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

Alexander Ly, Josine Verhagen, & Eric-Jan Wagenmakers

University of Amsterdam

Abstract

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

Keywords: Model selection; Bayes factors; Harold Jeffreys.

Consider the common scenario where a researcher entertains two competing hypotheses.

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.

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

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

In an attempt to make Jeffreys's default Bayes factor tests accessible to a wider audience we explain the basic principles that drove their development and then focus on two popular inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian t-test) and testing the nullity of a correlation. We illustrate Jeffreys's methodology using data sets from psychological studies. The paper is organized as follows: The first section provides some historical background, outlines Jeffreys's principles for scientific inference, and shows how the Bayes factor is a natural consequence of those principles. We decided to include Jeffreys's own words where appropriate, so as to give the reader an accurate impression of Jeffreys's ideas as well as his compelling style of writing. The second section outlines the ideas behind the Bayesian counterpart for the t-test, and the third section reinforces these ideas with a similar analysis for the Bayesian correlation test. For both the t-test and the correlation test, we also derive one-sided versions of Jeffreys's original tests. The fourth section concludes with a summary and a discussion.

Life and Work

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

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

For several decades, Jeffreys was one of only few scientists who actively developed, used, and promoted Bayesian methods. In recognition of Jeffreys's persistence in the face of relative isolation, E. T. Jaynes's dedication of his own book, "Probability theory: The logic of science", reads: "Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it" (Jaynes, 2003). In 1980, the seminal work of Jeffreys was celebrated in the 29-chapter book "Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys" (e.g., Geisser, 1980; Good, 1980; Lindley, 1980; Zellner, 1980). In one of its chapters, Dennis Lindley discusses ToP and argues that "The *Theory* is a wonderfully rich book. Open it at almost any page, read carefully, and you will discover some pearl." (Lindley, 1980, p. 37).[1]

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

Jeffreys's Perspective on Inference

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

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

[1] Lindley's statement resonates with our own experience, as evidenced in Wagenmakers, Lee, Rouder, and Morey (2014), available at **http://www.ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf**.

choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements" (Jeffreys, 1961, p. 5).

Jeffreys's second conviction is that in order to formalize induction one requires a logic of partial belief: "The idea of a reasonable degree of belief intermediate between proof and disproof is fundamental. It is an extension of ordinary logic, which deals only with the extreme cases." (Jeffreys, 1955, p. 275). This logic of partial belief, Jeffreys showed, needs to obey the rules of probability calculus in order to fulfill general desiderata of consistent reasoning –hence, degrees of belief can be thought of as probabilities (cf. Ramsey, 1926).

Jeffreys's third conviction, developed together with Dr. Dorothy Wrinch, is the simplicity postulate (Wrinch & Jeffreys, 1921), that is, the notion that scientific hypotheses can be assigned prior plausibility in accordance with their complexity, such that "the simpler laws have the greater prior probabilities" (e.g., Jeffreys, 1961, p. 47; see also Jeffreys, 1973, p. 38). In the case of testing a point null hypothesis, the simplicity postulate expresses itself through the recognition that the point null hypothesis represents a general law and, hence, requires a separate, non-zero prior probability, which contrasts with the treatment of an estimation problem. In his early work with Wrinch, Jeffreys argued that inductive reasoning demands that general laws are assigned non-zero prior probability (Wrinch & Jeffreys, 1923). This is explained clearly and concisely by Jeffreys himself:

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

> a general law is that for n large we are already expressing strong confidence that no general law is true. The way out is obvious. To make it possible to get a high probability for a general law from a finite sample the prior probability must have at least some positive value independent of n." Jeffreys (1980, p. 452)

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

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

Jeffreys also argued against the logical validity of p-values, famously pointing out that they depend on more extreme events that have not been observed: "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure." (Jeffreys, 1961, p. 385). In a later paper, Jeffreys clarifies this statement: "I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable;

that is, that it has not predicted something that has not happened." (Jeffreys, 1980, p. 453).

In sum, Jeffreys was convinced that induction is an extended form of logic; that this "logic of partial beliefs" needs to treat degrees of belief as probabilities; that simple laws or hypotheses require separate, non-zero prior probabilities, and that a useful and logically consistent method of hypothesis testing need to be comparative, and needs to be based on the data at hand rather then on data that were never observed. These convictions coalesced in Jeffreys's development of the Bayes factor, an attempt to provide a consistent method of model selection and hypothesis testing that remedies the weaknesses and limitations inherent to *p*-value statistical hypothesis testing.

The Bayes Factor

The concept of Bayes factors is entirely general; in particular, Bayes factors may be used to gauge the evidence for multiple models that are structurally very different. Jeffreys, however, was mostly concerned with the simple scenario featuring two nested models: a model $\mathcal{M}_0$ that instantiates a general law, as a point null hypothesis, and a model $\mathcal{M}_1$ that instantiates the negation of the point null hypothesis, relaxing the restriction imposed by the law. For instance, for the correlation test, $\mathcal{M}_0 : \rho = 0$ –the law says that the correlation is absent– and $\mathcal{M}_1$ is defined by specifying a prior distribution on ρ, for instance, $\mathcal{M}_1 : \rho \sim U[-1, 1]$. How should one grade the evidence that the observed data d provide for $\mathcal{M}_0$ versus $\mathcal{M}_1$?

From Bayes' rule it follows that the posterior probability for model $\mathcal{M}_0$ is given by:

$$P(\mathcal{M}_0 \mid d) = \frac{P(d \mid \mathcal{M}_0)P(\mathcal{M}_0)}{P(d)}, \tag{1}$$

and similarly for $\mathcal{M}_1$:

$$P(\mathcal{M}_1 \mid d) = \frac{P(d \mid \mathcal{M}_1)P(\mathcal{M}_1)}{P(d)}. \tag{2}$$

The above two expressions can be combined so that the common term $P(d)$ drops out, and this yields the key expression:

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

This equation has three crucial ingredients. First, the prior odds quantifies the relative plausibility of $\mathcal{M}_0$ and $\mathcal{M}_1$ before observing data d. Most researchers enter experiments with prior knowledge, prior experiences, and prior expectations, and these can in principle be used to determine the prior odds. Jeffreys preferred the assumption that both models are equally likely a priori, such that $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$. This is consistent with the Wrinch-Jeffreys simplicity postulate in the sense that prior mass 1/2 is assigned to a single point (e.g., $\mathcal{M}_0 : \rho = 0$, the general law), and the remaining 1/2 is spread out over the values that are allowed for the unrestricted parameter (e.g., $\mathcal{M}_1 : \rho \sim U[-1, 1]$). In general then, the prior odds quantify a researcher's initial skepticism about the hypotheses under test. The second ingredient is the posterior odds, which quantifies the relative plausibility of $\mathcal{M}_0$ and $\mathcal{M}_1$ after having observed data d. The third ingredient is the Bayes factor (Jeffreys, 1935): the extent to which data d update the prior odds to the posterior odds. The Bayes factor quantifies the relative probability of the observed data under each of the two competing hypotheses. For instance, when $\text{BF}_{10} = 9$, the observed data d are 9 times more likely to have occurred under $\mathcal{M}_1$ than under $\mathcal{M}_0$; when $\text{BF}_{10} = .20$, the observed data d are 5 times more likely to have occurred under $\mathcal{M}_0$ than under $\mathcal{M}_1$.

In what follows, we will focus on the Bayes factor as a measure of the change in relative belief brought about by the data. However, this does not mean that the prior odds are irrelevant or that they can safely be ignored. For instance, in the context of decision making it is evident that the prior odds remain important, for in order to select the action with the highest expected utility across a collection of uncertain events or states of the world, utilities need to be combined with posterior probabilities, and posterior probabilities depend on prior probabilities (e.g., Lindley, 1985). Prior odds are also important when either $\mathcal{M}_0$

or $\mathcal{M}_1$ is highly implausible. Laplace famously said that "The weight of evidence for an extraordinary claim must be proportioned to its strangeness.", an adage that Carl Sagan shortened to "extraordinary claims require extraordinary evidence." This idea is consistent with Equation 3 in the sense that if, say, $\mathcal{M}_1$ is an extraordinary claim, the prior odds would be very much in favor of $\mathcal{M}_0$, and the Bayes factor –the evidence from the data– needs to be very strong in order in to overcome the initial skepticism. Finally, prior odds are also important in situations that feature multiple comparisons, such as in neuroimaging or genetic association studies (e.g., Stephens & Balding, 2009).

A deeper consideration of Equation 3 reveals that the Bayes factor is defined as the ratio of weighted likelihoods, that is, likelihoods averaged over the parameter space and weighted by the prior distribution. The law of total probability implies that $P(d \mid \mathcal{M}_i) = \int P(d \mid \theta, \mathcal{M}_i) P(\theta \mid \mathcal{M}_i) \, \mathrm{d}\theta$, where θ denotes the parameter vector of model $\mathcal{M}_i$. For a comparison between simple point hypotheses, integration is not required and the Bayes factor reduces to a ratio of likelihoods. However, in many cases $\mathcal{M}_0$ has additional "nuisance" parameters, and in general the alternative hypothesis $\mathcal{M}_1$ is defined as a composite hypotheses, with a prior distribution over its parameters. These prior distributions $\pi_i(\theta) = P(\theta \mid \mathcal{M}_i)$ act as weighting functions for the likelihood $P(d \mid \theta)$; as we will discuss below, one of Jeffreys's goals was to create default Bayes factors by using prior distributions that obeyed a series of general desiderata.

