Archive for History of Exact Sciences (2023) 77:25–72
https://doi.org/10.1007/s00407-022-00298-3

History and nature of the Jeffreys–Lindley paradox

Eric-Jan Wagenmakers[1] · Alexander Ly[2]

Received: 1 December 2021 / Published online: 26 August 2022
© The Author(s) 2022

Abstract

The Jeffreys–Lindley paradox exposes a rift between Bayesian and frequentist hypothesis testing that strikes at the heart of statistical inference. Contrary to what most current literature suggests, the paradox was central to the Bayesian testing methodology developed by Sir Harold Jeffreys in the late 1930s. Jeffreys showed that the evidence for a point-null hypothesis $\mathcal{H}_0$ scales with $\sqrt{n}$ and repeatedly argued that it would, therefore, be mistaken to set a threshold for rejecting $\mathcal{H}_0$ at a constant multiple of the standard error. Here, we summarize Jeffreys's early work on the paradox and clarify his reasons for including the $\sqrt{n}$ term. The prior distribution is seen to play a crucial role; by implicitly correcting for selection, small parameter values are identified as relatively surprising under $\mathcal{H}_1$. We highlight the general nature of the paradox by presenting both a fully frequentist and a fully Bayesian version. We also demonstrate that the paradox does not depend on assigning prior mass to a point hypothesis, as is commonly believed.

Communicated by Jeremy Gray.

This research was supported by a Netherlands Organisation for Scientific Research (NWO) Grant to EJW (016.Vici.170.083), as well as an Advanced ERC Grant to EJW (743086 UNIFY). Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. Correspondence should be sent to Eric-Jan Wagenmakers, University of Amsterdam, Nieuwe Achtergracht 129 B, 1018 WT Amsterdam, The Netherlands. E-mail may be sent to EJ.Wagenmakers@gmail.com.

✉ Eric-Jan Wagenmakers
EJ.Wagenmakers@gmail.com

[1] University of Amsterdam, Amsterdam, The Netherlands

[2] Centrum Wiskunde and Informatica, University of Amsterdam, Amsterdam, The Netherlands

Introduction

The Jeffreys–Lindley paradox (e.g., Bartlett 1957; Jeffreys 1935; Lindley 1957) refers to the fact that, as sample size increases indefinitely and the p value remains constant at any non-zero value (e.g., $p = 0.005$), we inevitably arrive at a conflict between p values and Bayes factors, in the sense that the p value suggests that the point-null hypothesis $\mathcal{H}_0$ should be rejected, whereas the Bayes factor indicates that $\mathcal{H}_0$ decisively outpredicts the alternative hypothesis $\mathcal{H}_1$. This conflict will arise regardless of the p value under consideration and regardless of the prior distribution on the test-relevant parameter in $\mathcal{H}_1$ (under regularity conditions). Thus, a frequentist statistician may specify any non-zero α-level whatever, a Bayesian statistician may specify any continuous prior distribution on the test-relevant parameter under $\mathcal{H}_1$, and a third party could then infallibly construct data sets for which the point-null hypothesis $\mathcal{H}_0$ would be simultaneously rejected by the frequentist and accepted by the Bayesian.[1]

