa=1}$	$\text{BF}_{10;\kappa=1/3}$	$\text{BF}_{10;\kappa=1/10}$
1	1	1	1	1	1
2	1	1	1	1	1
3	∞	∞	2	1.2	1.05
4	∞	∞	∞	1.75	1.17
5	∞	∞	∞	3.20	1.36

11

12 **Example: The Bayesian correlation test**

13 We now apply Jeffreys's default Bayesian correlation test to a data set analyzed earlier
 14 by Stulp, Buunk, Verhulst, and Pollet (2013).

15 **Example 2** (Do taller electoral candidates attract more votes?). *Stulp et al. (2013) studied*
 16 *whether there exists a relation between the height of electoral candidates and their popularity*
 17 *among voters. Based on the data from $n = 46$ US presidential elections, Stulp et al. (2013)*
 18 *reported a positive linear correlation of $r = .39$ between X , the relative height of US presidents*
 19 *compared to their opponents, and Y , the proportion of the popular vote. A frequentist analysis*
 20 *yielded $p = .007$. Fig. 4 displays the data. Based in part on these results, Stulp et al. (2013,*
 21 *p. 159) concluded that "height is indeed an important factor in the US presidential elections",*
 22 *and "The advantage of taller candidates is potentially explained by perceptions associated*
 23 *with height: taller presidents are rated by experts as 'greater', and having more leadership*
 24 *and communication skills. We conclude that height is an important characteristic in choosing*
 25 *and evaluating political leaders."*

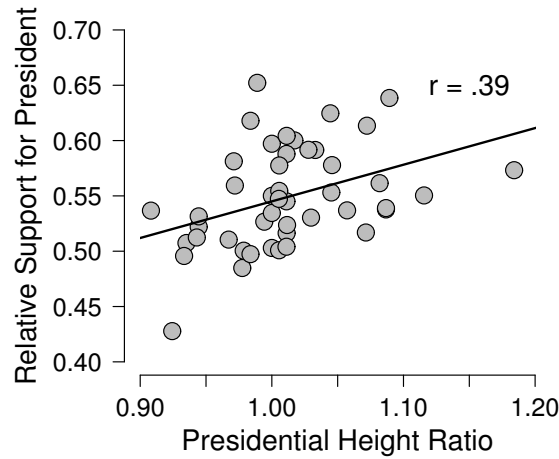


Figure 4. The data from $n = 46$ US presidential elections, showing the proportion of the popular vote for the president versus his relative height advantage against the closest competitor. The sample correlation equals $r = .39$, and, assuming an unrealistic sampling plan, the p -value equals .007. Jeffreys's default two-sided Bayes factor equals $BF_{10; \kappa=1}(n = 46, r = .39) = 6.33$, and the corresponding one-sided Bayes factor equals $BF_{+0; \kappa=1}(n = 46, r = .39) = 11.87$. See text for details.

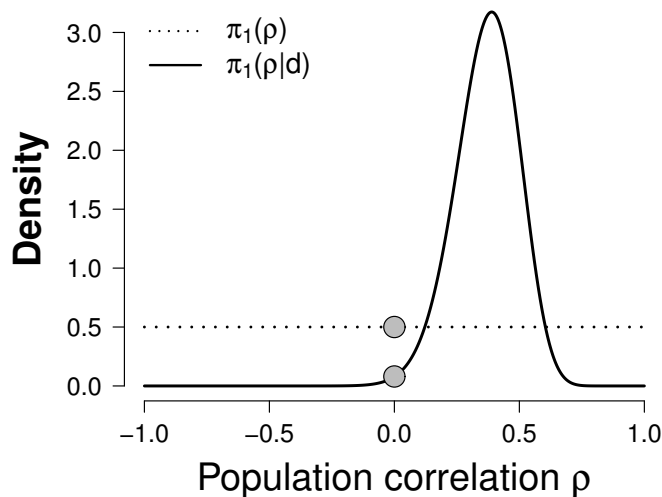


Figure 5. Posterior and prior distributions of the population correlation coefficient ρ for a two-sided default Bayes factor analysis of the height-popularity relation in US presidents (Stulp et al., 2013). The Jeffreys default Bayes factor of $BF_{10; \kappa=1} = 6.33$ equals the height ratio of the prior distribution $\pi_1(\rho)$ over the posterior distribution $\pi_1(\rho|d)$ at $\rho = 0$.