Equation 3 also shows that the Bayes factor differs from the p-value in a number of fundamental ways. First, the Bayes factor depends on the relative probability of the observed data d, and probabilities for unobserved, more extreme data are irrelevant. Second, the Bayes factor features a comparison between two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, instead of focusing only on one of the models. Third, the Bayes factor quantifies the relative degree of support that the observed data provide for $\mathcal{M}_0$ versus $\mathcal{M}_1$, and does so in a continuous fashion. In Appendix B of ToP, Jeffreys proposed a series of categories for evidential strength, labeling Bayes factors larger than 10 as "strong" evidence, and Bayes factors lower than 3 as anecdotal or "not worth more than a bare mention". Jeffreys (1961, pp. 256-257) describes a Bayes factor of 5.33 as "odds that would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper". This remark is still relevant because data for which p-values are larger than .01 (i.e., "$p < .05$, reject the null hypothesis") generally do not pass this threshold of evidence (Berger & Delampady, 1987; Edwards, Lindman, & Savage, 1963; Johnson, 2013; Sellke, Bayarri, & Berger, 2001; Wetzels et al., 2011).

Jeffreys's Programme for the Development of the Default Bayes Factor

In the work of Jeffreys that we discuss here, the Bayes factor comparison involves two models, the null hypothesis $\mathcal{M}_0$ and the alternative hypothesis $\mathcal{M}_1$. Each model $\mathcal{M}_i$ specifies how the data d relate to its parameters θ_i in terms of a likelihood function $L(\theta_i \,|\, d, \mathcal{M}_i)$ which is averaged with respect to a prior distribution $\pi_i(\theta_i)$ to yield $P(d \,|\, \mathcal{M}_i)$.

The role of the prior distributions is important. Subjective Bayesians argue that the prior distributions should be constructed so as to reflect one's prior beliefs – beliefs that differ from one substantive application to the next, and from one researcher to the next. In contrast, Jeffreys constructed the prior distributions methodically, in order to respect general desiderata about the discriminability of the competing models. The prior distributions proposed by Jeffreys are therefore based partly on the likelihood functions, that is, on the models under comparison. Jeffreys's goal was to develop a test that could be used to quantify evidence across a broad range of applications; substantive knowledge can be added but the tests proposed by Jeffreys can still be useful as a point of reference. In recognition of Jeffreys's methodical approach to constructing prior distributions we will henceforth refer to the prior distributions π_0 and π_1 as weighting functions.

Furthermore, $\mathcal{M}_0$ is nested under $\mathcal{M}_1$, meaning that the parameters present in $\mathcal{M}_0$ are also present in $\mathcal{M}_1$. Jeffreys's general approach was to set translation-invariant weighting functions π_0 on the common parameters. This already completes the specification of $\mathcal{M}_0$ and allows the calculation of its weighted likelihood $P(d \,|\, \mathcal{M}_0)$. What is left is the specification of the weighting function for the additional parameter that is present only in $\mathcal{M}_1$ (i.e., the parameter of interest). This specification requires special care, as priors that too wide inevitably reduce the weighted likelihood, resulting in a preference for $\mathcal{H}_0$ regardless of the observed data.

To determine the test-relevant weighting function over the additional parameter, Jeffreys used arguments from hypothetical data that were either completely uninformative or infinitely precise. Given uninformative data d, the weighting function $\pi_1(\theta_1)$ should be chosen such that $\text{BF}_{10} = 1$. Given infinitely precise data d that show an effect, the weighting function $\pi_1(\theta_1)$ should be chosen such that $\text{BF}_{10} = \infty$. Table 1 summarizes the Bayes factor tests developed by Jeffreys in Chapter V of ToP. In the following two sections we clarify Jeffreys's reasoning by discussing the development of the default Bayes factors for two scenarios that are particularly relevant for experimental psychology: testing the nullity of a normal mean and the testing the nullity of a correlation coefficient.

Table 1:: Default Bayes factor hypothesis tests proposed by Jeffreys in Chapter V of ToP.

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

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

In this section we first define the data and then associate these to the unknown parameters within each model through their likelihood functions. Next we outline the desiderata on the weighting function for the test-relevant parameter. We discuss Jeffreys's final choice and apply the resulting default Bayes factor to an example data set on cheating and creativity. We develop the one-sided adaptation of Jeffreys's test, after which we conclude the section on the t-test with some historical notes.

Normal Data

For the case at hand, experimental outcomes are assumed to follow a normal distribution characterized by the unknown population mean μ and standard deviation σ. Similarly, the observed data d from a normal distribution can be summarized by two numbers: the observed sample mean $\bar{x}$ and the average sums of squares $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$; hence we write $d = (\bar{x}, s^2)$. The main difference between the null model $\mathcal{M}_0 : \mu = 0$ and its relaxation $\mathcal{M}_1$ is reflected in the population effect size, which is defined as $\delta = \frac{\mu}{\sigma}$. However, as this effect size cannot be observed directly, the Fisherian statistician studies its imprint from the sampled version, namely, the t-statistic $t = \frac{\bar{x}}{s_\nu/\sqrt{\nu}}$, where s_ν refers to the sample standard deviation based on $\nu = n - 1$ degrees of freedom.

Likelihood Functions

Weighted likelihood for $\mathcal{M}_0$. A model defines a likelihood which structurally 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 restriction 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

$$L(\sigma \,|\, d, \mathcal{M}_0) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2}\left[\bar{x}^2 - s^2\right]\right), \tag{4}$$

$$L(\mu, \sigma \,|\, d, \mathcal{M}_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2}\left[(\bar{x} - \mu)^2 + s^2\right]\right), \tag{5}$$

where $L(\sigma \,|\, d, \mathcal{M}_0)$ and $L(\mu, \sigma \,|\, d, \mathcal{M}_1)$ refer to the likelihood of $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. Note that $L(\sigma \,|\, d, \mathcal{M}_0)$ is a function of σ alone, whereas $L(\mu, \sigma \,|\, d, \mathcal{M}_1)$ depends on two parameters, σ and μ.

To obtain the weighted likelihood under $\mathcal{M}_0$, we integrate out the dependence of σ from Eq. (4) as follows:

$$P(d \mid \mathcal{M}_0) = \int L(\sigma \mid d, \mathcal{M}_0) \pi_0(\sigma) \, \mathrm{d}\sigma, \tag{6}$$

where $\pi_0(\sigma)$ is a weighting function. Note that inference about σ cannot be used to discriminate the two models, since σ has the same interpretation as a scaling parameter in both $\mathcal{M}_0$ and $\mathcal{M}_1$. The choice of $\pi_0(\sigma)$ is therefore irrelevant for the Bayes factor as long as we use the same weighting function in both models.

As a default choice, Jeffreys set $\pi_0(\sigma)$ proportional to its reciprocal, that is, $\pi_0(\sigma) \propto \frac{1}{\sigma}$ which also known as the translation-invariant distribution derived from Jeffreys's rule (e.g., Ly, Verhagen, Grasman, & Wagenmakers, 2014). This choice for $\pi_0(\sigma)$ results in the following weighted likelihood for model $\mathcal{M}_0$:

$$P(d \mid \mathcal{M}_0) = \begin{cases} \dfrac{1}{2|\bar{x}|} & \text{n=1}, \quad (7a) \\[2ex] \dfrac{\Gamma\left(\frac{n}{2}\right)}{2\left(\pi n s_n^2\right)^{\frac{n}{2}}} \left(1 + \dfrac{t^2}{\nu}\right)^{\frac{-n}{2}} & \text{n} > 1, \quad (7b) \end{cases}$$

where t is the observed t-value and ν the degrees of freedom defined as before. Typically, only the right term $\left(1 + \frac{t^2}{\nu}\right)^{\frac{-n}{2}}$ of Eq. (7b) is reported, as the first term also appears in the marginal likelihood of $\mathcal{M}_1$ and hence cancels out in the Bayes factor. Hence, Eqn. (7a, 7a) specify the denominator of Jeffreys's Bayes factor BF_{10}.

Weighted likelihood for $\mathcal{M}_1$. We now focus on the numerator of Jeffreys's Bayes factor, $P(d \mid \mathcal{M}_1)$. The only aspect that distinguishes $\mathcal{M}_1$ from $\mathcal{M}_0$ is the treatment of the population mean μ, which under $\mathcal{M}_1$ is free to vary and therefore unknown, making the likelihood function Eq. (5) a function of both μ and σ.

To relate the observed t-value to the population effect size δ within $\mathcal{M}_1$, Jeffreys rewrote Eq. (5) in terms of the effect size δ and σ. The calculation of the weighted likelihood under $\mathcal{M}_1$ requires that we integrate out the dependence of both δ and σ from Eq. (5), that is, average the likelihood Eq. (5) with respect to a weighting function $\pi_1(\mu, \sigma) = \pi_1(\mu \,|\, \sigma)\pi_1(\sigma) = \pi_1(\delta)\pi_0(\sigma)$. By assigning the same weighting function to σ as was done in $\mathcal{M}_0$, we obtain:

$$P(d \,|\, \mathcal{M}_1) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n-1} \int_{-\infty}^{\infty} \exp\left(-\frac{n}{2} \left[\left(\frac{\bar{x}}{\sigma} - \delta \right)^2 + \left(\frac{s}{\sigma} \right)^2 \right] \right) \pi_1(\delta) \, \mathrm{d}\delta \, \mathrm{d}\sigma. \qquad (8)$$

The remaining task is to specify $\pi_1(\delta)$, the weighting function for the test-relevant parameter.

