Psychonomic Bulletin & Review (2023) 30:932–941
https://doi.org/10.3758/s13423-022-02205-x

THEORETICAL/REVIEW

Are P-values and Bayes factors valid measures of evidential strength?

Leendert Huisman[1]

Accepted: 17 October 2022 / Published online: 23 November 2022

Abstract
P-values and Bayes factors are commonly used as measures of the evidential strength of the data collected in hypothesis tests. It is not clear, however, that they are valid measures of that evidential strength; that is, whether they have the properties that we intuitively expect a measure of evidential strength to have. I argue here that measures of evidential strength should be stochastically ordered by both the effect size and the sample size. I consider the case that the data are normally distributed and show that, for that case, P-values are valid measures of evidential strength while Bayes factors are not. Specifically, I show that in a sharp Null hypothesis test the Bayes factor is stochastically ordered by the sample size only if the effect size or the sample size is sufficiently large. This lack of stochastic ordering lies at the root of the Jeffreys-Lindley paradox.

Keywords *P*-values · Bayes factors · Hypothesis testing · Evidential strength

Introduction

P-values (Fisher, 1973) and Bayes factors (Jeffreys, 1948) have been proposed and are used as measures of evidential strength in statistical hypothesis testing. They measure that strength in different ways, and their values are not always comparable. Consequently, the last few decades have seen a vigorous debate on which of these two measures is most appropriate for the task of measuring evidential strength (e.g., Dienes & Mclatchie, 2018; Wagenmakers et al., 2008). Considering the sometimes acrimonious nature of that debate, but also considering the recent and very practical problem with replication in many research areas (Camerer et al., 2018; Etz & Vandekerckhove, 2016; OSC, 2015), it is worth stepping back for a moment from the mere comparisons between these measures and reconsider their intrinsic validity as measures of evidential strength. Do they have the properties that we expect a genuine measure of evidential strength to have? Merely comparing their values will not tell us that.

Evidential strength is not a very well defined concept. Intuitively, it is the extent by which the collected data can change our opinion regarding the plausibility of a hypothesis of interest; that is, the extent to which, upon the acquisition of that evidence, the hypothesis becomes more plausible or less plausible, or maybe just less implausible or a bit more plausible. Strong evidence can have a large effect on how plausible or implausible we finally judge the hypothesis to be, while weak evidence has little effect. In order to quantify the concept of "evidential strength" we need a measure of evidential strength, a precise definition that can be computed from the data and that agrees, wherever possible, with our intuitions regarding evidential strength.

Evidential strength is important in hypothesis testing in which data are collected to gain information about the truth or falsehood of one or more selected hypotheses. In principle, multiple hypotheses could be considered, but, in typical applications, there is one hypothesis of central interest, the Null hypothesis, and one catch-all alternative hypothesis, usually the negation of the Null hypothesis. The data can be observations of natural phenomena, such as obtained in astronomy or biology, or outcomes of targeted experiments, such as in psychology or medicine. They are typically generated by probabilistic mechanisms – planet formation around a star, survival of the offspring of some animal of interest, response to a questionnaire, the effectiveness of a new drug, etc., and the hypotheses concern those probabilistic mechanisms.

Because of the variety of possible data that can be collected in different observations or experiments, the data themselves are typically not used directly to make

✉ Leendert Huisman
lmhuisman35@gmail.com

[1] South Burlington, USA

informed statements about the hypotheses. Instead, the values of functions of the observations are calculated that allow for a more or less uniform interpretation across different types of data. These functions are intended to be measures of the evidential strength of the data and depend on the Null hypothesis. They depend on the data and are, therefore, statistics with their own probabilistic distributions that can be derived from the assumed distributions of the data. P-values and Bayes factors are examples of such functions

P-values are measures of the incompatibility between the Null hypothesis and the observed data; between what was expected and what was observed. They have a long history but have also been attacked as inadequate for the purpose of measuring the strength of evidence (e.g., Hubbard & Lindsay, 2008; Wagenmakers, 2007); they have several known shortcomings. For example, since they are measures of incompatibility, they can only indicate how strongly the data undermine the Null hypothesis. Furthermore, P-values come with no provision for measuring the strength of evidence in favor of any hypothesis if it turns out that the Null hypothesis is strongly rejected. In addition, they have shown themselves to be open to misunderstandings, misuses, and abuses (e.g., Goodman, 2008; Greenland et al., 2016). Consequently, there has recently been an increasing push to deprecate the use of P-values in hypothesis testing (e.g., Trafimow & Marks, 2015).

Bayes factors compare how well the Null and alternative hypotheses predict the data, and they can measure the strength of the evidence both for and against the Null hypothesis if the alternative hypothesis is the negation of the Null hypothesis. They were introduced by Jeffreys (1948) and have been suggested as replacement of P-values (e.g., Goodman, 1999; Kass & Raftery, 1995; Morey et al., 2016), but questions have been raised recently (Tendeiro & Kiers, 2019) about their appropriateness as measures of evidential strength. Their well-recognized main shortcoming, other than computational complexity, is that they require prior probability distributions for the constituents of the Null and alternative hypotheses when those hypotheses are composite, as well as the prior probabilities of the (possible composite) Null and alternative hypotheses themselves.

The validity of P-values and Bayes factors as measures of evidential strength can, of course, be studied from many different perspectives. Here, I consider one perspective that focuses on the notion of strength and on how that strength should vary when different parameters of the hypothesis test are varied. The most obvious parameter of a hypothesis test that affects evidential strength is the size of the sample that is used in the test. If that size increases – if more data are collected – the strength of the evidence should increase with it, whether the evidence points at the truth or the falsehood of the Null hypothesis. In particular, the evidence should become overwhelmingly strong in the limit of very large samples;[1] it should indicate with near certainty that the Null hypothesis is true if it is true and false if it is false.

A hypothesis test depends on another parameter than the sample size that is equally relevant to the question of the validity of proposed measures of evidential strength. When the Null hypothesis is false, there is a discrepancy between the true state of affairs and what is being hypothesized about that state of affairs. Of course, the size of that discrepancy is fixed by the actual probabilistic mechanism that produces the data, but we can consider the question of what would happen if the discrepancy were larger than it actually is. In that counter-factual case, the test should produce stronger evidence, even if it were otherwise the same.[2] Furthermore, if the difference between the Null hypothesis and reality is very large, the evidential strength of the data should indicate with near certainty that the Null hypothesis is false.