1 *For the Stulp et al. (2013) election data Jeffreys's exact correlation Bayes factor Eq. (27)*
 2 *yields $BF_{10;\kappa=1} = 6.33$, indicating that the observed data are 6.33 times more likely under \mathcal{M}_1*
 3 *than under \mathcal{M}_0 . This result is visualized in Fig. 5 using the Savage-Dickey density ratio test.*
 4 *With equal prior odds, the posterior probability for \mathcal{M}_0 remains an arguably non-negligible*
 5 *14%. ◇*

6 The one-sided extension of Jeffreys's exact correlation Bayes factor

7 Whereas the function A fully determines the two-sided Bayes factor $BF_{10;\kappa}(n, r)$, the
 8 function B takes on a prominent role when we compare the null hypothesis \mathcal{M}_0 against the
 9 one-sided alternative \mathcal{M}_+ with $\rho > 0$.

10 To extend Jeffreys's exact correlation Bayes factor to a one-sided version, we retain
 11 the prior on the common parameters θ_0 . For the test-relevant prior $\pi_+(\rho | \kappa)$ we restrict ρ to
 12 non-negative values, which due to symmetry of $\pi_1(\rho | \kappa)$ is specified as

$$\pi_+(\rho; \kappa) = \begin{cases} 2\pi_1(\rho; \kappa) & \text{for } 0 \leq \rho \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

13 Recall that A is an even function of ρ ; combined with the doubling of the prior for ρ this
 14 leads to a one-sided Bayes factor that can be decomposed as

$$BF_{+0;\kappa}(n, r) = \underbrace{BF_{10;\kappa}(n, r)}_{\int_0^1 A(n, r | \rho) \pi_+(\rho; \kappa) d\rho} + \underbrace{C_{+0;\kappa}(n, r)}_{\int_0^1 B(n, r | \rho) \pi_+(\rho; \kappa) d\rho}. \quad (29)$$

15 The function $C_{+0;\kappa}(n, r)$ can be written as

$$C_{+0;\kappa}(n, r) = \frac{2^{\frac{3\kappa-2}{\kappa}} r \kappa}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})((n-1)\kappa + 2)} \left[\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \right]^2 {}_3F_2\left(1, \frac{n}{2}, \frac{n}{2}; \frac{3}{2}, \frac{2+\kappa(n+1)}{\kappa}; r^2\right), \quad (30)$$

16 where ${}_3F_2$ is a generalized hypergeometric function (Gradshteyn & Ryzhik, 2007, p. 1010)
 17 with three upper and two lower parameters.

18 The function $C_{+0;\kappa}(n, r)$ is positive whenever r is positive, since B as a function of
 19 ρ is then positive on the interval $(0, 1)$; consequently, for positive values of r the restricted,
 20 one-sided alternative hypothesis \mathcal{M}_+ is supported more than the unrestricted, two-sided hy-
 21 pothesis \mathcal{M}_1 , that is, $BF_{+0;\kappa}(n, r) > BF_{10;\kappa}(n, r)$. On the other hand, $C_{+0;\kappa}(n, r)$ is negative
 22 whenever r is negative; for such cases, $BF_{+0;\kappa}(n, r) < BF_{10;\kappa}(n, r)$.

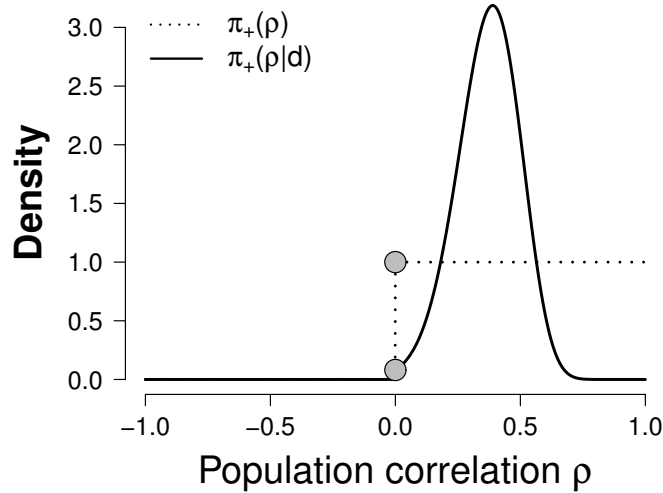


Figure 6. Posterior and prior distributions of the population correlation coefficient ρ for a one-sided default Bayes factor analysis of the height-popularity relation in US presidents Stulp et al. (2013). The Jeffreys default Bayes factor of $BF_{+0;\kappa=1} = 11.87$ equals the height ratio of the prior $\pi_+(\rho)$ over the posterior $\pi_+(\rho|d)$ at $\rho = 0$. The prior $\pi_+(\rho)$ is zero for negative values of ρ . Furthermore, note that the prior distribution $\pi_+(\rho)$ is twice as high for $\rho \geq 0$ compared to $\pi_1(\rho)$ in Fig. 5.