Desiderata on the Weighting Function for δ Based on Extreme Data

As is shown below, Jeffreys proposed his weighting function $\pi_1(\delta)$ based on desiderata obtained from hypothetical, extreme data.

Predictive matching: Symmetric $\pi_1(\delta)$. The first "extreme" case Jeffreys discusses is when $n = 1$; this automatically yields $s^2 = 0$ regardless of the value of $\bar{x}$. Jeffreys noted that a single datum cannot provide support for $\mathcal{M}_1$, as any deviation of $\bar{x}$ from zero can also be attributed to our lack of knowledge of σ. Hence, nothing is learned from only one observation and consequently the Bayes factor should equal 1 whenever $n = 1$.

To ensure that $\mathrm{BF}_{10} = 1$ whenever $n = 1$, Jeffreys entered $n = 1$, thus, $s^2 = 0$ into Eq. (8) and noted that $P(d \,|\, \mathcal{M}_1)$ equals $P(d \,|\, \mathcal{M}_0)$, Eq. (7a) whenever $\pi_1(\delta)$ is taken to be symmetric around zero.

Information consistency: Heavy-tailed $\pi_1(\delta)$. The other extreme case Jeffreys studied is when the data are infinitely precise with $n > 1$ and a sample mean away from zero: $\bar{x} \neq 0$ and $s^2 = 0$. Note that this implies an infinite observed effect size $t = \infty$, something that should produce infinite support in favor of $\mathcal{M}_1$ over $\mathcal{M}_0$, that is, $\mathrm{BF}_{10} = \infty$. For infinitely precise data, the weighted likelihood under $P(\mathcal{M}_0)$ is finite: $P(\mathcal{M}_0 \,|\, d) = \frac{\Gamma(\frac{n}{2})}{2(n\pi\bar{x}^2)^{\frac{n}{2}}}$ for $\mathcal{M}_0$. To obtain $\mathrm{BF}_{10} = \infty$ the weighted likelihood under $\mathcal{M}_1$ needs to be infinite. Jeffreys noted that this occurs whenever the test-relevant weighting function $\pi_1(\delta)$ is taken to be heavy-tailed.

Jeffreys's choice: The standard Cauchy distribution. The Cauchy distribution with scale γ is the most well-known distribution to be both symmetric around zero and heavy-tailed:

$$\pi(\delta\,;\,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{\delta}{\gamma}\right)^{2}\right)}. \tag{9}$$

Jeffreys suggested to use the simplest version, the standard Cauchy distribution with $\gamma = 1$, as the test-relevant weighting function $\pi_1(\delta)$.

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

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

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

Example: The Bayesian Between-Subject t-Test

To illustrate the default Bayesian t-test we extend Eq. (10) to a between-subjects design and apply the test to a psychological data set. The development above is easily generalized to a between-subject design in which observations are assumed to be drawn from two separate normal populations. The difference scores are then once again normally distributed and we can apply Eq. (10) with $\nu = n - 2$ degrees of freedom.

Example 1 (Does cheating enhance creativity?)**.** *Gino and Wiltermuth (2014, Experiment 2) reported that the act of cheating enhances creativity. This conclusion was based on five exper-*

iments. Here we analyze the results from Experiment 2 in which, having been assigned either to a control condition or to a condition in which they were likely to cheat, participants were rewarded for correctly solving each of 20 math and logic multiple-choice problems. Next, participants' creativity level was measured by having them complete 12 problems from the Remote Association Task (RAT; Mednick, 1962).

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

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

$$BF_{10} = \frac{\pi_1(\delta = 0)}{\pi_1(\delta = 0 \mid d)}. \tag{11}$$

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

In this example, both the Bayesian and Fisherian analysis gave the same qualitative result. Nevertheless, the Bayes factor is more conservative, and some researchers may be surprised that, for the same data, $p = .008$ and $P(\mathcal{M}_0 \mid d) = .17$. Indeed, there are many cases in which the Bayesian and Fisherian analyses disagree qualitatively as well as quantitatively (e.g., Wetzels et al., 2011).

The One-Sided Extension of Jeffreys's Bayes Factor

Some reflection suggests that the authors' hypothesis from Example 1 is more specific – the authors argued that cheating leads to more creativity, not less. To take into account

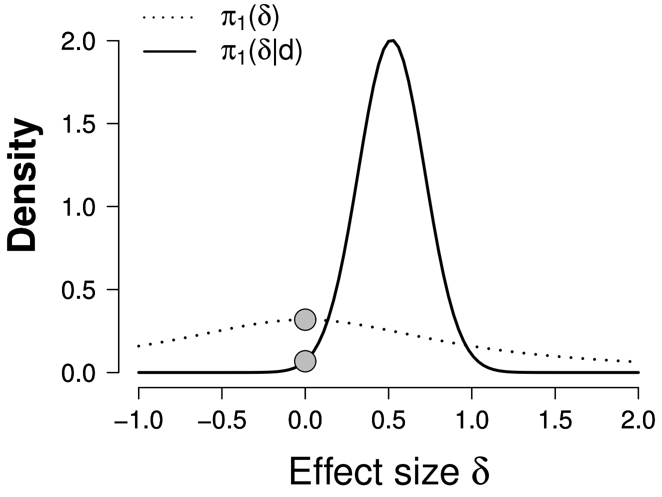

Figure 1. : Posterior and prior weighting functions on effect size for a two-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $\mathrm{BF}_{10\,;\,\gamma=1} \approx 4.60$ equals the ratio of the prior weighting function $\pi_1(\delta)$ over the posterior weighting function $\pi_1(\delta \mid d)$ at $\delta = 0$.

the directionality of the hypothesis we need a one-sided adaptation of Jeffreys's Bayes factor $\mathrm{BF}_{10\,;\,\gamma=1}$. The comparison that is made is then between the model of no effect $\mathcal{M}_0$ and one denoted by M_+ in which the effect size δ is assumed to be positive. We decompose BF_{+0} as follows:

$$\mathrm{BF}_{+0} = \underbrace{\frac{P(d \,|\, M_+)}{P(d \,|\, \mathcal{M}_1)}}_{\mathrm{BF}_{+1}} \underbrace{\frac{P(d \,|\, \mathcal{M}_1)}{P(d \,|\, \mathcal{M}_0)}}_{\mathrm{BF}_{10}}, \tag{12}$$

where BF_{+1} is the Bayes factor that compares the unconstrained model $\mathcal{M}_1$ to the positively restricted model M_+ (Morey & Wagenmakers, 2014). The objective comparison between M_+ and $\mathcal{M}_1$ is then to keep all aspects the same $\pi_+(\sigma) = \pi_1(\sigma) = \pi_0(\sigma)$ except for the distinguishing factor of δ being restricted to positive values within M_+. For the test-relevant weighting function we restrict $\pi_1(\delta)$ to positive values of δ, which by symmetry of the Cauchy distribution means that $\pi_+(\delta)$ accounts doubly for the likelihood when δ is positive and nullifies it when δ is negative (Klugkist, Laudy, & Hoijtink, 2005).

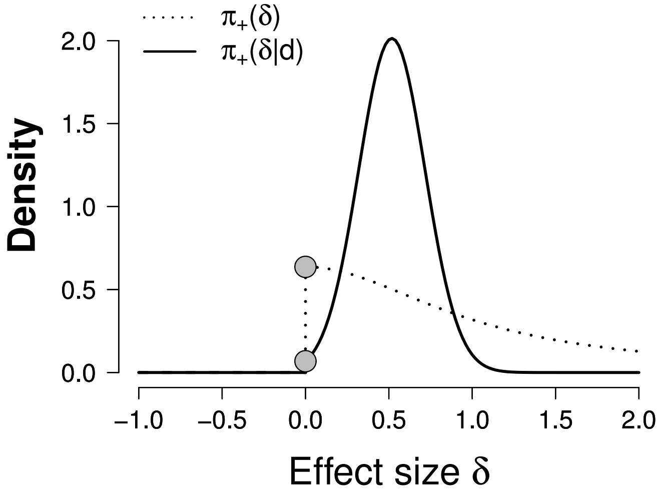

Figure 2. : Posterior and prior weighting functions on effect size for a one-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $\mathrm{BF}_{+0} = 9.18$ equals the ratio of the prior weighting function $\pi_1(\delta)$ over the posterior weighting function $\pi_1(\delta \mid d)$ at $\delta = 0$. The weighting function $\pi_+(\delta)$ is zero for negative values of δ. Furthermore, note that the weights for $\delta \geq 0$ are doubled compared to $\pi_1(\delta)$ in Fig. 1.

Example 1 (One-Sided Test for the Gino and Wiltermuth Data, Continued)**.** *For the data from Gino and Wiltermuth (2014, Experiment 2) the one-sided adaptation of Jeffreys's Bayes factor Eq.* (10) *yields $BF_{+0} = 9.18$. Because almost all of the posterior mass is consistent with the authors' hypothesis, the one-sided Bayes factor is almost twice the two-sided Bayes factor. The result is visualized through the Savage-Dickey ratio in Fig. 2.* ♢

Discussion on the t-test

Discussion on the updated weights. In this section we showcased Jeffreys's philosophy in selecting the instrumental weighting functions for grading the support that the data provide for $\mathcal{M}_0$ versus $\mathcal{M}_1$. By construction, Jeffreys's Bayes factor resulting from $\pi_1(\sigma) = \pi_0(\sigma) \propto \sigma^{-1}$ and from $\delta \sim \mathcal{C}(0, 1)$ is predictively matched at $n = 1$ and information consistent. That is, it indicates no support for either model, $\mathrm{BF}_{01} = 1$, whenever there is a single datum $n = 1$ and it indicates infinite support for $\mathcal{M}_1$ whenever the data are overwhelming (i.e., $s^2 = 0$ and $\bar{x} \neq 0$, or equivalently, $t = \infty$).