In this article, I address the question of whether P-values and Bayes factors have the properties of indicating larger strength when either the sample size or the discrepancy becomes larger. If these measures have those properties, I will call them valid. Whether or not P-values and Bayes factors are valid in that sense may depend on the details of the models that describe the probabilistic mechanisms that produce the data. I consider only the simple model in which the data are normally distributed, and I confine myself to sharp Null hypotheses. Moreover, I focus on the case of a false Null hypothesis, because P-values do not measure the strength of the evidence in support of the Null hypothesis. It turns out that, for that simple model, P-values are valid. Bayes factors, on the other hand, are not valid unless the discrepancy or the sample size is sufficiently large. In fact, the observed values of the Bayes factors may be highly misleading, seeming to indicate that the evidence supports the Null hypothesis even though it is false. Moreover, and more seriously, this support of the Null hypothesis, even though it is false, may actually increase when the sample size is still small and more data is collected. This failure of Bayes factors raises serious questions as to their appropriateness as measures of evidential strength, in particular in situations in which both the discrepancy and the sample size are small.

I present the necessary technical details in in the first section, as well as the essential statistical properties of P-values and Bayes factors. The statistical properties of P-values are, of course, well known, and I just summarize them here. The

[1] We can argue that tests that do not lead to correct definite conclusions regarding the truth or falsehood of the Null hypothesis, no matter how large the sample, are not good tests. If the tests are capable of leading to such conclusions when the sample size becomes very large, measures of evidential strength should of course respect that capability.

[2] This amounts to a requirement on the power of the test.

properties of Bayes factors are not difficult to establish but seem to be less well known. I give only as much details as is necessary to establish the main results. The data whose evidential strength needs to be determined are statistical in nature, and the phrase "measure of evidential strength" needs to be properly interpreted by taking that statistical nature into account. I argue that the distributions of proposed measures of evidential strength need to have certain properties for those measures to qualify as valid. In the second section, I discuss two such properties: measures of evidential strength should be stochastically ordered in the right way by the sample size and by the discrepancy between the Null hypothesis and the real state of affairs. I summarize the conclusions in the third section and I attempt to put the lack of validity of the Bayes factor as a measure of evidential strength in the broader context of Bayesian statistics.

Notations and definitions

In this section, I briefly introduce the statistics background of the analysis to be presented in subsequent sections. Everything in this section is well known (or, at least, easily established), and the main purpose of this section is to define notations and present important results.

The sample space consists of the possible outcomes of an experiment.[3] The experiment need not consist of a single act of data acquisition. It can consist of a number of repetitions of the same basic experiment or the recording of some observations on a number of distinct objects (people, situations, physical objects, etc.). The number of repetitions or distinct objects, the sample size, are indicated by 'n'. In the basic problem of Null hypothesis testing considered in this article, only the sample average of the individual outcomes,

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

will be needed, where x_i is the outcome of the i^{th} individual experiment.

The actual outcomes of the experiments are determined by some probabilistic mechanism and the goal of the experiments is to obtain information about that mechanism. The starting point of all statistical inferences is the sequence of observed outcomes, and the set of hypotheses concerning the probabilistic mechanism that produced those outcomes. For frequentists, this set is fully described by

$$M_F =_{\text{def}} < \{f\}, \Omega >, \tag{2}$$

where $\{f\}$ indicates a collection of probability densities[4] on the sample space, this collection being indexed by the members of Ω, the space of hypotheses.

To keep the mathematics simple, I limit the discussion to the standard case of normally distributed data with

$$f\left(x;\theta,\sigma_0^2\right) = \frac{1}{\sqrt{\left(2\pi\sigma_0^2\right)}} e^{-\frac{1}{2}\frac{(x-\theta)^2}{\sigma_0^2}}. \tag{3}$$

The standard deviation σ_0 is fixed and known. The true value of θ, indicated by θ^*, is unknown, and the hypotheses will concern its value. I only consider the sharp Null hypothesis that $\theta^* = 0$. The alternative hypothesis, indicated by H_1, is then that $\theta^* \neq 0$. Furthermore, I only consider the case that the Null hypothesis is false. The experimental quantity of interest is the sample average. It, too, is normally distributed, with mean θ^* and variance $\sigma^2 = {\sigma_0}^2/n$.

A discrepancy between a Null hypothesis and the real state of the world is a vaguely defined quantity, but it can be made precise in the present case of normally distributed data. It is convenient to define the effect size

$$\delta = \frac{\theta^*}{\sigma_0}. \tag{4}$$

The discrepancy between the sharp Null hypothesis and reality is then $|\delta|$.

The P-value will be indicated by 'P_S'. As is well known,

$$P_S = 2\Phi(-|m|/\sigma), \tag{5}$$

where Φ is the standard normal cumulative distribution[5]. As P_S is a statistic, it has a cumulative probability distribution under the true but unknown effect size δ, indicated by '$\text{Prob}_\delta(P_S \leq p)$', for any $p \in [0, 1]$. This cumulative distribution can be calculated easily[6] because P_S not exceeding p implies that $|m|$ is large and m itself is either positive or negative. Under the effect size δ, m/σ has the normal distribution with mean $\delta\sqrt{n}$ and variance 1, and I indicate its cumulative distribution by $\Phi_{\delta\sqrt{n}}$. The latter can easily be calculated from the standard normal distribution using $\Phi_{\delta\sqrt{n}}(z) = \Phi(z - \delta\sqrt{n})$. P_S equals p when $|m|/\sigma$ equals $-\Phi^{-1}(½\ p)$, where Φ^{-1} is the inverse of Φ. Since ½ p does not exceed 0.5, $\Phi^{-1}(½\ p)$ is non-positive and, for P_S to be less than or equal to p, m/σ should either not exceed $\Phi^{-1}(½\ p)$ or be at least as large as $-\Phi^{-1}(½\ p)$. Consequently,

[3] From now on, I simply say "experiment" rather than "observation or experiment."

[4] I will assume that the sample space is continuous. Discrete sample spaces can, of course, be handled easily by replacing probability densities by probability functions.

[5] There is a suppressed dependence on the sample size n because $\sigma^2 = {\sigma_0}^2/n$.