1 **Example 2** (One-Sided Correlation Test for the US President Data, Continued). As shown
 2 in Fig. 6, for the Stulp et al. (2013) data the one-sided Jeffreys's exact correlation Bayes
 3 factor Eq. (29) yields $BF_{+0;\kappa=1} = 11.87$, indicating that the observed data are 11.87 times
 4 more likely under \mathcal{M}_+ than under \mathcal{M}_0 . Because almost all posterior mass obeys the order-
 5 restriction, $BF_{+0} \approx 2 \times BF_{10}$ – its theoretical maximum. \diamond

6

7 Using the same arguments as above, we can define the Bayes factor for a test between
 8 \mathcal{M}_- and \mathcal{M}_0 , which is in fact given by $BF_{-0;\kappa}(n, r) = BF_{+0;\kappa}(n, -r)$ due to the fact that
 9 B is an odd function of ρ . In effect, this implies that $BF_{+0;\kappa}(n, r) + BF_{-0;\kappa}(n, r) = 2 \times$
 10 $BF_{10;\kappa}(n, r)$, where the factor of two follows from symmetry of $\pi_1(\rho; \kappa)$ in the definition of
 11 $\pi_+(\rho; \kappa)$. Additional information on the coherence (Mulder, 2014) of the Bayes factor for
 12 order restrictions is available elsewhere in this special issue (e.g., Mulder, in press).

13 Discussion on the correlation test

14 As mentioned earlier, the previous analysis cannot be found in Jeffreys (1961) as Jeffreys
 15 did not derive the functions A and B explicitly. In particular, Jeffreys (1961, Eqn. (8, 9),
 16 p. 291) suggested that the integral of the likelihood Eq. (19) with respect to the translation-
 17 invariant parameters $\pi_0(\theta_0)$ yields

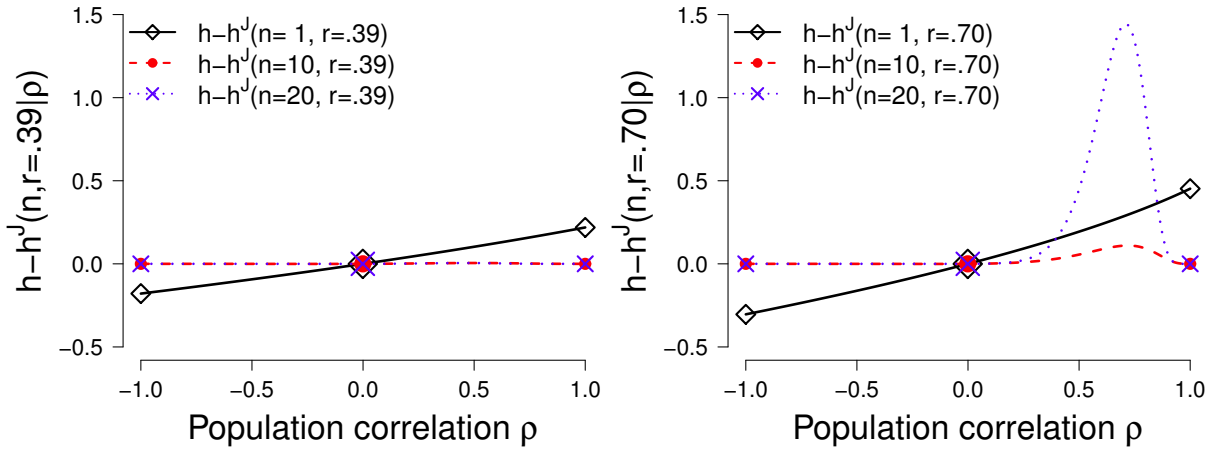


Figure 7. Error of approximation between the exact function h and Jeffreys's approximation h^J . The left panel shows that for a modest sample correlation (i.e., $r = .39$, as in the example on the height of US presidents) Jeffreys's approximation is quite accurate; moreover, the error decreases as n grows, and the curve of $n = 10$ overlaps with that of $n = 20$. However, the right panel shows that for a sample correlation of $r = .70$ the error increases with n , but only for some values of ρ . Furthermore, note that Jeffreys's approximation h^J does not yield $h^J(n = 1, r) = 1$ for every possible r .

$$h^J(n, r | \rho) = \frac{(1 - \rho^2)^{\frac{n-1}{2}}}{(1 - r\rho)^{\frac{2n-3}{2}}}, \quad (31)$$