In both models, the data yield updated weights for σ, which are typically not the same,

$\pi_0(\sigma \mid d) \neq \pi_1(\sigma \mid d)$, due to the differences in Eq. (4) and Eq. (5) — even though they were given the same prior weights $\pi_1(\sigma) = \pi_0(\sigma)$. Each updated weighting function $\pi_i(\sigma \mid d)$ should be interpreted as a posterior in estimating σ within their own context, the model M_i. The Bayes factor BF_{10} conveys which of these updated weights $\pi_i(\sigma \mid d)$ received more support from the data.

On the other hand, $\mathcal{M}_0$ provides the same deterministic value of $\delta = 0$ regardless of the observed data. The weighting function on δ was chosen based on between-model considerations with the purpose to infer its presence and not its value. This implies that $\pi_1(\delta \mid d)$ updated from the standard Cauchy distribution might not necessarily be the best default result for estimating the value of δ, unless n is sufficiently large.

In ToP, (Jeffreys, 1961, p. 245) introduced the concept of testing and the distinction with estimation, as follows:

> "In the problems of the last two chapters we were concerned with the estimation of the parameters in a law, the form of the law itself being given. We are now concerned with the more difficult question: in what circumstances do observations support a change of the form of the law itself? This question is really logically prior to the estimation of the parameters, since the estimation problem presupposes that the parameters are relevant."

Therefore, one may argue as follows. The estimation of parameters becomes important and informative only after the presence of the phenomenon has been established convincingly. This will generally happen only when n is not small. When n is not small, from a practical perspective, the choice between different sets of reasonable priors is inconsequential. Of course, "many problems that arise in practice are genuinely estimation problems." Jeffreys (1961, p. 245) but the consideration of such problems is outside the scope of the present article.

Historical note. It took several decades before Jeffreys's Bayes factor was adopted by Zellner and Siow (1980), who used the multivariate Cauchy distribution to generalize the method to the linear regression framework. One practical drawback of their proposal was the fact that the numerical integration required to calculate the Bayes factor becomes computationally demanding as the number of covariates grows.

Liang, Paulo, Molina, Clyde, and Berger (2008) established a computationally efficient alternative to Zellner and Siow (1980) by first decomposing the multivariate Cauchy distribution into a mixture of gamma and normal distributions followed by computational simplifications introduced by Zellner (1986). As a result, only a single numerical integral needs to be performed regardless of the size of the model.

The form of the numerator in Eq. (10) is in fact inspired by Liang et al. (2008) and introduced to psychology by Rouder et al. (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009). The combination $\pi_0(\sigma) \propto \sigma^{-1}$ and $\delta \sim \mathcal{C}(0, 1)$ was dubbed the JZS-prior in honor of Jeffreys, Zellner and Siow; this is understandable in the framework of linear regression, although it should be noted that all ideas for the t-test were already present in the second edition of ToP (Jeffreys, 1948, pp. 242–248).

Model selection consistency. In addition to predictive matching and information consistency, Liang et al. (2008) showed that Zellner and Siow's (1980) generalization of Jeffreys's work is also model selection consistent, which implies that as the sample size n increases indefinitely, the support that the data d provide for the correct data-generating model (i.e., $\mathcal{M}_0$ or $\mathcal{M}_1$) grows without bound. Hence, Jeffreys's default Bayes factor Eq. (10) leads to the correct decision whenever the sample size is sufficiently large.

Other desirable properties on the weighting functions. The JZS-priors are not the only weighting functions that possess the properties mentioned above. In particular, Bayarri, Berger, Forte, and García-Donato (2012) formalized Jeffreys's criteria for the construction of a default Bayes factor and defined four additional desiderata on the test-relevant weighting function, namely: intrinsic prior consistency, invariance criteria, measurement invariance and group invariance. Based on these criteria they derived a weighting function that is known as the robustness prior, which yields Bayes factors that have desirable properties similar to Jeffreys's. An elaborate comparison is beyond the scope of this paper.

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

The previous section showed that Jeffreys constructed the weighting functions for the Bayesian t-test based on a careful analysis of how $\mathcal{M}_0$ and $\mathcal{M}_1$ relate the observed data to the unknown parameters using their likelihood functions. We suspect that Jeffreys constructed the weighting functions for the Bayesian correlation test in a similar fashion. We also believe that Jeffreys forgot to mention that some of his intermediate calculations for the correlation test were (very good) approximations, resulting in a Bayes factor $\mathrm{BF}_{10}^{\mathrm{J}}$ that departs from his original intention. In what follows, we give the re-computed Jeffreys's correlation Bayes factor that can be expressed in closed form, which we refer to as Jeffreys's exact correlation Bayes factor.

For the above reasons we divert from the narrative of Jeffreys (1961, Paragraph 5.5) and instead prefer to: (a) explain Jeffreys's reasoning with a structure analogous to that of the previous section; and (b) give the exact results instead. The first subsection below introduces the data and the models involved in a correlation test. Next we specify the weighting functions on the common parameters, allowing us to compute the weighted likelihood for $\mathcal{M}_0$. Then we reconstruct the reasoning behind Jeffreys's choice for the test-relevant weighting function on the parameter of interest ρ; with this weighting function in place we can calculate the weighted likelihood for $\mathcal{M}_1$, and, hence, the default Bayes factor. This default Bayes factor is then applied to a concrete data set. Next we adapt the exact Bayesian hypothesis correlation test for one-sided testing. We end with a short discussion, highlighting the main differences between the exact Bayes factor and Jeffreys's approximate Bayes factor.

Bivariate Normal Data

The Pearson correlation coefficient quantifies the strength of a linear relation between a pair (X, Y) of continuous, normally distributed random variables. To test the nullity of the population correlation it is helpful to summarize the data for X and Y separately in terms of their respective 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 $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, t^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, respectively. The sample correlation coefficient r then defines an observable measure of the linear relationship between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{nst}. \tag{13}$$

This sample correlation coefficient r is an imperfect reflection of the unobservable population correlation coefficient ρ. Hence, the data can be summarised by the five quantities $d = (\bar{x}, s^2, \bar{y}, t^2, r)$.

The statistical models $\mathcal{M}_0$ and $\mathcal{M}_1$ both assume that the pair (X, Y) follows a bivariate normal distribution, $(X, Y) \sim \mathcal{N}_2(\vec{\mu}, \Sigma)$, where $\vec{\mu} = (\mu_x, \mu_y)$ is the vector of the population means and Σ a two-by-two covariance matrix given by:

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}, \tag{14}$$

where σ^2, τ^2 are the respective population variances of X and Y, and ρ denotes the population correlation coefficient defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma\tau} = \frac{E(XY) - \mu_x\mu_y}{\sigma\tau}. \tag{15}$$

Likelihood Functions

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

$$\begin{aligned} L(\theta_0 \mid d, \mathcal{M}_0) =& (2\pi\sigma\tau)^{-n} \exp\Big(-\frac{n}{2}\Big[\Big(\frac{\bar{x}-\mu_x}{\sigma}\Big)^2 + \Big(\frac{\bar{y}-\mu_y}{\tau}\Big)^2\Big]\Big) \\ & \times \exp\Big(-\frac{n}{2}\Big[\Big(\frac{s}{\sigma}\Big)^2 + \Big(\frac{t}{\tau}\Big)^2\Big]\Big). \qquad (16) \\ L(\theta_1 \mid d, \mathcal{M}_1) =& (2\pi\sigma\tau\sqrt{1-\rho^2})^{-n} \exp\Big(-\frac{n}{2(1-\rho^2)}\Big[\frac{(\bar{x}-\mu_x)^2}{\sigma^2} - 2\rho\frac{(\bar{x}-\mu_x)(\bar{y}-\mu_y)}{\sigma\tau} + \frac{(\bar{y}-\mu_y)^2}{\tau^2}\Big]\Big) \\ & \times \exp\Big(-\frac{n}{2(1-\rho^2)}\Big[\Big(\frac{s}{\sigma}\Big)^2 - 2\rho\Big(\frac{rst}{\sigma\tau}\Big) + \Big(\frac{t}{\tau}\Big)^2\Big]\Big). \qquad (17) \end{aligned}$$

To obtain the weighted likelihood under $\mathcal{M}_0$ we have to integrate out the model parameters $\theta_0 = (\mu_x, \mu_y, \sigma, \tau)$ from Eq. (16), that is:

$$P(d \mid \mathcal{M}_0) = \int\int\int\int L(\theta_0 \mid d, \mathcal{M}_0)\pi(\mu_x, \mu_y, \sigma, \tau)\, \mathrm{d}\mu_x\, \mathrm{d}\mu_y\, \mathrm{d}\sigma\, \mathrm{d}\tau, \qquad (18)$$

where $\pi_0(\theta_0) = \pi_0(\mu_x, \sigma, \mu_y, \tau)$ is a weighting function for the common parameters θ_0. Note that inference about these parameters cannot be used to discriminate $\mathcal{M}_0$ from $\mathcal{M}_1$, since the parameters $\mu_x, \mu_y, \sigma, \tau$ within $\mathcal{M}_0$ have the same interpretation as the corresponding parameters in $\mathcal{M}_1$. The choice for $\pi_0(\theta_0)$ is therefore irrelevant for the Bayes factor as long as we use the same weighting function for the common parameters in $\mathcal{M}_1$.