[6] I assume this expression for the cumulative distribution of the P-value in the sharp Null hypothesis test is well known, but I have not been able to find a reference for it.

$$\mathrm{Prob}_{\boldsymbol{\delta}}(P_S \le p) = \Phi_{\boldsymbol{\delta}\sqrt{n}}\left(\Phi^{-1}\left(\frac{1}{2}p\right)\right) + 1 - \Phi_{\boldsymbol{\delta}\sqrt{n}}\left(-\Phi^{-1}\left(\frac{1}{2}p\right)\right).$$

Using the relationship between $\Phi_{\delta\sqrt{n}}$ and Φ, we finally find

$$\mathrm{Prob}_{\delta}(P_S \le p) = \Phi\left(\Phi^{-1}\left(\frac{1}{2}p\right) + \delta\sqrt{n}\right) + \Phi\left(\Phi^{-1}\left(\frac{1}{2}p\right) - \delta\sqrt{n}\right). \quad (6)$$

For Bayesians, the set of hypotheses is slightly more complex:

$$M_B =_{\mathrm{def}} < \{f\}, \Omega, \Psi >, \quad (7)$$

in which $\{f\}$ and Ω have the same meaning as before, and ψ indicates a probability density on Ω, that is, the space of possible values of θ. ψ is generally referred to as the prior." It represents the strengths of the beliefs the agent has in the various hypotheses in Ω. I make the simplifying assumption that ψ is the weighted average of a point mass centered on $\theta = 0$ and a normal distribution[7] with mean 0 and variance τ^2. The weight of the point mass is indicated by 'γ'.

Bayes factors are contrastive in the sense that they compare how well different hypotheses predict the data. Of course, a posterior probability can easily be obtained once the Bayes factor and the prior probability of the model are known. The posterior probability of H_0 equals

$$\psi(H_0|\,m) = \frac{B_{01}\gamma}{B_{01}\gamma + 1 - \gamma}, \quad (8)$$

with the Bayes factor

$$B_{01} = \frac{f(m|H_0)}{f(m|H_1)}. \quad (9)$$

B_{01} is a function of the hypothesis and the data, but I will suppress that dependence. I use B_{01} rather than the perhaps more standard $B_{10} = 1/B_{01}$ because B_{01} shares with P_S the property that the evidence against the Null hypothesis is stronger the smaller B_{01}: the standard interpretation of B_{01} is that the data support H_0 if $B_{01} > 1$ and undermine it if $B_{01} < 1$.

In the case of the sharp Null hypothesis, I indicate the Bayes factor by 'B_S', with (Rouder et al., 2009, Note 3[8])

$$B_S = \frac{\tau}{T} e^{-\frac{1}{2}\frac{m^2}{\sigma^2}\frac{T^2}{\sigma^2}}. \quad (10)$$

in which

$$T^{-2} = \sigma^{-2} + \tau^{-2}. \quad (11)$$

The range of possible B_S values is $[0, \tau/T]$. The cumulative distribution of B_S given δ can be calculated in essentially the same way as the cumulative distribution of P_S. B_S is small if m^2/σ^2 is large, and B_S equals b when m^2/σ^2 equals k(b) with

$$k(b) = 2\frac{\sigma^2}{T^2}\left(\ln\frac{\tau}{T} - \ln(b)\right), \quad (12)$$

a non-negative function of b. We then find

$$\mathrm{Prob}_{\theta*}(B_S \le b) = \Phi\left(-\sqrt{k(b)} + \delta\sqrt{n}\right) + \Phi\left(-\sqrt{k(b)} - \delta\sqrt{n}\right). \quad (13)$$

Note that, as functions of δ, $\mathrm{Prob}_{\delta}(P_S \le p)$ and $\mathrm{Prob}_{\delta}(B_S \le b)$ are very similar: both are symmetric in δ, showing that the two cumulative distribution functions depend only on $|\delta|$. Also note that $\Phi^{-1}(½p)$ does not depend on n and that k(b) depends on n only weakly: for large n, T goes to σ, so σ^2/T^2 goes to 1, τ/T goes to $\tau\sqrt{n}/\sigma_0$, and k(b) goes to ln(n) plus a constant. Consequently, for sufficiently large values of either $|\delta|$ or n, the $\delta\sqrt{n}$ term dominates in the arguments of the cumulative distributions. This implies that both $\mathrm{Prob}_{\delta}(P_S \le p)$ and $\mathrm{Prob}_{\delta}(B_S \le b)$ go to 1 for any value of p or b, no matter how small, if either $|\delta|$ or n goes to infinity. In other words, both P_S and B_S go to 0 in probability if the sample size or the effect size becomes very large. The only exception occurs when $\delta = 0$, because then $\mathrm{Prob}_{\delta}(P_S \le p)$ equals p for all n and $\mathrm{Prob}_{\delta}(B_S \le b)$ goes to 0 for all b.

It will be useful to consider the actual values of B_S as well, and it is convenient to use the expected value of $\ln(B_S)$ for that purpose. We easily obtain from Eq. (10) that

$$\ln(B_S) = -\frac{1}{2}\frac{m^2}{\sigma^2}\frac{T^2}{\sigma^2} + \ln\left(\frac{\tau}{T}\right). \quad (14)$$

Since m/σ has the standard normal distribution with mean $\delta\sqrt{n}$, m^2/σ^2 has the non-central chi-squared distribution with non-centrality parameter $n\delta^2$. The expected value of m^2/σ^2 is then $1 + n\delta^2$, and

$$\text{Expected value of } \ln(B_S) = -\frac{1}{2}(1 + n\delta^2)\frac{T^2}{\sigma^2} + \ln\frac{\tau}{T}. \quad (15)$$

Measuring evidential strength

The P-value and Bayes factor defined in the preceding section have been and are being used as standard measures of the strength of the evidence, the vector of trial outcomes $\{x_i\}$. In the introduction, I argued that, in order for P-values and Bayes factors to be valid measures of evidential strength, they should have the following property: if the Null

[7] A more standard assumption, going back to Jeffreys (1948), is to use a Cauchy distribution for the prior, but that choice leads to mathematical complications that are not germane to the problem at hand.

[8] I use a slightly different notation than Rouder et al.

hypothesis is false, they should be smaller the larger the effect size or the sample size. I do not address the question of whether these measures have the desired properties in the general case, but only in the case that the data are normally distributed (Eq. (3)), and look first at the effect size. Henceforth, τ will be set equal to σ_0 because that choice simplifies the mathematics[9] and because the corresponding prior seems to be reasonably vague in comparison with the effect sizes of interest.