1 which in fact approximates the true test-relevant likelihood function $h = A + B$ very well for
 2 modest values of $|r|$ (cf. Jeffreys, 1961, p. 175) — this is illustrated in Fig. 7 which plots the
 3 error $h - h^J$. Specifically, the left panel of Fig. 7 shows that when $r = .39$, as in the example
 4 on the height of US presidents, there is virtually no error when $n = 10$. The right panel of
 5 Fig. 7, however, shows that when $r = .70$, the error increases with n , but only for values of
 6 ρ from about .30 to about .95. From Jeffreys's approximation h^J one can define Jeffreys's
 7 integrated Bayes factor (Boekel et al., 2015; Wagenmakers, Verhagen, & Ly, in press):

$$\begin{aligned} \text{BF}_{10}^{\text{J,I}}(n, r) &= \frac{1}{2} \int_{-1}^1 h_J(n, r | \rho) d\rho \\ &= \frac{\sqrt{\pi} \Gamma\left(\frac{n+1}{2}\right)}{2 \Gamma\left(\frac{n+2}{2}\right)} {}_2F_1\left(\frac{2n-3}{4}, \frac{2n-1}{4}; \frac{n+2}{2}; r^2\right). \end{aligned} \quad (32)$$

8 Jeffreys (1961, p. 175) noticed the resulting hypergeometric function, but as these functions
 9 were hard to compute, Jeffreys went on to derive a practical approximation for the users of

his Bayes factor. The final Bayes factor that Jeffreys recommended for the comparison \mathcal{M}_1 versus \mathcal{M}_0 is therefore an approximation of an approximation and given by

$$\text{BF}_{10}^{\text{J}}(n, r) = \sqrt{\frac{\pi}{2n-3}}(1-r^2)^{\frac{4-n}{2}}. \quad (33)$$

For the US presidents data from Example 2 all three Bayes factors yield virtually the same evidence (i.e., $\text{BF}_{10;\kappa=1}(n=46, r=.39) = 6.331$, $\text{BF}_{10}^{\text{J,I}}(n=46, r=.39) = 6.329$, and $\text{BF}_{10}^{\text{J}}(n=46, r=.39) = 6.379$). Table 2 shows that the three Bayes factors generally produce similar outcomes, even for large values of r (cf. Robert et al., 2009). Jeffreys's approximation of an approximation turns out to be remarkably accurate, especially because there is rarely the need to determine the Bayes factor exactly. Jeffreys (1961, p. 432) remarks:

In most of our problems we have asymptotic approximations to K [i.e., BF_{01}] when the number of observations is large. We do not need K with much accuracy. Its importance is that if $K > 1$ the null hypothesis is supported by the evidence; if K is much less than 1 the null hypothesis may be rejected. But K is not a physical magnitude. Its function is to grade the decisiveness of the evidence. It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are 10^4 or 10^{10} to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice.

Table 2

A comparison of Jeffreys's exact Bayes factor (i.e., $\text{BF}_{10;\kappa=1}$) to Jeffreys's approximate integrated Bayes factor (i.e., $\text{BF}_{10}^{\text{J,I}}$) and to Jeffreys approximation of the approximate integrated Bayes factor (i.e., $\text{BF}_{10}^{\text{J}}$) reveals the high accuracy of the approximations, even for large values of r .

n	$\text{BF}_{10;\kappa=1}(n, .7)$	$\text{BF}_{10}^{\text{J,I}}(n, .7)$	$\text{BF}_{10}^{\text{J}}(n, .7)$	$\text{BF}_{10;\kappa=1}(n, .9)$	$\text{BF}_{10}^{\text{J,I}}(n, .9)$	$\text{BF}_{10}^{\text{J}}(n, .9)$
5	1.1	1.1	0.9	2.8	2.8	1.5
10	3.6	3.6	3.2	84.6	83.7	62.7
20	67.5	67.2	63.7	197,753.0	196,698.0	171,571.5

18

Hence, the main advantage of having obtained the exact Bayes factor based on the true test-relevant likelihood function h may be that it justifies Jeffreys's approximation $\text{BF}_{10}^{\text{J}}(n, r)$. The true function h also provides insight in the one-sided version of Jeffreys's test, and it provides a clearer narrative regarding Jeffreys's motivation in model selection and hypothesis testing in general. Moreover, it allows us to show that Jeffreys's exact Bayes factor is model selection consistent.

24

1 **Model selection consistency.** To show that Jeffreys's correlation Bayes factor is
 2 model selection consistent, we use the sampling distribution of the maximum likelihood es-
 3 timate (MLE). As r is the MLE we know that it is asymptotically normal with mean ρ and
 4 variance $\frac{1}{n(1-\rho^2)^2}$, where ρ is the true value. In particular, when the data are generated under
 5 \mathcal{M}_0 , thus, $\rho = 0$, we know that $r \sim \mathcal{N}\left(0, \frac{1}{n}\right)$ when n is large. In order to show that the
 6 support for a true \mathcal{M}_0 grows without bound as the number of data points n increases, the
 7 Bayes factor $\text{BF}_{10;\kappa}(n, r)$ needs to approach zero as n increases.