As a default choice, Jeffreys specified $\pi_0(\theta_0)$ as $\pi_0(\mu_x, \mu_y, \sigma, \tau) = 1 \cdot 1 \cdot \frac{1}{\sigma}\frac{1}{\tau}$, which implies the assignment of translation-invariant distributions –derived from Jeffreys's rule– for each of the parameters independently (Ly et al., 2014). With this choice for $\pi_0(\theta_0)$ we obtain the following weighted likelihood of model $\mathcal{M}_0$:

$$P(d \mid \mathcal{M}_0) = 2^{-2} n^{\frac{1-2n}{2}} \pi^{1-n} (st)^{1-n} \Big[\Gamma\left(\frac{n-1}{2}\right)\Big]^2. \qquad (19)$$

Note that the weighted likelihood does not depend on the sample correlation coefficient r.

Weighted likelihood for $\mathcal{M}_1$. The only aspect that distinguishes $\mathcal{M}_0$ from $\mathcal{M}_1$ is the treatment of the population correlation ρ, which is free to vary and, thus, unknown in $\mathcal{M}_1$. Hence, to compute the weighted likelihood for $\mathcal{M}_1$ we have to integrate out both the four common parameters θ_0 and ρ from Eq. (17) with respect to a weighting function $\pi_1(\theta_1)$. Since ρ does not change with the common parameters $\mu_x, \mu_y, \sigma, \tau$, Jeffreys decided on a weighting function π_1 that can be factored into two independent components $\pi_1(\theta_0, \rho) = \pi_1(\theta_0)\pi_1(\rho)$. With $\pi_1(\theta_0) = \pi_0(\theta_0)$ we obtain the following expression for the weighted likelihood of $\mathcal{M}_1$:

$$P(d \,|\, \mathcal{M}_1) = P(d \,|\, \mathcal{M}_0) \int_{-1}^{1} h(n, r \,|\, \rho)\pi_1(\rho)\mathrm{d}\rho, \tag{20}$$

where h is a function of n, r, ρ that can be written as $h = A + B$, where A is an even and B an odd function of ρ, see Eq. (22) and Eq. (24) below. The fact that $P(d \,|\, \mathcal{M}_0)$ occurs as a component in $P(d \,|\, \mathcal{M}_1)$ implies that the common parameters are not only treated similarly as in $\mathcal{M}_0$, but that their similar treatment also leads to the same evidential value.

The Bayes factor, therefore, simplifies to

$$\mathrm{BF}_{10} = \frac{P(d \,|\, \mathcal{M}_1)}{P(d \,|\, \mathcal{M}_0)} = \frac{P(d \,|\, \mathcal{M}_0) \int_{-1}^{1} h(n, r \,|\, \rho)\pi_1(\rho)\mathrm{d}\rho}{P(d \,|\, \mathcal{M}_0)} = \int_{-1}^{1} h(n, r \,|\, \rho)\pi_1(\rho)\mathrm{d}\rho. \tag{21}$$

Note that whereas $P(d \,|\, \mathcal{M}_0)$ does not depend on ρ or the statistic r (see Eq. (19)), the function h does not depend on the statistics $\bar{x}, s^2, \bar{y}, t^2$ that are associated with the common parameters. Thus, the evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$ resides within n, r alone and is quantified by integrating ρ out of $h(n, r \,|\, \rho)$ with respect to a weighting function $\pi_1(\rho)$. The desirable properties of $\pi_1(\rho)$ can be derived from the two functions that together constitute h. The function A that is relevant for the comparison $\mathcal{M}_1$ versus $\mathcal{M}_0$ is specified as

$$A(n, r \,|\, \rho) = (1 - \rho^2)^{\frac{n-1}{2}} \, {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2} \,;\, \frac{1}{2} \,;\, (r\rho)^2\right), \tag{22}$$

where ${}_2F_1$ is Gauss' hypergeometric function (Oberhettinger, 1972, section 15) with two upper parameters and one lower parameter, generalizing the exponential function as follows (Gradshteyn & Ryzhik, 2007, p 9.114):

$$ {}_2F_1\left(a, b\,;\, c\,;\, z\right) = 1 + \frac{a \cdot b}{\gamma \cdot 1} z + \frac{a(a+1)b(b+1)}{c(c+1) \cdot 1 \cdot 2} z^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2) \cdot 1 \cdot 2 \cdot 3} z^3 + \dots \tag{23} $$

Observe that A is a symmetric function of ρ when n, r are given. The second function B is relevant for the one-sided test and is given by

$$ \begin{aligned} B(n, r \,|\, \rho) =& \frac{1}{2r\rho} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \right]^2 (1-\rho^2)^{\frac{n-1}{2}} \\ & \times \left[\left(1 - 2n(r\rho)^2\right) {}_2F_1\left(\frac{n}{2}, \frac{n}{2}\,;\, \frac{1}{2}\,;\, (r\rho)^2\right) - \left(1 - (r\rho)^2\right) {}_2F_1\left(\frac{n}{2}, \frac{n}{2}\,;\, \frac{-1}{2}\,;\, (r\rho)^2\right) \right], \end{aligned} \tag{24} $$

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

Selecting the Weights on the Population Correlation

As a default weighting function for ρ, Jeffreys chose the uniform distribution, that is, $\pi_1(\rho) = U[-1, 1]$. In this subsection we elaborate on what we suspect to be Jeffreys's reasons for this choice.

Predictive matching: A proper and symmetric $\pi_1(\rho)$. Note that we cannot infer the correlation of a bivariate normal distribution whenever we have only a single data pair (x, y); r is undefined when $n = 1$. Furthermore, when $n = 2$ we automatically get $r = 1$ or $r = -1$. As such, nothing is learned up to $n = 2$ and we therefore require that, for these cases, $\mathrm{BF}_{10} = 1$ or $\int h(n, r \,|\, \rho)\pi_1(\rho)\mathrm{d}\rho = 1$, see Eq. (21).

Using $n = 1$ we see that $h(1, r \,|\, \rho) = 1$ for every ρ and r from which we conclude that

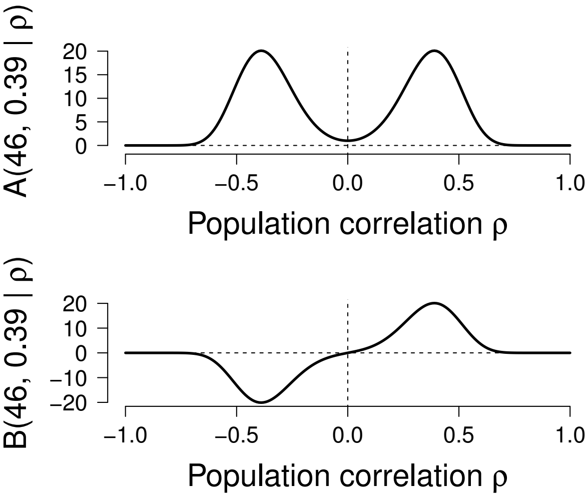

Figure 3. : $A(n, r \mid \rho)$ is an even function of ρ, and $B(n, r \mid \rho)$ is an odd function of ρ. Together, A and B determine the function h from Eq. (21): $h(n, r \mid \rho) = A(n, r \mid \rho) + B(n, r \mid \rho)$. For this illustration, we used $n = 46$ and $r = 0.39$ based on the example data discussed below.

we require $\pi_1(\rho)$ to integrate to one. Similarly, using $n = 2$ we obtain $A(2, r \mid \rho) = 1$ at $r = 1$ and $r = -1$; also, recall that $B(2, r \mid \rho)$ is an odd function of ρ regardless of the value of r, see Fig. 3. Thus, with $\pi_1(\rho)$ a proper weighting function, we obtain $\int_{-1}^{1} A(2, r \mid \rho)\pi_1(\rho)\mathrm{d}\rho = 1$, which implies that $\int_{-1}^{1} B(n, r \mid \rho)\pi_1(\rho)\mathrm{d}\rho = 0$. This occurs whenever $\pi_1(\rho)$ is symmetric around zero.

The symmetric beta distribution and Jeffreys's choice: The uniform distribution. Jeffreys proposed the uniform distribution on ρ after rejecting the translation-invariant distribution because it is inadequate to test $\mathcal{M}_0 : \rho = 1$ or $\mathcal{M}_0 : \rho = -1$ (Jeffreys, 1961, p 290).[2] The uniform distribution is a member of the so-called symmetric beta distributions

$$\pi_1(\rho\,; \alpha) = \frac{2^{1-2\alpha}}{\mathcal{B}(\alpha, \alpha)}(1 - \rho^2)^{\alpha - 1}, \tag{25}$$

where $\mathcal{B}(\alpha, \alpha)$ is a beta function, see Appendix A for details. Each $\alpha > 0$ yields a candidate weighting function. In particular, Eq. (25) with $\alpha = 1$ yields the uniform distribution

[2] Robert et al. (2009) already noted that such a test is rather uncommon as we are typically interested in the point null hypothesis $\mathcal{M}_0 : \rho = 0$. Our reason to reject the translation-invariant distribution on ρ stems from the fact that it cannot be normalized on $(-1, 1)$.

of ρ on $(-1, 1)$. Furthermore, $\gamma = \frac{1}{\alpha}$ can be thought of as a scale parameter as in Eq. (9).