The requirement that the evidence against the Null hypothesis be stronger the larger the effect size is analogous to what makes a good thermometer: the hotter the object whose temperature we are interested in, the higher the temperature reading should be. Unfortunately, the relationship between the effect size and the outcome of an experiment is not as simple as that between the temperature of some object and the thermometer reading; in the latter case the relationship is one to one, while in the former it is not. The statistic m can, in principle, take on any value, and, consequently, the P-value and the Bayes factor can take on any value in their respective ranges as well. The best we can hope for is that these measures are *typically* lower when the effect size is larger, where what is meant by "typically" needs further specification.

Since lower values of these measures are associated with stronger evidence against the Null hypothesis, the most obvious and also the strongest specification is that these measures have the property of being "typically lower when the effect size is larger" when smaller values of the measures become more probable when the effect size increases. More precisely, they should have the property that, for any possible value t of the measure and any non-zero value of $|\delta|$, $\text{Prob}_\delta(\text{observed measure} \leq t)$ increases when $|\delta|$ increases: for any t in the range of the observed measure and any non-zero δ and δ^* with $|\delta| < |\delta^*|$, $\text{Prob}_\delta(\text{observed measure} \leq t) < \text{Prob}_{\delta^*}(\text{observed measure} \leq t)$. More formally and more compactly, the measures should be stochastically ordered by the absolute effect size.

Consider now Eqs. (6) and (13), and assume a fixed value of the sample size and fixed values of p and b, respectively. It is then clear that the P-value and the Bayes factor are stochastically ordered by the effect size, because their cumulative distribution functions are increasing functions of $|\delta|$, as shown in the Appendix (Eq. A.2). As a specific and important example of this stochastic ordering, consider the probability that the Bayes factor is less than 1. This particular value of b is important because it is the separator in the range of Bayes factors between support for the alternative hypothesis ($b < 1$) and support for the Null hypothesis ($b > 1$). The probability that B_S is less than 1 increases when $|\delta|$ increases, indicating that low values of the Bayes factor become more likely and that support for the Null hypothesis, if any, decreases while that for the alternative hypothesis increases.

These results show that P-values and Bayes factors are valid measures of evidential strength when the sample size is kept fixed while the effect size is varied (at least when the data are normally distributed). P-values, as we shall see, maintain their validity when the sample size is varied rather than the effect size, but Bayes factors do not. As with the effect size, both the P-value and the Bayes factor should stochastically decrease, if the Null hypothesis is false, when the sample size n becomes larger. The dependence on n arises from the variance $\sigma^2 = {\sigma_0}^2/n$. The cumulative distribution of the P-value depends on n only via the product $\delta\sqrt{n}$ (see Eq. (6)), so the dependence of the cumulative distribution on $\sqrt{n}$ is the same as that on $|\delta|$. In other words, the P-value is stochastically ordered by $\sqrt{n}$ the same way as it is by $|\delta|$: it decreases stochastically when n increases and converges to 0 in probability when n goes to infinity.

The Bayes factor does not have that property. It is true that it goes to 0 in probability when the Null hypothesis is false and n goes to infinity, but its behavior at small sample sizes can be highly misleading. This is shown clearly by the probability that the Bayes factor is less than 1. Depending on k(1), n and δ, that probability may or may not exceed 0.5, a probability of 0.5 indicating indifference between large and small values of B_S. If the probability is less than 0.5, small values of the Bayes factor are less likely than large values and the Bayes factor is likely to indicate that the Null hypothesis is true, even when it is false. I return to that case later. Whatever the value of the probability, however, if the Bayes factor were a valid measure of the strength of evidence, small values of the Bayes factor should become more likely if the Null hypothesis is false and the sample size increases; the probability of finding a small value of the Bayes factor should increase with n.

But that is not what happens, as shown in Fig. 1. The figure shows $\text{Prob}_\delta(B_S \leq 1)$ as a function of the sample size n for four different values of the effect size δ. In the upper left panel, showing $\text{Prob}_\delta(B_S \leq 1)$ as a function of n when $\delta = 0.05$, the probability starts out at less than 0.5 and decreases when n increases. This implies that, for small sample sizes, B_S is likely to be large (larger than 1, at least) and is likely to become larger when n increases. Only when n is sufficiently large does $\text{Prob}_\delta(B_S \leq 1)$ start to increase.

When the effect size increases (remaining three panels), the behavior of $\text{Prob}_\delta(B_S \leq 1)$ changes in two ways. First, the minimum occurs at progressively smaller sample sizes, referred to as 'n_{min}'. Second, the value of $\text{Prob}_\delta(B_S \leq 1)$ at n_{min} increases. The respective decrease and increase continue until, at δ larger than about 0.5, no minimum occurs at

[9] With that choice, $T^2 = {\sigma_0}^2/(n+1)$ and $k(b) = (1+1/n)(\ln(n+1) - 2\ln(b))$.