Although the paradox is often associated with Lindley (1957), and sometimes with Bartlett (1957), it was already derived, demonstrated, explained, and emphasized by Sir Harold Jeffreys in his articles and books on Bayesian hypothesis testing from the second half of the 1930s (i.e., Jeffreys 1935, p. 205; 1936a, p. 345 and p. 353; 1936b, p. 417; 1937a, p. 494; 1937b, pp. 250–251 and p. 259; 1937c, p. 1004; 1938a, pp. 377–381; 1938b, p. 161; 1938c, p. 148; 1938e, p. 114; 1938d, p. 310; 1939, pp. 194–195 and pp. 359–360—see p. 248 and pp. 435–436 in 1961). The paradox has remained a source of inspiration for statisticians and philosophers alike (e.g., Bernardo 1980, 2011; Berrar and Dubitzky 2017; Colquhoun 2019; Cornfield 1966; Cousins 2017; Edwards et al. 1963; Good 1980a; Jefferys 1990; Leamer 1978; Nasir et al. 2020; Ormerod et al. 2017; Robert 2014; Royall 1986; Senn 2001; Shafer 1982; Spanos 2013; Sprenger 2013; Villa and Walker 2017; Yin and Shi 2020; Wagenmakers 2007; Zellner 1971/1996, Chapter 10), but we believe that the neglect of Jeffreys's original work on the paradox has led to considerable confusion. Indeed, the paradox has caused statisticians to question the usefulness of Bayesian statistics as a whole (e.g., Shafer 1982; Spanos 2013), to reject Bayes factor hypothesis testing in favor of Bayesian parameter estimation (e.g., Bernardo 1980), and to develop alternative forms of Bayesian hypothesis testing (e.g., Aitkin 1991; Andrews 1994; de Bragança Pereira et al. 2008; Kamary et al. 2014; Vehtari et al. 2017). We do not wish to disparage this work but we do believe the original arguments by Jeffreys have been underappreciated if not entirely forgotten (for a notable exception see Cousins 2017).

The goal of this paper is, therefore, threefold. First, we aim to demonstrate the extent to which the paradox had already been treated by Jeffreys prior to the 1957 articles by Lindley and by Bartlett. The appendix lists Jeffreys's discussions of the paradox after 1957. Contrary to popular belief, our analysis reveals that the paradox played a central role in Jeffreys's system of Bayes factor hypothesis tests, and did so from the outset. Although Jeffreys often downplayed the practical ramification of the paradox for moderate sample sizes, he also repeatedly stressed that his Bayesian hypothesis

[1] Note that the Jeffreys–Lindley paradox is a veridical paradox: it is an *apparent* contradiction (e.g., Jeffreys 1938d, p. 310). A sufficiently knowledgeable and confident statistician may, therefore, rightly proclaim that the Jeffreys–Lindley paradox is not at all paradoxical (to them). Veridical paradoxes are in the eye of the beholder. See also Cousins (2017) and Pericchi (2011).

test depended not just on how many standard errors the maximum likelihood estimate is away from zero (as in the classical method) but also involved a $\sqrt{n}$ term. Crucially, this means that the criterion for "significance" in Jeffreys's tests is not given by a constant multiple of the standard error. Jeffreys presented almost every Bayes factor he proposed in the same form, with a $\sqrt{n}$ factor outside of an exponential term and a multiple-of-the-standard-error factor inside the exponential term; these expressions leave no doubt about the large-n conflict between Jeffreys's Bayes factors and p values. Moreover, throughout his published work Jeffreys highlighted the effect of sample size on his tests by means of tables; he discussed the reasons for the appearance of the $\sqrt{n}$ term, and he explicitly stated that including this term was both desirable and dictated by the application of Bayesian probability theory to the problem of hypothesis testing. The common notion that Jeffreys mentioned the paradox only in passing is, therefore, seriously incorrect.

The second goal of this paper is to revive Jeffreys's original line of argumentation, which was that the paradox, instead of being "certainly embarrassing to the Bayesian" (Szabó and van der Vaart 2019, p. 17), or "difficult to accept" (Bernardo 2009, p. 174) was rather the inevitable consequence of any reasonable definition of evidence. In other words, Jeffreys felt that no sensible measure of evidence can be based on a constant multiple of the standard error, independent of sample size.

The third goal of this manuscript is to highlight the general nature of the paradox. Specifically, we demonstrate that the paradox can be given both a fully frequentist interpretation and a fully Bayesian interpretation. Moreover, and in contrast to popular belief, we show that the essence of the paradox does not depend on the fact that the model comparison involves a sharp null hypothesis $\mathcal{H}_0$ with a point-mass at zero. Instead, the paradox will manifest itself for any Bayes factor where the prior distribution for effect size under the sceptic's $\mathcal{H}_0$ is more heavily concentrated around zero than the prior distribution for effect size under the proponent's $\mathcal{H}_1$, a condition so mild as to be almost tautological.