8 We exploit the smoothness of $\text{BF}_{10;\kappa}(n, r)$ by Taylor expanding it up to third order
 9 in r . By noting that the leading term of the Taylor expansion $\text{BF}_{10;\kappa}(n, 0)$ has a factor
 10 $\Gamma\left(\frac{(n-1)\kappa+2}{2\kappa}\right) / \Gamma\left(\frac{n\kappa+2}{2\kappa}\right)$ we conclude that it converges to zero as n grows. The proof that the
 11 Bayes factor $\text{BF}_{10;\kappa}$ is also model selection consistent under \mathcal{M}_1 follows a similar approach
 12 by a Taylor approximation of second order and consequently concluding that $\text{BF}_{10;\kappa}(n, r)$
 13 diverges to ∞ as n grows indefinitely.

14 Conclusion

15 We hope to have demonstrated that the Bayes factors proposed by Harold Jeffreys have
 16 a solid theoretical basis, and, moreover, that they can be used in empirical practice to answer
 17 one particularly pressing question: what is the degree to which the data support either the
 18 null hypothesis \mathcal{M}_0 or the alternative hypothesis \mathcal{M}_1 ? As stated by Jeffreys (1961, p. 302):

19 “In induction there is no harm in being occasionally wrong; it is inevitable
 20 that we shall be. But there is harm in stating results in such a form that they do
 21 not represent the evidence available at the time when they are stated, or make it
 22 impossible for future workers to make the best use of that evidence.”

23 It is not clear to us what inferential procedures other than the Bayes factor are able
 24 to represent evidence for \mathcal{M}_0 versus \mathcal{M}_1 . After all, the Bayes factor follows directly from
 25 probability theory, and this ensures that it obeys fundamental principles of coherence and
 26 common sense (e.g., Wagenmakers, Lee, Rouder, & Morey, 2014).

27 It needs to be acknowledged that the Bayes factor has been subjected to numerous
 28 critiques. Here we discuss two. First, one may object that the test-relevant prior distribution
 29 for the parameter of interest has an overly large influence on the Bayes factor (Liu & Aitkin,
 30 2008). In particular, uninformative, overly wide priors result in an undue preference for
 31 \mathcal{M}_0 , a fact that Jeffreys recognized at an early stage. The most principled response to this
 32 critique is that the selection of appropriate priors is an inherent part of model specification.
 33 Indeed, the prior offers an opportunity for the implementation of substantively different model
 34 (Vanpaemel, 2010).

35 In this manuscript, we showcased this ability when we adjusted the prior to implement
 36 a directional, one-sided alternative hypothesis. In general, the fact that different priors result
 37 in different Bayes factors should not come as a surprise. As stated by Jeffreys (1961, p. x):

1 “The most beneficial result that I can hope for as a consequence of this work
2 is that more attention will be paid to the precise statement of the alternatives
3 involved in the questions asked. It is sometimes considered a paradox that the
4 answer depends not only on the observations but on the question; it should be a
5 platitude.”

6 The second critique is that in practice, all models are wrong. At first glance this
7 appears not to be a problem, as the Bayes factor quantifies the support for \mathcal{M}_0 versus \mathcal{M}_1 ,
8 regardless of whether these models are correct. However, it is important to realize that the
9 Bayes factor is a relative measure of support. A Bayes factor of $\text{BF}_{10} = 100,000$ indicates
10 that \mathcal{M}_1 receives much more support from the data than does \mathcal{M}_0 , but this does not mean
11 that \mathcal{M}_1 is any good in an absolute sense (e.g., Andraszewicz et al., 2015; Anscombe, 1973).
12 In addition, it has recently been suggested that when both models are misspecified, the Bayes
13 factor may perform poorly in the sense that it is too slow to select the best model (van Erven,
14 Grünwald, & de Rooij, 2012). However, the Bayes factor does have a predictive interpretation
15 that does not depend on one of the model being true (Wagenmakers, Grünwald, & Steyvers,
16 2006); similarly, the model preferred by the Bayes factor will be closest (with respect to the
17 Kullback-Leibler divergence) to the true data-generating model (Berger, 1985; Jeffreys, 1980).
18 More work on this topic is desired and expected.