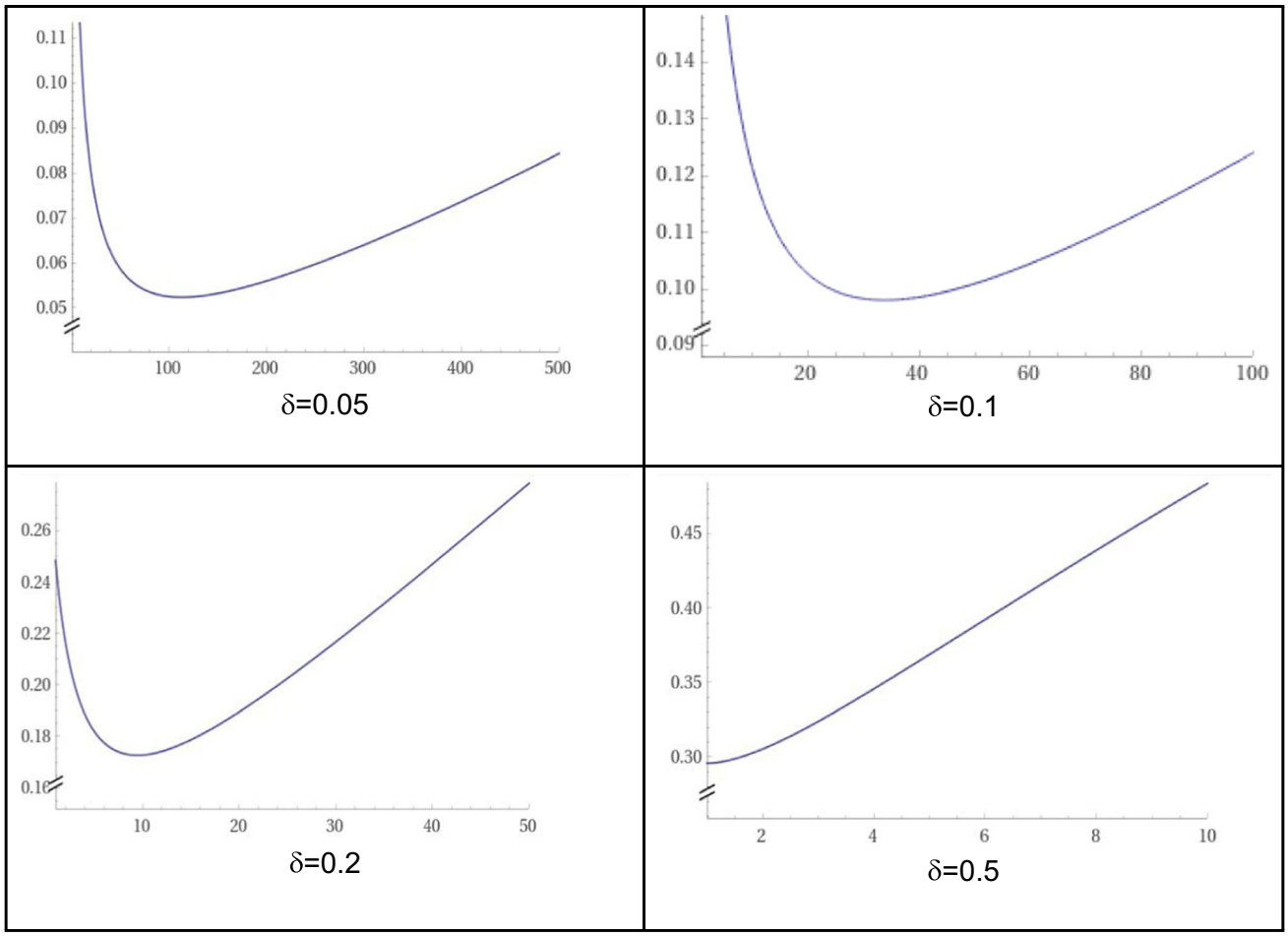

Fig. 1 Probability that B_S is less than 1 as function of the sample size, for various values of δ. Note the varying vertical and horizontal scales

all (see Appendix). On the other hand, at small effect sizes n_{min} can be substantial. For example, it is about 100 at δ = 0.05, 35 at δ = 0.1 and around 10 at δ = 0.2.

What these panels do not show is the actual value of B_S. We can get an idea of how these values behave[10] by considering the expected value of $\ln(B_S)$ as functions of the sample size for various values of the effect size (see Fig. 2). To get an idea of what these values mean, it may be helpful to note that $\ln(B_S) = 3$ corresponds to $B_S \approx 20$. The figure clearly shows that smaller values of δ correspond to larger maximum values of (the expected value of the logarithm of) B_S, and to larger ranges of sample sizes at which B_S is still substantial. The sample size at which the maximum expected value of $\ln(B_S)$ occurs increases rapidly when δ becomes smaller. By taking the derivative of Eq. (15) with respect to n, the maximum expected value is found to occur at $n+2 = \delta^{-2}$. It becomes arbitrarily large when δ decreases because the expected value at its maximum goes to $\frac{1}{2}\ln(\delta^{-2}-1) \approx -\ln(\delta)$. In other words, the support for the Null hypothesis at small sample sizes and effect sizes is not just nominal in the sense that $Prob_\delta(B_S \leq 1)$ is larger than 1. The support can in fact be very strong in the sense that B_S is large.

It may seem that, for n larger than n_{min}, B_S does behave properly, but in fact it still behaves contrary to how a valid measure of evidential strength ought to behave. After all, $Prob_\delta(B_S \leq 1)$ at sample sizes exceeding n_{min} can still be less than $Prob_\delta(B_S \leq 1)$ as evaluated at n equal to 1. Also, comparing Figs. 1 and 2 shows that the expected value on $\ln(B_S)$ can still be large even at sample sizes much larger than n_{min}. B_S can only be said to be valid once the sample size is large enough that $Prob_\delta(B_S \leq 1)$ exceeds the value it had at n = 1. Let us call that sample size n_{eq}. It shows

[10] Plots of ln(BS) are shown in Tendeiro & Kiers (2020, their figure 9) for different values of δ. These plots were obtained by setting x/σ_0 equal to δ. A similar plot is shown in Keysers et al. (2020, their Extended data Fig. 1b). Keysers et al. provide a log-log plot of $1/B_S$ rather than of B_S and they use a slightly different prior distribution on the alternative hypothesis space, but, qualitatively, the results are the same.