Statistical background

In the early twentieth century, Sir Ronald Fisher promoted the idea of null hypothesis significance testing (NHST) using p values. Informally, the p value is the chance under the null hypothesis of finding a test statistic at least as extreme as the one obtained (e.g., Wasserstein and Lazar 2016). The idea of NHST is loosely similar to that of a proof by contradiction: to show that there exists an effect, one assumes the opposite (i.e., the null model $\mathcal{H}_0$) and demonstrates that the data make this assumption unlikely (Wagenmakers et al. 2017). In NHST, the data are believed to cast doubt on $\mathcal{H}_0$ when the obtained p value is sufficiently small. Fisher deemed a p value of 0.05 or lower sufficient grounds to reject the null hypothesis. In Chapter 3 of *Statistical Methods for Research Workers*, Fisher discusses the normal distribution and notes that

> "The value for which P = .05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus

> formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty." (p. 45, Fisher 1934)

Despite constant criticism from within the statistical community, Fisher's rule has since been institutionalised in academic practice. Researchers routinely conclude that results constitute a significant deviation from the null model whenever $p < 0.05$, that is, whenever the observed value of the test statistic falls more than two standard errors away from the value postulated by the null model.[2]

As an alternative to p value significance testing, Sir Harold Jeffreys developed and advocated a series of Bayesian hypothesis tests whose key outcome is now known as the *Bayes factor* (e.g., Kass and Raftery 1995). The philosophical foundation of the Bayes factor goes back to Jeffreys's work with Dorothy Wrinch in the early 1920s (Wrinch and Jeffreys 1919, 1921, 1923), but the concrete statistical development was initiated and largely completed by Jeffreys in the second half of the 1930s (e.g., Jeffreys 1935, 1939; for a modern appreciation, see Etz and Wagenmakers 2017; Howie 2002; Ly et al. 2016a, b, 2020; Robert et al. 2009). To learn from data Jeffreys proposed to assign prior model probabilities $P(\mathcal{M}_0)$ and $P(\mathcal{M}_1)$ to the null hypothesis and the alternative hypothesis, respectively. In light of data y, these probabilities can then be updated to posterior model probabilities using Bayes' rule. The ratio of the posterior models probabilities then leads to

$$\underbrace{\frac{P(\mathcal{M}_1 \mid y)}{P(\mathcal{M}_0 \mid y)}}_{\text{posterior model odds}} = \underbrace{\frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_0)}}_{\mathrm{BF}_{10}(y)} \times \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{prior model odds}}, \tag{1}$$

where $p(y \mid \mathcal{M}_k)$ is known as the marginal likelihood, that is, the likelihood function of the free parameters θ_k under $\mathcal{M}_k$ integrated out with respect to a prior distribution $\pi(\theta \mid \mathcal{M}_k)$:

$$p(y \mid \mathcal{M}_k) := \int_{\Theta_k} f(y \mid \theta_k, \mathcal{M}_k)\pi(\theta_k \mid \mathcal{M}_k)\, \mathrm{d}\theta_k. \tag{2}$$

The purpose of the Bayes factor $\mathrm{BF}_{10}(y)$ is to "grade the decisiveness of the evidence" (Jeffreys 1961, p. 432). In contrast to the p value, this pertains to both $\mathcal{M}_0$ and $\mathcal{M}_1$. Specifically, a $\mathrm{BF}_{10}(y)$ much larger than 1 indicates evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$; a $\mathrm{BF}_{10}(y)$ near 0 indicates evidence for $\mathcal{M}_0$ over $\mathcal{M}_1$ (i.e., "evidence of absence"); and a $\mathrm{BF}_{10}(y)$ near 1 indicates that the data are insufficiently diagnostic ("absence of evidence"; Keysers et al. 2020). Note that for the construction of a Bayes factor a *pair of priors* needs to be selected, one for each model. Jeffreys did so with great care for

[2] As noted by Cornfield (1966, pp. 18–19), the α-level (i.e., the critical value below which the p value is deemed to indicate a statistically significant deviation from the null model) is often viewed as a "universal yardstick" and the underlying intuition is that "All hypotheses rejected at the same critical level have equal amounts of evidence against them." (i.e., Cornfield's "α-postulate", which he sought to undercut).

various statistical models and documented the results in his magnum opus *Theory of Probability* (Jeffreys 1939, 1948, 1961).[3] As will become apparent below, one of the defining features of the Bayes factor is that it does not depend on a constant multiple of the standard error. The additional involvement of sample size is what generates the paradox.