19 In mathematical psychology, the Bayes factor is a relatively popular method of model
20 selection, as it automatically balances the tension between parsimony and goodness-of-fit,
21 thereby safeguarding the researcher against overfitting the data and preferring models that are
22 good at describing the obtained data, but poor at generalizing and prediction (Myung, Forster,
23 & Browne, 2000; Myung & Pitt, 1997; Wagenmakers & Waldorp, 2006). Nevertheless, with
24 the recent exception of the Bayes factor t -test, the Bayes factors proposed by Jeffreys (1961)
25 have not received much attention, neither by statisticians nor mathematical psychologists.
26 One of the reasons for this unfortunate fact is that Jeffreys notation is more accustomed
27 to philosophers of logic (Geisser, 1980). In order to make Jeffreys’s work somewhat more
28 accessible, Appendix D provides a table with a modern-day translation of Jeffreys’s notation.
29 In addition, any scholar new to the work of Jeffreys is recommended to first read the extended
30 modern summary by Robert et al. (2009).

31 We would like to stress that a Jeffreys Bayes factor is not a mere ratio of likelihood
32 functions averaged with respect to a subjective elicited prior $\pi_i(\theta_i)$ obtained from a within-
33 model perspective. Jeffreys’s development of the Bayes factor resembles an experimental
34 design for which one studies where the likelihood functions overlap, how they differ, and
35 in what way the difference can be apparent from the data. These consideration then yield
36 priors from which a Bayes factor needs to be computed. The computations are typically hard
37 to perform and might not yield analytic results. These computational issues were a major
38 obstacle for the Bayesian community, however, Jeffreys understood that analytic solutions

1 are not always necessary for good inference; moreover, he was able to derive approximate
2 Bayes factors, allowing his exposition of Bayesian inductive reasoning to transcend from a
3 philosophical debate into practical tools for scientific scrutiny.

4 Modern-day statisticians and mathematical psychologists may lack Jeffreys's talent to
5 develop default Bayes factors, but we are fortunate enough to live in a time in which computer-
6 driven sampling methods known as Markov chain Monte Carlo (MCMC: e.g., Gamerman &
7 Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996) are widely available. This removes
8 the computational obstacles one needs to resolve after the priors are specified. These tools
9 makes Jeffreys's method of testing more attainable than ever before.

References

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

- Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*, 521–543.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.
- Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.
- Bayarri, M., Berger, J., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics*, *40*(3), 1550–1577.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Verlag.
- Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*(203), 526–536.
- Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A. J., Brown, S. D., & Forstmann, B. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133.
- Bolt, B. (1982). The constitution of the core: seismological evidence. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *306*(1492), 11–20.
- Cook, A. (1990). Sir Harold Jeffreys. 2 April 1891–18 March 1989. *Biographical Memoirs of Fellows of the Royal Society*, *36*, 302–333.
- Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.
- Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.
- Geisser, S. (1980). The contributions of Sir Harold Jeffreys to Bayesian inference. In A. Zellner & B. Kadane Joseph (Eds.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 13–20). Amsterdam: North-Holland.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.
- Gino, F., & Wiltermuth, S. S. (2014). Evil genius? How dishonesty can lead to greater creativity. *Psychological Science*, *4*, 973–981.
- Good, I. J. (1980). The contributions of Jeffreys to Bayesian statistics. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 21–34). Amsterdam, The Netherlands: North-Holland Publishing Company.
- Gradshteyn, I. S., & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (7th ed.). Academic Press.
- Huzurbazar, V. S. (1991). Sir Harold Jeffreys: Recollections of a student. *Chance*, *4*(2), 18–21.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

- 1 Jeffreys, H. (1924). *The earth, its origin, history and physical constitution*. Cambridge University
2 Press.
- 3 Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press.
- 4 Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the*
5 *Cambridge Philosophical Society*, 31(2), 203–222.
- 6 Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings*
7 *of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007), 453–
8 461.
- 9 Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, UK: Oxford University Press.
- 10 Jeffreys, H. (1955). The present position in probability theory. *The British Journal for the Philosophy*
11 *of Science*, 5, 275–289.
- 12 Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
- 13 Jeffreys, H. (1973). *Scientific inference* (3rd ed.). Cambridge, UK: Cambridge University Press.
- 14 Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner & B. Kadane Joseph
15 (Eds.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys*
16 (pp. 451–453). Amsterdam: North-Holland.
- 17 Jeffreys, H., & Jeffreys, B. S. (1946). *Methods of mathematical physics*. Cambridge, UK: Cambridge
18 University Press.
- 19 Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy*
20 *of Sciences of the United States of America*, 110, 19313–19317.
- 21 Klugkist, I., Laudy, O., & Hoiijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian
22 approach. *Psychological Methods*, 10(4), 477.
- 23 Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge
24 University Press.
- 25 Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the
26 Laplace—Metropolis estimator. *Journal of the American Statistical Association*, 92(438), 648–
27 655.
- 28 Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for
29 Bayesian variable selection. *Journal of the American Statistical Association*, 103(481).
- 30 Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.
- 31 Lindley, D. V. (1980). Jeffreys's contribution to modern statistical thought. In A. Zellner (Ed.),
32 *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 35–
33 39). Amsterdam, The Netherlands: North-Holland Publishing Company.
- 34 Lindley, D. V. (1991). Sir Harold Jeffreys. *Chance*, 4(2), 10–14, 21.
- 35 Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal*
36 *of Mathematical Psychology*, 52, 362–375.
- 37 Ly, A., Marsman, M., Verhagen, A., Grasman, R., & Wagenmakers, E.-J. (2015). A tutorial on Fisher
38 information. *Manuscript submitted for publication at the Journal of Mathematical Psychology*.
- 39 Marin, J.-M., & Robert, C. P. (2010). On resolving the savage–dickey paradox. *Electronic Journal of*
40 *Statistics*, 4, 643–654.
- 41 Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3),
42 220–232.
- 43 Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and