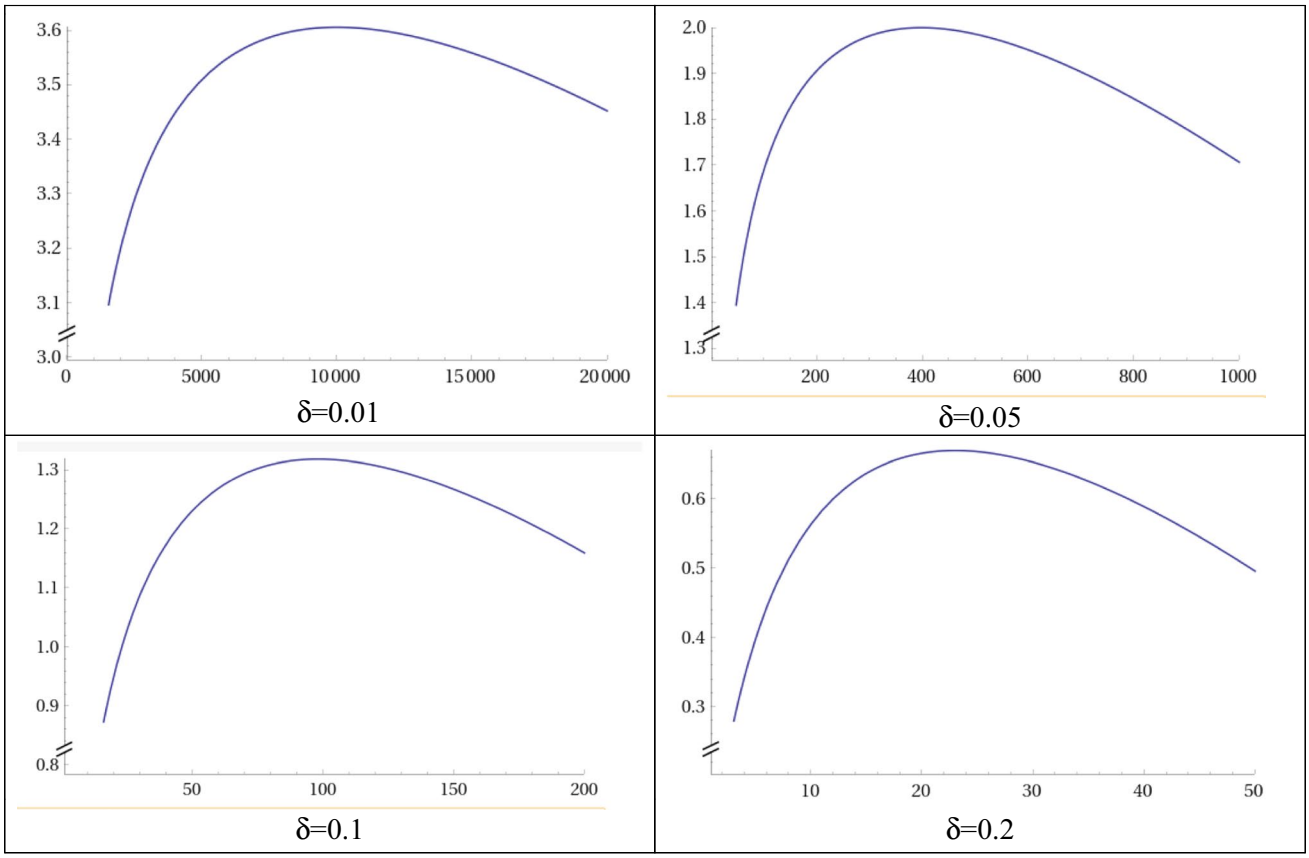

Fig. 2 Expected values of $\ln(B_S)$ as a function of the sample size, for various values of δ. Note the varying vertical and horizontal scales

the true range of sample sizes for which B_S is not a valid measure of evidential strength and can be quite large. Using (13) twice, with b = 1, we find

$$\Phi\left(-\sqrt{k(1)}+\delta\sqrt{n_{eq}}\right)+\Phi\left(-\sqrt{k(1)}-\delta\sqrt{n_{eq}}\right) = \Phi\left(-\sqrt{2\ln(2)}+\delta\right)+\Phi\left(-\sqrt{2\ln(2)}-\delta\right), \quad (16)$$

where $k(1) = (1+n_{eq}^{-1})\ln(1+n_{eq})$. If there is a minimum, n_{min} becomes larger when δ goes to zero (see Fig. 1). Since n_{eq} is larger than n_{min}, it will become larger too when δ goes to zero. To compute n_{eq} using Eq. (16) requires a numerical equation solver. The required computations can be simplified by remembering that, for very small δ, n_{eq} will be large and k(1) will be roughly equal to $\ln(n_{eq})$. In addition, δ can then be ignored compared to $\sqrt{(2\ln(2))}$. As an example, for $\delta = 0.1$, $n_{eq} \approx 276$, which is considerably larger than $n_{min} \approx 35$ found earlier. Likewise, for $\delta = 0.05$, $n_{eq} \approx 1613$, much larger than $n_{min} \approx 100$.

The anomalous dependence of $\text{Prob}_\delta(B_S \leq 1)$ on the sample size shown in Fig. 1 proves that B_S is not stochastically ordered by n. There is a substantial range of sample sizes, all the way up to n_{eq}, for which support for the Null hypothesis does not stochastically decrease with the sample size, even when the Null hypothesis is false. This makes B_S an invalid measure of evidential strength, at least for small effect sizes and for sample sizes that are not too large. Whether B_S can be considered to be a valid measure of evidential strength when the effect size or the sample size is large can, of course, not be determined using arguments such as the ones given here.

Let us return finally to the case that $\text{Prob}_\delta(B_S \leq 1)$ is less than 0.5 when n = 1. The Bayes factor is then likely to be larger than 1 and to support the Null hypothesis, even when the effect size is not zero. Within the Bayesian paradigm, this may be unavoidable. But that implies that, like the P-value, the Bayes factor does not in any obvious way provide support *for* the Null hypothesis. For any sample size, there is a range of effect sizes that cannot be meaningfully distinguished from δ = 0. This range can be estimated as follows.

Assume that n is fixed at some value that is not too small. If we then investigate what happens to $\text{Prob}_\delta(B_S \leq 1)$ when we vary the effect size δ, we find that, at some effect size δ_{max}, the probability equals 0.5 and increases with increasing effect size. Figure 1 demonstrates that such a δ_{max} always

exists because, for any sample size n, we can find an effect size such that n_{min} equals n. At that sample size, $Prob_{\delta}(B_S \leq 1)$ is less than 0.5 and, to get a larger probability, the effect size has to increase. The maximum effect size can be obtained from Eq. (13), because, whatever the sign of δ, one of the two cumulative distributions on the right will be much smaller than the other one. In other words, we can write approximately $\Phi(-\sqrt{k} + |\delta_{max}|\sqrt{n}) \approx 0.5$. That is, $-\sqrt{k}+|\delta_{max}|\sqrt{n} = 0$, or

$$\delta^2_{max} \approx \frac{1}{n} \ln(n). \tag{17}$$

This range can be quite large. For example, when n = 100, $\delta_{max} \approx 0.2$, a not uncommon effect size (Szucs & Ioannidis, 2017)).

Conclusions and comments

P-values are statistics that measure the discrepancy between the Null hypothesis and the actual state of affairs. They are valid measures of evidential strength in the sense that they behave the way we intuitively think measures of evidential strength ought to behave: when the Null hypothesis is false, P-values are stochastically ordered by both the effect size and the sample size.