Jeffreys's exact Bayesian correlation test. Jeffreys's Bayes factor now follows from the integral in Eq. (21) with $\pi_1(\rho)$ as in Eq. (25), which Jeffreys did not solve explicitly. Nevertheless, a closed form expression for this integral exists and is given by

$$\begin{aligned} \mathrm{BF}_{10\,;\,\alpha}(n, r) &= \int_{-1}^{1} h(n, r \mid \rho)\pi_1(\rho\,;\,\alpha)\mathrm{d}\rho \\ &= \int_{-1}^{1} A(n, r \mid \rho)\pi(\rho\,;\,\alpha)\mathrm{d}\rho + \underbrace{\int_{-1}^{1} B(n, r \mid \rho)\pi(\rho\,;\,\alpha)\mathrm{d}\rho}_{0} \\ &= \frac{2^{1-2\alpha}\sqrt{\pi}}{\mathcal{B}(\alpha, \alpha)} \frac{\Gamma\left(\frac{n+2\alpha-1}{2}\right)}{\Gamma\left(\frac{n+2\alpha}{2}\right)} {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}\,;\, \frac{n+2\alpha}{2}\,;\, r^2\right). \end{aligned} \tag{26}$$

Jeffreys's exact correlation Bayes factor $\mathrm{BF}_{10\,;\,\gamma=1}(n, r)$ now awaits a user's observed r-value and the number of sample pairs n.

Model selection consistency. To show that Jeffreys's correlation Bayes factor is model selection consistent, we use the sampling distribution of the maximum likelihood estimate (MLE). As r is the MLE we know that it is asymptotically normal with mean ρ and variance $\frac{1}{n(1-\rho^2)^2}$, where ρ is the true value. In particular, when the data are generated under $\mathcal{M}_0$, 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 support for a true $\mathcal{M}_0$ grows without bound as the number of data points n increases, the Bayes factor $\mathrm{BF}_{10\,;\,\alpha}(n, r)$ needs to approach zero as n increases.

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

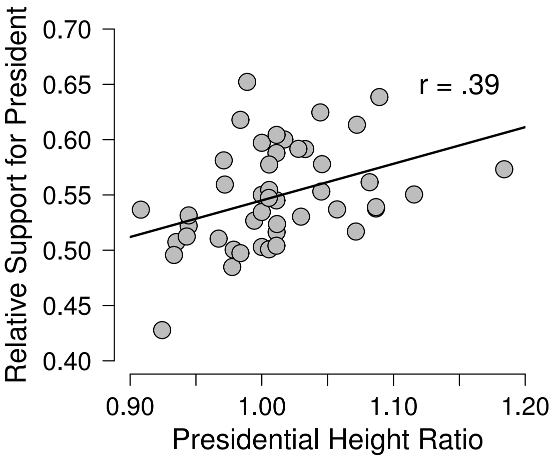

Figure 4. : The data from $n = 46$ US presidential elections, showing the proportion of the popular vote for the president versus his relative height advantage against the closest competitor. The sample correlation equals $r = .39$, and, assuming an unrealistic sampling plan, the p-value equals .007.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

Example: The Bayesian Correlation Test

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

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

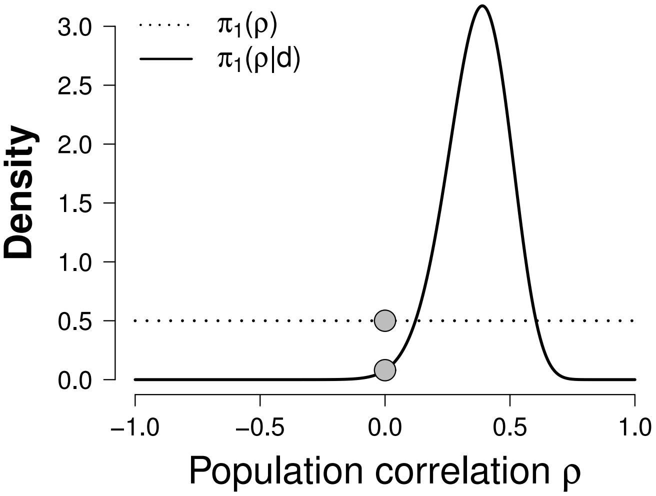

Figure 5. : Posterior and prior weighting functions on the population correlation coefficient ρ for a two-sided default Bayes factor analysis of the height-popularity relation in US presidents Stulp et al. (2013). The Jeffreys default Bayes factor of $BF_{10\,;\,\alpha=1} = 6.33$ equals the ratio of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

Before turning to the Bayes factor analysis, note that the calculation of the p-value is intrinsically tied to probability density function of the data, which in turn depends on the sampling plan. The sampling plan in this example is unknown, as the the data are given to us by external forces, one election result at a time, until the time when the US decides on a different form of government or until it ceases to exist altogether. Hence, when nature provides the data and the sampling plan is unknown, the p-value is unknown as well (Berger & Berry, 1988; Lindley, 1993).

In contrast to the p-value, the Bayes factor does not depend on the sampling plan (i.e., the intentions with which the data have been collected) as it conditions on the observed data and only uses that part of the modeled relationship where the parameters and data interact (Berger & Wolpert, 1988). For the Stulp et al. (2013) election data Jeffreys's exact correlation Bayes factor Eq. (26) *yields* $BF_{10\,;\,\alpha=1} = 6.33$*, indicating that the observed data are* 6.33 *times more likely under* $\mathcal{M}_1$ *than under* $\mathcal{M}_0$*. This result is visualized in Fig. 5 using the Savage-Dickey density ratio test. With equal prior odds, the posterior probability for* $\mathcal{M}_0$ *remains a non-negligible 14%.*

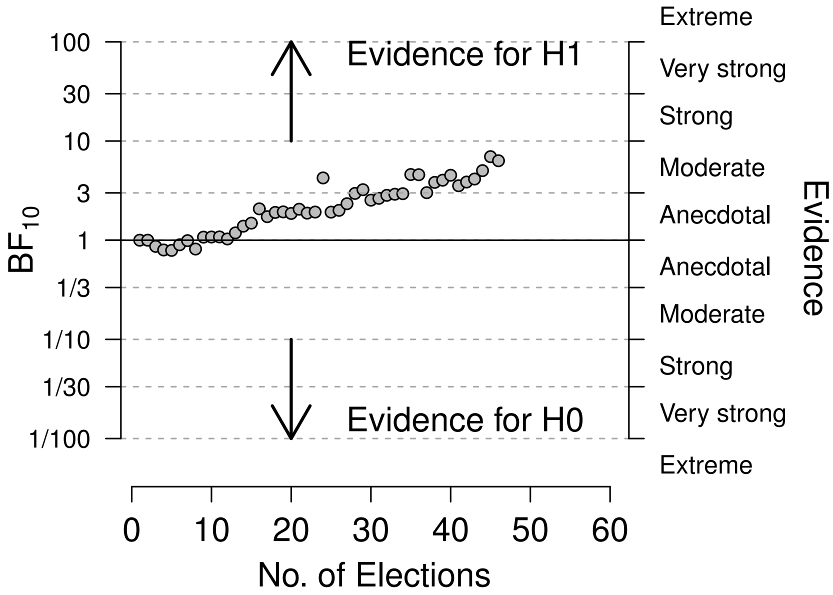

Figure 6. : Development of Jeffreys's default two-sided correlation Bayes factor for the presidential election data reported in Stulp et al. (2013). The category labels on the right Y-axis are inspired by the classification scheme presented by Jeffreys (1961, Appendix B).

The fact that the Bayes factor is independent of the intention with which the data have been collected is of considerable practical relevance. Not only does this independence allow researchers to interpret Bayes factors for data sets collected without a well-defined sampling plan, it also means that researchers may monitor the Bayes factor as the data come in. As pointed out by (Edwards et al., 1963, p. 193), from a Bayesian perspective "(...) the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." (for a recent discussion see Sanborn & Hills, 2014; Rouder, 2014). In Bayesian inference, the sequential analysis of experimental data requires no correction or adjustment (e.g., Andraszewicz et al., in press; Berger & Mortera, 1999; Wagenmakers, 2007). For the example of the US presidents, the development of the Bayesian factor is shown in Eq. (6). ◊

The One-Sided Extension of Jeffreys's Exact Correlation Bayes Factor

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

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

$$\pi_+(\rho\,|\,\alpha) = \begin{cases} 2\pi_1(\rho\,|\,\alpha) & \text{for } 0 \leq \rho \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

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

$$\mathrm{BF}_{+0\,;\,\alpha}(n,r) = \underbrace{\mathrm{BF}_{10\,;\,\alpha}(n,r)}_{\int_0^1 A(n,r\,|\,\rho)\pi_+(\rho\,|\,\alpha)\mathrm{d}\rho} + \underbrace{C_{+0\,;\,\alpha}(n,r)}_{\int_0^1 B(n,r\,|\,\rho)\pi_+(\rho\,;\,\alpha)\mathrm{d}\rho} . \tag{28}$$

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

$$
\begin{aligned}
C_{+0\,;\,\alpha}(n,r) = & \frac{-2^{1-2\alpha} r}{(n+2\alpha-1)(n+2\alpha+1)\mathcal{B}(\alpha,\alpha)} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \right]^2 \\
& \times \Bigg[n^2 r^2 \, {}_3F_2\left(1, \frac{n+2}{2}, \frac{n+2}{2} \,;\, \frac{1}{2}, \frac{n+2\alpha+3}{2} \,;\, r^2\right) \\
& + 2n^3 r^2 \, {}_3F_2\left(1, \frac{n+2}{2}, \frac{n+2}{2} \,;\, \frac{3}{2}, \frac{n+2\alpha+3}{2} \,;\, r^2\right) \\
& - \Big(n^2(n+2\alpha+1)\Big) \, {}_3F_2\left(1, \frac{n+2}{2}, \frac{n+2}{2} \,;\, \frac{3}{2}, \frac{n+2\alpha+1}{2} \,;\, r^2\right) \\
& + 2n^2 - 2\alpha(1-2n) + n - 1 \Bigg],
\end{aligned}
\tag{29}
$$

where ${}_3F_2$ is a generalized hypergeometric function (Gradshteyn & Ryzhik, 2007, Section 9.14) with three upper and two lower parameters.