- 1 point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124.
- 2 Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior
3 specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171.
- 4 Mulder, J. (in press). Bayes factors for testing order-constrained hypotheses on correlations. *Journal*
5 *of Mathematical Psychology*.
- 6 Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate
7 linear models: objective model selection using constrained posterior priors. *Journal of Statistical*
8 *Planning and Inference*, *140*(4), 887–906.
- 9 Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*,
10 *47*, 90–100.
- 11 Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of*
12 *Mathematical Psychology*, *44*(1–2).
- 13 Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian
14 approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.
- 15 Oberhettinger, F. (1972). Hypergeometric functions. In M. Abramowitz & I. A. Stegun (Eds.),
16 *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 555–
17 566). New York: Dover Publications.
- 18 O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference*
19 *(2nd ed.)*. London: Arnold.
- 20 Pericchi, L. R., Liu, G., & Torres, D. (2008). Objective Bayes factors for informative hypothe-
21 ses: "Completing" the informative hypothesis and "splitting" the Bayes factor. In H. Hoijtink,
22 I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131–154).
23 New York: Springer Verlag.
- 24 Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathe-*
25 *matics and other logical essays* (pp. 156–198). London: Kegan Paul.
- 26 Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited.
27 *Statistical Science*, 141–172.
- 28 Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for
29 accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.
- 30 Senn, S. (2009). Comment. *Statistical Science*, *24*(2), 185–186.
- 31 Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about
32 the importance of height of US presidents. *The Leadership Quarterly*, *24*(1), 159–171.
- 33 Swirles, B. (1991). Harold Jeffreys: Some reminiscences. *Chance*, *4*(2), 22–23, 26.
- 34 van Erven, T., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: A pre-
35 dictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal*
36 *of the Royal Statistical Society B*, *74*, 361–417.
- 37 Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal*
38 *of Mathematical Psychology*, *54*, 491–498.
- 39 Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the
40 selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.
- 41 Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R. D. (2014). Another statistical paradox,
42 or why intervals cannot be used for model comparison. *Manuscript submitted for publication*.
- 43 Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing

- 1 for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60, 158–189.
- 2 Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (in press). How to quantify the evidence for the
3 absence of a correlation. *Behavior Research Methods*.
- 4 Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey,
5 R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In
6 S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges
7 and proposed solutions*. John Wiley and Sons.
- 8 Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applica-
9 tions [Special issue]. *Journal of Mathematical Psychology*, 50(2).
- 10 Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011).
11 Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests.
12 *Perspectives on Psychological Science*, 6, 291–298.
- 13 Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support
14 for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian
15 *t* test. *Psychonomic Bulletin & Review*, 16(4), 752–760.
- 16 Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical
17 Magazine*, 42, 369–390.
- 18 Zellner, A. (1980). Introduction. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics:
19 Essays in honor of Harold Jeffreys* (pp. 1–10). Amsterdam, The Netherlands: North-Holland
20 Publishing Company.
- 21 Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior
22 distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*,
23 6, 233–243.
- 24 Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In M. Bernardo
25 Jose, H. DeGroot Morris, V. Lindley Dennis, & F. Smith Adrian (Eds.), *Bayesian statistics:
26 Proceedings of the first international meeting held in Valencia* (Vol. 1, pp. 585–603). Springer.

Appendix A

The Default Bayes Factor Hypothesis Tests Proposed by Jeffreys in ToP

27

Appendix B

Hypergeometric Functions

28 The hypergeometric function (Oberhettinger, 1972, section 15) with two upper paramete-
29 rs and one lower parameter generalizes the exponential function as follows (Gradshteyn
30 & Ryzhik, 2007, p 1005):

$${}_2F_1(a, b; c; z) = 1 + \frac{a \cdot b}{c \cdot 1} z + \frac{a(a+1)b(b+1)}{c(c+1) \cdot 1 \cdot 2} z^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2) \cdot 1 \cdot 2 \cdot 3} z^3 + \dots \quad (34)$$

Table A1

Default Bayes factor hypothesis tests proposed by Jeffreys (1961) in Chapter V of "Theory of Probability" (third edition).