Of course, the present definition of validity of measures of evidential strength does not imply that P-values are reliable measures of that strength. As is well known, they have many shortcomings. For example, by their very nature they cannot provide evidence for the Null hypothesis. Also, they do not incorporate either the prior likelihood of the Null hypothesis or the power of hypothesis tests to detect discrepancies between the Null hypothesis and reality.

These shortcomings have led many researchers to consider Bayes factors as alternative measures of evidential strength. They are valid measures in the sense that they are stochastically ordered by the effect size, at least in the simple scenario we have studied in this article. But they are not generally valid measures, because, unlike P-values, they need not be stochastically ordered by the sample size.

Bayes factors can in fact be highly misleading: when the Null hypothesis is false, the probability that the Bayes factor is less than 1 should increase with the sample size. Instead, when the sample size is still small and the effect size not too large, it decreases. The decrease continues until, for sample sizes on the order of δ^{-2}, the probability finally starts to increase. It does not become larger than its value at n = 1 until the sample size n is larger than the solution of the implicit Eq. (16). Likewise, the expected value of $\ln(B_S)$ remains high until the sample size is much larger than n_{min}.

This is a serious problem because, when the Null hypothesis is false, a valid measure of evidential strength should indicate stronger evidence against the Null hypothesis when more data are collected. Instead, when data are beginning to be collected and the effect size, although nonzero, is not too large, the Bayes factor shows increasing support for the Null hypothesis (even though it is false), and does not distinguish between a true Null hypothesis and a false one until the sample size is sufficiently large.

The range of effect sizes for which these various problems can occur is admittedly rather small, and it might be argued that the smallness of that range negates the lack of validity of the Bayes factor as a measure of evidential strength. In many research areas, after all, small effect sizes are not important and the Bayes factor being misleading may be unfortunate but not disastrous. That argument, however, is fallacious for a variety of reasons.

First, there are research areas where any deviation from the Null hypothesis, no matter how small, is important. For example, when studying the properties of elementary particles such as their magnetic moments, any discrepancy between the outcomes of theoretical calculations, representing the Null hypothesis, and experiments is of great interest, even when those discrepancies occur in the seventh significant digit (Abi et al., 2021).

Second, a measure of evidential strength is a statistical tool that uses the outcomes of hypothesis tests in order to justify certain conclusions regarding the truth or falsehood of the Null hypothesis. However, each tool has a range of applicability and the user of that tool needs to understand in what range that tool can be used safely. For example, a home thermometer that fails to give correct readings when the temperature drops below 10 °C may still be an acceptable thermometer, but only if the homeowner is aware of that limitation and does not care to know exactly how cold it is when it is colder than 10 °C. Likewise, the problems listed above with Bayes factors are of no consequence to someone who can afford to always work with very large sample sizes. Nevertheless, the limitation should be understood in order to estimate how large a sample is needed to avoid problems.

Finally, even if a researcher really does not care about small effect sizes and can defend that attitude as reasonable and maybe even desirable, she should perform a correct hypothesis test by formulating a small interval Null hypothesis that incorporates the small effect sizes she does not care about. However, it is not clear that switching to an interval Null hypothesis will make the resulting Bayes factor valid. That validity still has to be demonstrated, but I do not address that problem here.

There is another reason for taking this lack of validity seriously: it underlies the Jeffreys-Lindley paradox

(Bartlett, 1957; Lindley, 1957). This paradox is generated by considering a collection of hypothesis tests of the same statistical phenomenon with different sample sizes, but with the same value of m/σ observed in each test. Consequently, the P-value is the same in all the tests, but the Bayes factors differ. In fact, the larger the sample size, the larger the Bayes factor.[11] Furthermore, in the standard account m/σ is sufficiently large and the corresponding P-value sufficiently small that it provides strong evidence against the Null hypothesis. Lindley's setup demonstrates that it is possible for an outcome of a hypothesis test to be such that the P-value strongly suggests that the Null hypothesis is false while the Bayes factor equally strongly suggests that it is true. The standard interpretation of this paradox is that it demonstrates the inadequacy of P-values as measures of evidential strength.

But we can now see that this standard interpretation is erroneous: the cause of the paradox is the Bayes factor, for it may give misleading results for small effect sizes. As has been noted by many commentators on the Jeffreys-Lindley paradox, m needs to decrease when n increases if m/σ is supposed to stay constant. In fact, it needs to be fairly small if the P-value is set at, say, 0.05 (in which case $m\sqrt{n} \approx 2\sigma_0$). A useful estimator of δ is m/σ_0, the Maximum Likelihood estimator, and I indicate it by η. η will become small too when the sample size increases because it is approximately equal to $2/\sqrt{n}$ when the P-value is set to 0.05.

That the Bayes factor may give misleading results when the effect size is small can now be shown in a number of ways. First, $\delta_{max}{}^2$ is roughly equal to $n^{-1}\ln(n)$ (Eq. (17)), which is considerably larger than η^2 when n is large.[12] In other words, the estimated effect size η is well within the range of effect sizes where the Bayes factor gives misleading results.

Second, n_{eq} depends on the effect size and becomes very large when the latter becomes small. Let us consider a specific hypothesis test in the collection of tests in the Jeffreys-Lindley paradox with a large sample size n. It turns out that, for sufficiently small δ, n is smaller than n_{eq}, as can be demonstrated using Eq. (16). Replacing δ by $\eta = 2/\sqrt{n}$, we get

$$\Phi\left(-\sqrt{k}+2\sqrt{\left(\frac{n_{eq}}{n}\right)}\right)+\Phi\left(-\sqrt{k}-2\sqrt{\left(\frac{n_{eq}}{n}\right)}\right)=2\Phi\left(-\sqrt{2\ln 2}\right). \tag{18}$$

Eq. (18) is still an equation for n_{eq}, but now one in which the effect size is labeled by the corresponding sample size in the collection of hypothesis tests. With some straightforward numerical experimentation, we find that n_{eq} is larger than n for n greater than roughly 1,500 (corresponding to $\eta \approx 0.05$), and that the ratio n_{eq}/n increases with increasing n (that is, decreasing η). In other words, n as defined in the Jeffreys-Lindley paradox and in the truly paradoxical limit of very small η, is well within the range of sample sizes for which the Bayes factor is invalid as a measure of evidential strength.