To set the stage, we start by discussing the 1957 article from Dennis Lindley. We then list Jeffreys's work on the paradox as expressed in a series of 16 articles and two books from 1935 to 1957. In order to drive home the point that the paradox was central to Jeffreys's tests, our treatment aims to be comprehensive. The included quotations are unusual both in their number and in their length, but we believe this is necessary in order to (1) irrevocably refute the common misconception that Jeffreys had ignored or neglected the paradox; (2) support the claim that the paradox in fact presents a defining feature of the Bayes factor hypothesis test; (3) demonstrate the different ways in which Jeffreys explained why a measure of evidence cannot depend on a constant multiple of the standard error.

The 1957 contribution by Lindley

Dennis Lindley (1957) started his famous article *A statistical paradox* as follows:

> "An example is produced to show that, if H is a simple hypothesis and x the result of an experiment, the following two phenomena can occur simultaneously:
> (i) a significance test for H reveals that x is significant at, say, the 5% level;
> (ii) the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as 95%.
> Clearly the common-sense interpretations of (i) and (ii) are in direct conflict. The phenomenon is fairly general with significance tests and casts doubts on the meaning of a significance level in some circumstances." (p. 187)

Later in the article, Lindley elaborates:

> "Now in our example we have taken situations in which the significance level is fixed because, as explained above, we wish to see whether its interpretation as a measure of lack of conviction about the null hypothesis does mean the same in different circumstances. The Bayesian probability is all right, by the arguments above; and since we now see that it varies strikingly with n for fixed significance level, in an extreme case producing a result in direct conflict with the significance level, the degree of conviction is not even approximately the same in two situations with equal significance levels. *5% in to-day's small sample does not mean the same as 5% in to-morrow's large one.*" (Lindley 1957, p. 189, italics added for emphasis)

Lindley explicitly acknowledges the fact that Jeffreys noted the paradox earlier:

[3] For his Bayes factor tests, Jeffreys proposed prior distributions that did not reflect strong advance knowledge and that obeyed several logical desiderata (e.g., Bayarri et al. 2012; Consonni et al. 2018; Ly et al. 2016a). Note that the paradox manifests itself regardless of how the prior distribution is defined, under regularity conditions.

> "The paradox is not, in essentials, new, although few statisticians are aware of it. The difference between the two approaches has been noted before by Jeffreys (see, in particular, 1948, Appendix), who is the originator of significance tests based on Bayes's theorem and a concentration of prior probability on the null value. But Jeffreys is concerned to emphasize the similarity between his tests and those due to Fisher and the discrepancies are not emphasized." (p. 190)

We believe that Lindley's assessment requires revision. Below we demonstrate that Jeffreys repeatedly emphasized the theoretical difference between the two approaches throughout 16 articles and two books published from 1935 to 1957.

The contributions by Jeffreys from 1935 to 1957

In order to follow the quotations from the works cited below, note that Jeffreys uses K to refer to the Bayes factor for $\mathcal{H}_0$ over $\mathcal{H}_1$, that is, $K \equiv \text{BF}_{01}$. In addition, Jeffreys denotes $\mathcal{H}_0$ by q and $\mathcal{H}_1$ by $\sim q$ or q'. For a modern-day reader, it may be confusing that Jeffreys used Greek letters for observed data—in particular, he often used θ to denote observed data rather than an unobserved parameter. Finally, Jeffreys often conditioned all probability statements on background knowledge, which he denoted by h or H ('history')—not to be mistaken for the modern-day use of H for 'hypothesis'. A complete translation of Jeffreys's notation can be found in Table D.4 of Ly et al. (2016a).