Tests	Pages
Binomial rate	256 – 257
Simple contingency	259 – 265
Consistency of two Poisson parameters	267 – 268
Whether the true value in the normal law is zero, σ unknown	268 – 274
Whether a true value is zero, σ known	274
Whether two true values are equal, standard errors known	278 – 280
Whether two location parameters are the same, standard errors not supposed equal	280 – 281
Whether a standard error has a suggested value σ_0	281 – 283
Agreement of two estimated standard errors	283 – 285
Both the standard error and the location parameter	285 – 289
Comparison of a correlation coefficient with a suggested value	289 – 292
Comparison of correlations	293 – 295
The intraclass correlation coefficient	295 – 300
The normal law of error	314 – 319
Independence in rare events	319 – 322

Appendix C

The Stretched Beta Density

- 1 By the change of variable formula, we obtain the stretched beta density of ρ on $(-1, 1)$ with
 2 parameters $\alpha, \beta > 0$

$$\frac{1}{2\mathcal{B}(\alpha, \beta)} \left(\frac{\rho+1}{2}\right)^{\alpha-1} \left(\frac{1-\rho}{2}\right)^{\beta-1}, \quad (35)$$

- 3 where $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function that generalizes $\binom{n}{k}$ to real numbers. By setting
 4 $\beta = \alpha$ this yields the symmetric beta density of ρ on $(-1, 1)$ with parameters $\alpha > 0$

$$\frac{2^{-2\alpha+1}}{\mathcal{B}(\alpha, \alpha)} (1 - \rho^2)^{\alpha-1}. \quad (36)$$

- 5 The reparametrisation we used in text is given by simply substituting $\alpha = 1/\kappa$ allowing us to
 6 interpret κ as a scale parameter.

Appendix D

Translation of Jeffreys's Notation in ToP

Table D1

Translation of the notation introduced by Jeffreys (1961, pp. 245–267). The treatment of α and β as new or old parameters differs from context to context in Jeffreys (1961).

Jeffreys's notation	Modern notation	Interpretation
q	\mathcal{M}_0	Null hypothesis or null model
q'	\mathcal{M}_1	Alternative hypothesis or alternative model
H		Background information (mnemonic: “history”)
$P(q H)$	$P(\mathcal{M}_0)$	Prior probability of the null model
$\int f(\alpha)d\alpha$	$\int \pi(\theta)d\theta$	Prior density on the parameter θ
$P(q'd\alpha H)$	$P(\mathcal{M}_1, \theta)$	Probability of the alternative model and its parameter
$P(q aH)$	$\pi_0(\theta_0 x)$	Posterior density on the parameter within \mathcal{M}_0
$P(q'd\alpha aH)$	$\pi_1(\theta_1 x)$	Posterior density on the parameter within \mathcal{M}_1
K	$\text{BF}_{01}(d)$	The Bayes factor in favor of the null over the alternative
α', β	$\theta_0 = \alpha, \theta_1 = \begin{pmatrix} \alpha' \\ \beta \end{pmatrix}$	“Alternative” parameter $\theta_1 = \begin{pmatrix} \text{function of the old parameter} \\ \text{new parameter} \end{pmatrix}$
$f(\beta, \alpha')$	$\pi_1(\eta \theta_0)$	Prior of the new given the old prior within \mathcal{M}_1
$g_{\alpha\alpha}d\alpha'^2 + g_{\beta,\beta}d\beta^2$	$I(\vec{\theta})$	Fisher information matrix
$P(q, db H) = f(b)db$	$\pi_0(\theta_0)$	Prior density of the common parameters within \mathcal{M}_0
$P(q'dbd\alpha H) = f(b)dbd\alpha$	$\pi_1(\theta_1)$	Prior density of the parameters within \mathcal{M}_1
$P(\theta q, b, H)$	$f(d \theta_0, \mathcal{M}_0)$	The likelihood under \mathcal{M}_0
$P(\theta q', b, \alpha, H)$	$f(d \theta_0, \eta, \mathcal{M}_1)$	Likelihood under \mathcal{M}_1
$P(q db \theta H)$	$\pi_0(\theta_0 d)$	Posterior of the parameters within \mathcal{M}_0
$P(q' db d\alpha \theta H)$	$\pi_1(\theta_1 d)$	Posterior of the parameters within \mathcal{M}_1