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

Example 2 (One-Sided Correlation Test for the US President Data, Continued)**.** *As shown in Fig. 7, for the Stulp et al. (2013) data the one-sided Jeffreys's exact correlation Bayes factor Eq.* (28) *yields* $BF_{+0\,;\,\alpha=1} = 11.87$*, indicating that the observed data are* 11.87 *times more likely under* $\mathcal{M}_+$ *than under* $\mathcal{M}_0$*. Because almost all posterior mass obeys the order-restriction,* $BF_{+0} \approx 2 \times BF_{10}$ *– its theoretical maximum.* ♢

Using the same arguments as above, we can define the Bayes factor for a test between M_- and $\mathcal{M}_0$, which is in fact given by $\mathrm{BF}_{-0\,;\,\alpha}(n,r) = \mathrm{BF}_{+0\,;\,\alpha}(n,-r)$ due to the fact that B is an odd function of ρ. In effect, this implies that $\mathrm{BF}_{+0\,;\,\alpha}(n,r) + \mathrm{BF}_{-0\,;\,\alpha}(n,r) = 2 \times \mathrm{BF}_{10\,;\,\alpha}(n,r)$, where the factor of two follows from symmetry of $\pi_1(\rho\,;\,\alpha)$ in the definition of $\pi_+(\rho\,;\,\alpha)$. Hence, only if there is evidence for the presence of ρ, that is $\mathrm{BF}_{10\,;\,\alpha}(n,r) > 1$, can this be spread out over the mutually exclusive models M_+ or M_- with a factor of two to

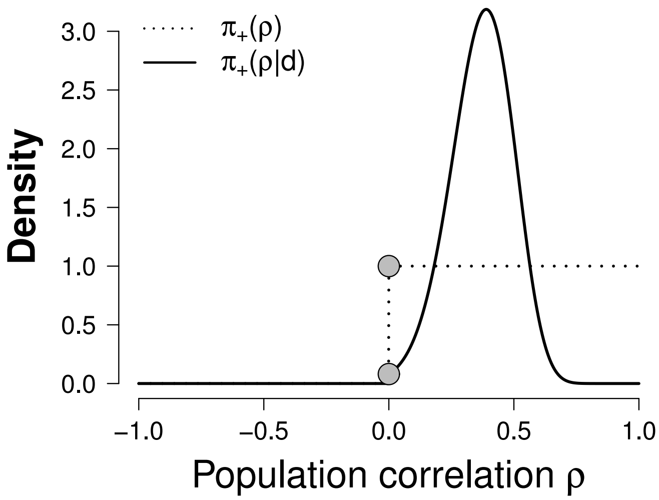

Figure 7. : Posterior and prior weighting functions on the population correlation coefficient ρ for a one-sided default Bayes factor analysis of the height-popularity relation in US presidents Stulp et al. (2013). The Jeffreys default Bayes factor of $\mathrm{BF}_{+0\,;\,\alpha=1} = 11.87$ equals the ratio of the prior weighting function $\pi_+(\rho)$ over the posterior weighting function $\pi_+(\rho \,|\, d)$ at $\rho = 0$. The weighting function $\pi_+(\rho)$ is zero for negative values of ρ. Furthermore, note that the weights for $\rho \geq 0$ are doubled compared to $\pi_1(\rho)$ in Fig. 5.

reward the more specific theory. Additional information on the coherence of the Bayes factor for order restrictions is available elsewhere (e.g., Mulder, in press).

Discussion on the Correlation Test

Comparison between the exact correlation Bayes factor and Jeffreys's approximate Bayes factor. The previous analysis cannot be found in Jeffreys (1961) as Jeffreys did not derive the functions A and B explicitly. In particular, Jeffreys (1961, Eqn. (8, 9), p. 291) claimed that the integral of the likelihood Eq. (17) with respect to the translation-invariant parameters $\pi_0(\theta_0)$ yields

$$h^J(n, r \,|\, \rho) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(1-r\rho)^{\frac{2n-3}{2}}}, \tag{30}$$

which in fact approximates the true function $h = A + B$ very well for modest values of $|r|$ (cf. Jeffreys, 1961, p. 175) — this is illustrated in Fig. 8 which plots the error $h - h^J$.

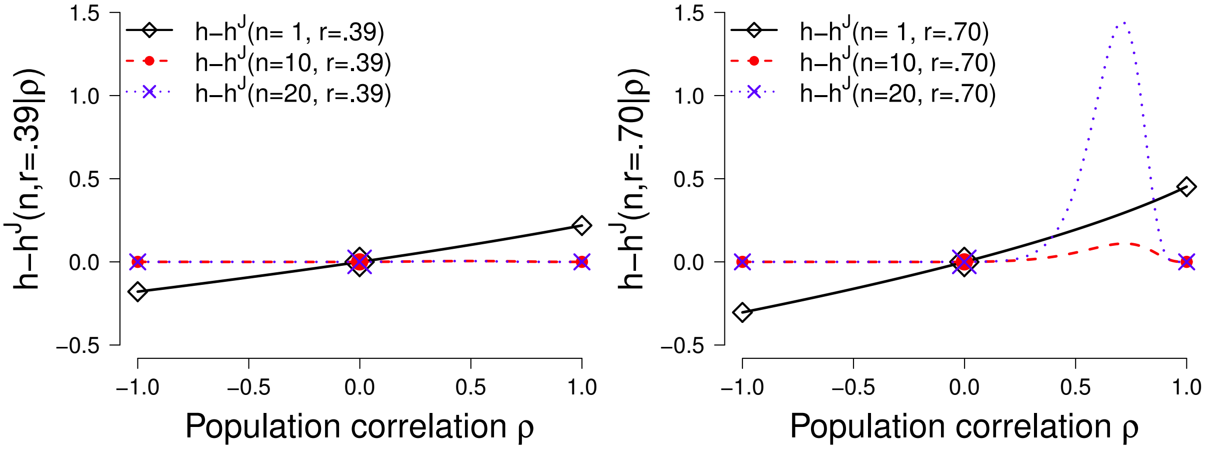

Figure 8. : Error of approximation between the exact function h and Jeffreys's approximation h^J. The left panel shows that for a modest sample correlation (i.e., $r = .39$, as in the example on the height of US presidents) Jeffreys's approximation is quite accurate; moreover, the error decreases as 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.

Specifically, the left panel of Fig. 8 shows that when $r = .39$, as in the example on the height of US presidents, there is virtually no error when $n = 10$. The right panel of Fig. 8, however, shows that when $r = .70$, the error increases with n, but only for values of ρ from about .3 to about .95. From Jeffreys's approximation h^J one can define Jeffreys's integrated Bayes factor (Boekel et al., in press; Wagenmakers, Verhagen, & Ly, 2014)[3]:

$$\begin{aligned}\mathrm{BF}_{10}^{\mathrm{J,I}}(n,r) &= \frac{1}{2}\int_{-1}^{1} h_J(n,r,\rho)\mathrm{d}\rho \\ &= \frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}{}_2F_1\left(\frac{2n-3}{4},\frac{2n-1}{4}\,;\,\frac{n+2}{2}\,;\,r^2\right).\end{aligned} \tag{31}$$

Jeffreys (1961, p. 175) noticed the resulting hypergeometric function, but as these functions 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

[3] The latter manuscript is available online at `http://www.ejwagenmakers.com/submitted/BayesianBathingHabits.pdf`.