What exacerbates the paradox is that, for increasingly small values of the effect size (increasingly large values of the sample size), typical values of B_S become arbitrarily large (compare Fig. 2). In other words, not only do the Bayes factors in the tests in the Jeffreys-Lindley paradox nominally support the Null hypothesis, contrary to the verdict of the P-values, they do so increasingly strongly when the sample size increases and the estimated effect size becomes very small.

The Bayes factor as employed in the Jeffreys-Lindley paradox is invalid. It supports the Null hypothesis even though the latter may be false, and it supports it to an anomalous extent. The paradox is not an argument against the validity of P-values as measures of evidential strength. It merely illustrates the misleading behavior of Bayes factors when the Null hypothesis is sharp and the actual effect size is small.

The lack of validity of the Bayes factor extends to Bayesian statistics in general. Oftentimes, when in doubt about the meaning of a Bayes factor, switching to the posterior probability of the Null hypothesis will clarify that meaning. Such a switch does not help with the lack of validity of the Bayes factor, however, because the posterior odds of the Null hypothesis are equal to the Bayes factor multiplied by a constant factor, the prior odds of that hypothesis. As a result, whatever misleading behavior is shown by the Bayes factor will merely be repeated by the posterior odds: they will be as misleading as the Bayes factor itself.

The attraction of Bayesian statistics is that it provides for a coherent way of updating one's beliefs upon the acquisition of more information. But, as the case of the sharp Null hypothesis shows, coherency is not the same as reliability. The posterior probability, after outcome m has been obtained, is coherent and may very well correctly reflect what we ought to believe once that outcome was obtained. But the Bayesian paradigm does not guarantee that we are better informed about the actual state of the world upon the acquisition of the new data. It is true that asymptotically – that is, for a sufficiently large sample size – the Bayes factor and the posterior probability will show that the Null hypothesis is false if it is indeed false, but initially, for small to moderately large sample sizes, the data provide increasing support for the Null hypothesis, even when it is false.

[11] Compare Keysers (2020 Extended data Fig.1a), which shows $1/B_S$ as a function of n for different values of m/σ.

[12] To be specific, when $n >> e4 \approx 55$.

Appendix

Both $\text{Prob}_\delta(P_S \le p)$ and $\text{Prob}_\delta(B_S \le b)$ have the form $\Phi(h+\delta\sqrt{n}) + \Phi(h-\delta\sqrt{n})$, where h is a negative function of either p or b, and, in the case of the Bayes factor, also of the sample size n. The derivative of this expression with respect to δ is

$$\frac{d}{d\delta}\left(\Phi\left(h+\delta\sqrt{n}\right)+\Phi\left(h-\delta\sqrt{n}\right)\right) = f_0\left(h+\delta\sqrt{n}\right)\sqrt{n} - f_0\left(h-\delta\sqrt{n}\right)\sqrt{n}, \tag{A.1}$$

where

$$f_0(z) \quad = \quad \frac{1}{\sqrt{(2\pi)}} e^{-z^2/2}$$

is the standard normal distribution function. When we use the explicit expression for that distribution function, the right-hand side of Eq. (A.1) becomes

$$\sqrt{\frac{n}{2\pi}} e^{-(h^2+\delta^2 n)/2}\left(e^{-h\delta\sqrt{n}} - e^{h\delta\sqrt{n}}\right). \tag{A.2}$$

Since h is negative, this expression is positive when δ is positive and negative when δ is negative, showing that $\text{Prob}_\delta(P_S \le p)$ and $\text{Prob}_\delta(B_S \le b)$ become larger when |δ| becomes larger.

The derivative with respect to n is more complex because h may be a function of n. I consider only the case of the Bayes factor, in which case $h = -\sqrt{k}$ (see Eq. (12)). The derivative of $\text{Prob}_\delta(B_S \le b)$ is then seen to be

$$f_0\left(-\sqrt{k}+\delta\sqrt{n}\right)\left(\frac{-1}{2\sqrt{k}}k' + \frac{\delta}{2\sqrt{n}}\right) + f_0\left(-\sqrt{k}-\delta\sqrt{n}\right)\left(\frac{-1}{2\sqrt{k}}k' - \frac{\delta}{2\sqrt{n}}\right), \tag{A.3}$$

with k' the derivative of k with respect to n.

We are interested in the value of n for which this derivative vanishes, that is, the value of n for which $\text{Prob}_\delta(B_S \le b)$ changes from decreasing into increasing. We find this value by setting the derivative to 0. Using the explicit expression for f_0 as in the derivation of Eq. (A.2), we find that the location of the minimum value of $\text{Prob}_\delta(B_S \le b)$ is determined by

$$(e^a + e^{-a})\frac{k'}{\sqrt{k}} = (e^a - e^{-a})\frac{\delta}{\sqrt{n}}, \tag{A.4}$$

with $a = \sqrt{k}\delta\sqrt{n}$.

A minimum occurs at a solution of Eq. (A.4), but such a solution need not exist for n at least equal to 1. If we set n = 1, we determine the maximum effect size for which a minimum does in fact occur. At larger effect sizes, no minimum occurs at non-trivial values of the sample size ($n \ge\varepsilon 1$). When b = 1 and n = 1, k = 2*ln(2) and k' = 1-ln(2). Numerically solving Eq. (A.4) then gives $|\delta| \approx 0.496$.

References

Abi, B., et al. (2021). Measurement of the positive muon anomalous magnetic moment to 0.46 ppm. *Physical Review Letters, 126*, 141801.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44*, 533–553.

Camerer, C. F., et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior, 2*, 637–644.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*, 207–218.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE, 11*(2), e0149794.

Fisher, S. R. A. (1973). *Statistical Methods for Research Workers* (14th ed.). Hafner Publishing Company.

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine, 130*, 1005–1013.

Goodman, S.(2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology, 45*, 134–140.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*, 337–350.

Hubbard, R., & Lindsay, R. M. (2008). Why *P* values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*, 69–88.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773–795.

Keysers, C., Gazzola, V., & Wagenmakers, E. J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience, 23*, 788–799.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18.

OSC. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

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

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797.

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24*, 774–795.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1–2.

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

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. (2008). Bayesian Versus Frequentist Inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses* (pp. 181–210). Springer.