$\mathcal{M}_1$ versus $\mathcal{M}_0$ is therefore an approximation of an approximation and given as

$$\mathrm{BF}_{10}^{\mathrm{J}}(n, r) = \sqrt{\frac{\pi}{2n-3}}(1-r^2)^{\frac{n-4}{2}}. \tag{32}$$

For the US presidents data from Example 2 all three Bayes factors yield virtually the same evidence (i.e, $\mathrm{BF}_{10\,;\,\alpha=1}(n = 46, r = .39) = 6.331$, $\mathrm{BF}_{10}^{\mathrm{J,I}}(n = 46, r = .39) = 6.329$, and $\mathrm{BF}_{10}^{\mathrm{J}}(n = 46, r = .39) = 6.379$). 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., $\mathrm{BF}_{10\,;\,\alpha=1}$) to Jeffreys's approximate integrated Bayes factor (i.e., $\mathrm{BF}_{10}^{\mathrm{J,I}}$) and to Jeffreys approximation of the approximate integrated Bayes factor (i.e., $\mathrm{BF}_{10}^{\mathrm{J}}$) reveals the high accuracy of the approximations, even for large values of r.

n	$\mathrm{BF}_{10\,;\,\alpha=1}(n, .7)$	$\mathrm{BF}_{10}^{\mathrm{J,I}}(n, .7)$	$\mathrm{BF}_{10}^{\mathrm{J}}(n, .7)$	$\mathrm{BF}_{10\,;\,\alpha=1}(n, .9)$	$\mathrm{BF}_{10}^{\mathrm{J,\,I}}(n, .9)$	$\mathrm{BF}_{10}^{\mathrm{J}}(n, .9)$
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

Hence, the main advantage of having obtained the exact Bayes factor based on the true function h may be that it justifies Jeffreys's approximation $\mathrm{BF}_{10}^{\mathrm{J}}(n, r)$. 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.

Information consistency and model selection consistency revisited. As the sample correlation has $\nu = n - 2$ degrees of freedom, we require that the Bayes factor diverts to infinity whenever $r = 1$ and $n = 3$; however, with $\alpha = 1$ we have $\mathrm{BF}_{10\,;\,\alpha=1}(n = 3, r = 1) = 2$ from which it follows that Jeffreys's choice does not lead to a Bayes factor that is information consistent. An analysis of the Bayes factor Eq. (26) with $r = 1$ and $n = 3$ reveals that the Bayes factor diverts to infinity only when $\alpha \leq 0.5$. We therefore tentatively suggest that the Bayes factor with $\alpha = 0.5$ may be better calibrated for unambiguous data. In practice, however, we never encounter unambiguous data and a subjective calibration might be more realistic. We therefore chose not to specify a particular value for α in the Bayes factor Eq. (26), although both $\alpha = 0.5$ and $\alpha = 1$ may serve as good benchmarks.

Conclusion

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

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

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

It needs to be acknowledged that the Bayes factor has been subjected to numerous critiques. Here we discuss two. First, one may object that the test-relevant weighting function (i.e., the prior distribution on the parameter of interest) has an overly large influence on the Bayes factor (Liu & Aitkin, 2008). In particular, uninformative, overly wide priors result in an undue preference for $\mathcal{M}_0$, a fact that Jeffreys recognized at an early stage. The most principled response to this critique is that the selection of appropriate weighting functions or priors is an inherent part of model specification. Indeed, the prior offers an opportunity for the implementation of substantively different model (Vanpaemel, 2010). In this manuscript, we showcased this ability when we adjusted the prior to implement a directional, one-sided alternative hypothesis. In general, the fact that different priors result in different Bayes factors should not come as a surprise. As stated by Jeffreys (1961, p. x):

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

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

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

We would like to stress that a Jeffreys Bayes factor is not a mere ratio of likelihood functions averaged with respect to a subjective weighting function $\pi_i(\theta_i)$ obtained from a within-model perspective. Jeffreys's development of the Bayes factor resembles an experimental design for which one studies where the likelihood functions overlap, how they differ, and in what way the difference can be apparent from the data. These consideration then yield weighting functions from which a Bayes factor needs to be computed. The computations are typically hard to perform and might not yield closed form results. These computational issues were a major obstacle for the Bayesian community, however, Jeffreys understood that closed form solutions are not always necessary for good inference; moreover, he was able to derive approximate Bayes factors, allowing his exposition of Bayesian inductive reasoning to transcend from a philosophical debate into practical tools for scientific scrutiny.

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

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., & Wagenmakers, E.-J. (in press). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

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.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*, 542–554.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526–536.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A. J., Brown, S. D., & Forstmann, B. (in press). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*.

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.; A. Jeffrey & D. Zwillinger, Eds.). 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.

Jeffreys, H. (1924). *The earth, its origin, history and physical constitution*. Cambridge University Press.

Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In *Proceedings of the Cambridge philosophical society* (Vol. 31, pp. 203–222).

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1955). The present position in probability theory. *The British Journal for the Philosophy of Science*, *5*, 275–289.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1973). *Scientific inference* (3rd ed.). Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner & B. Kadane Joseph (Eds.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 451–453). Amsterdam: North-Holland.

Jeffreys, H., & Jeffreys, B. S. (1946). *Methods of mathematical physics*. Cambridge, UK: Cambridge University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481).

Lindley, D. V. (1980). Jeffreys's contribution to modern statistical thought. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 35–39). Amsterdam, The Netherlands: North-Holland Publishing Company.

Lindley, D. V. (1985). *Making decisions* (2nd ed.). London: Wiley.

Lindley, D. V. (1991). Sir Harold Jeffreys. *Chance*, *4*(2), 10–14, 21.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Ly, A., Verhagen, A., Grasman, R., & Wagenmakers, E.-J. (2014). A tutorial on Fisher information. *Manuscript submitted for publication at the Journal of Mathematical Psychology.*

Marin, J.-M., & Robert, C. P. (2010). On resolving the savage–dickey paradox. *Electronic Journal of Statistics*, *4*, 643–654.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232.

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124.

Mulder, J. (in press). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Journal of Mathematical Psychology*.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, *44*(1–2).

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Oberhettinger, F. (1972). Hypergeometric functions. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 555–566). New York: Dover Publications.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.

Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan Paul.

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, 141–172.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Senn, S. (2009). Comment. *Statistical Science*, *24*(2), 185–186.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *10*, 681–690.

Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, *24*(1), 159–171.

Swirles, B. (1991). Harold Jeffreys: Some reminiscences. *Chance*, *4*(2), 22–23, 26.

van Erven, T., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society B*, *74*, 361–417.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

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

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R. D. (2014). Another statistical paradox, or why intervals cannot be used for model comparison. *Manuscript submitted for publication.*

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (2014). How to quantify the evidence for the absence of a correlation. *Manuscript submitted for publication.*

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. ??–??). ??: John Wiley and Sons.

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, *50*(2).

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, *16*(4), 752–760.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Wrinch, D., & Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *45*, 368–374.

Zellner, A. (1980). Introduction. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 1–10). Amsterdam, The Netherlands: North-Holland Publishing Company.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, *6*, 233–243.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In M. Bernardo Jose, H. DeGroot Morris, V. Lindley Dennis, & F. Smith Adrian (Eds.), *Bayesian statistics: Proceedings of the first international meeting held in Valencia* (Vol. 1, pp. 585–603). Springer.

Appendix A

The Shifted Beta Density

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

$$\frac{1}{2\mathcal{B}(\alpha,\beta)}\left(\frac{\rho+1}{2}\right)^{\alpha-1}\left(\frac{1-\rho}{2}\right)^{\beta-1}, \tag{33}$$

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

$$\frac{2^{-2\alpha+1}}{\mathcal{B}(\alpha,\alpha)}(1-\rho^2)^{\alpha-1}. \tag{34}$$

Appendix B

Translation of Jeffreys's Notation in ToP

Table B1:: Translation of the notation introduced by Jeffreys (1961, pp. 245–251).

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")
$P(q \mid H)$	$P(\mathcal{M}_0)$	Prior probability of the null model
$\int f(\alpha)\mathrm{d}\alpha$	$\int \pi(\theta)\mathrm{d}\theta$	Prior density on the parameter θ
$P(q'\mathrm{d}\alpha \mid H)$	$P(\mathcal{M}_1,\theta)$	Probability of the alternative model and its parameter
$P(\mathrm{d}\alpha \mid qH)$	$\pi_0(\theta_0)$	Prior density on the parameter within $\mathcal{M}_0$
$P(\mathrm{d}\alpha \mid q'\alpha H)$	$\pi_1(\theta_1)$	Prior density on the parameter within $\mathcal{M}_1$
$P(q \mid aH)$	$\pi_0(\theta_0 \mid x)$	Posterior density on the parameter within $\mathcal{M}_0$
$P(q'\mathrm{d}\alpha \mid aH)$	$\pi_1(\theta_1 \mid x)$	Posterior density on the parameter within $\mathcal{M}_1$
K	BF_{01}	The Bayes factor in favor of the null over the alternative
α',β	$\theta_0=\alpha,\ \theta_1=\tbinom{\alpha'}{\beta}$	"Alternative" parameter $\theta_1=\tbinom{\text{function of the old parameter}}{\text{new parameter}}$
$g_{\alpha\alpha}\mathrm{d}\alpha'^2+g_{\beta,\beta}\mathrm{d}\beta^2$	$I(\vec{\theta})$	Fisher information matrix
$P(\mathrm{d}\alpha\mathrm{d}\beta \mid q\alpha H)$	$\pi_0(\theta_0)$	Prior density on the parameter within $\mathcal{M}_0$
$P(d\alpha\mathrm{d}\beta \mid q'\alpha'\beta H)$	$\pi_1(\theta_1)$	Prior density on the parameter within $\mathcal{M}_1$
$P(q \mid abH)$	$\pi_0(\theta_0 \mid x)$	Posterior density on the parameter within $\mathcal{M}_0$
$P(q' \mid abH)$	$\pi_1(\theta_1 \mid x)$	Posterior density on the parameter within $\mathcal{M}_1$
$f(\beta,\alpha')$	$\pi_1(\beta \mid \alpha')$	Prior of the new given the old prior within $\mathcal{M}_1$