1. The 1934 letter to Fisher

The first hint that Jeffreys is interested in developing a Bayesian significance test is found in a 1934 letter to Fisher:

> "The sort of thing that bothers me is this. In seismology we get times of transmission to various distances, and fit a polynomial of degree 3, say, to them. The significance of the last term really involves the prior probability that such a term will be present. The usual thing is to keep it if it is some arbitrary multiple of its standard error, but I think it ought to be possible to frame a rule with *some* sort of argument behind it..."
> Sir Harold Jeffreys, in a letter to Sir Ronald Fisher, 1934 (Bennett 1990, p. 156)[4]

The Bayes factor rule that Jeffreys later derived turned out to be different from "the usual thing": the strength of the Bayes factor is not proportional to a constant multiple of the standard error, but also involves sample size. This is the paradox. Thus, the 1934 letter to Fisher shows that the seeds of the paradox were sown even before Jeffreys had started to develop his tests.

[4] Curiously, the letter as given in Bennett is incomplete. The original, complete letter can be found at https://digital.library.adelaide.edu.au/dspace/bitstream/2440/67780/109/1934-03-21.pdf.

2. The 1935 article *Some tests of significance, treated by the theory of probability*

This was the first article in which Jeffreys developed a series of concrete Bayes factor hypothesis tests. The introductory paragraph immediately sets up the key issue, in similar fashion to the 1934 letter to Fisher:

> "It often happens that when two sets of data obtained by observation give slightly different estimates of the true value we wish to know whether the difference is significant. The usual procedure is to say that it is significant if it exceeds a certain rather arbitrary multiple of the standard error; but this is not very satisfactory, and it seems worth while to see whether any precise criterion can be obtained by a thorough application of the theory of probability." (Jeffreys 1935, p. 203)

First Jeffreys turns to a comparison of two proportions:

> "Suppose that two different large, but not infinite, populations have been sampled in respect of a certain property. One gives x specimens with the property, y without; the other gives x' and y' respectively. The question is, whether the difference between x/y and x'/y' gives any ground for inferring a difference between the corresponding ratios in the complete population." (Jeffreys 1935, p. 203)

Jeffreys (p. 204, Eq. 11) then shows that the posterior odds for q over $\sim q$ is given by

$$\frac{P(q \mid \theta, h)}{P(\sim q \mid \theta, h)} = \frac{(x+x')!\,(y+y')!\,(x+y+1)!\,(x'+y'+1)!}{x!\,y!\,x'!\,y'!\,(x+x'+y+y'+1)!},$$

where θ denotes the observed data and h ('history') denotes background knowledge. For large samples, Jeffreys obtains the following approximation (p. 205, Eq. 15):

$$\frac{P(q \mid \theta, h)}{P(\sim q \mid \theta, h)} \sim \left\{\frac{(x+x'+y+y')(x+y)(x'+y')}{2\pi(x+x')(y+y')}\right\}^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\frac{(x+x'+y+y')(xy'-x'y)^2}{(x+x')(y+y')(x+y)(x'+y')}\right\}.$$

Jeffreys then continues and identifies the phenomenon that lies at the heart of the paradox:

> "The theory therefore shows that a small difference between the sampling ratios may establish a high probability that the ratios in the main populations are equal, while a large one may show that they are different. This is in accordance with ordinary practice, but has not, so far as I know, been related to the general theory before. *In one respect, however, there is a departure from ordinary practice*. It would be natural to define a standard error of $xy' - x'y$ in terms of the coefficient of its square in the exponential; but the range of values of the exponent that make the ratio of the posterior probabilities greater than 1 is not a constant, since it depends on the outside factor, which increases with the sizes of the samples.

Table 1 Table reproduced from Jeffreys (1935, p. 205)

$x + y$	$P(q)/P(\sim q)$	$x' - y'$	$(x' - y')/(x + y)^{\frac{1}{2}}$
40	3.57	14.3	2.26
100	5.65	26.4	2.64
200	7.97	40.8	2.89
400	11.3	61.5	3.07
1000	17.8	107.3	3.39
10,000	56.4	401	4.01
100,000	178	1440	4.57

> This variability is of course connected directly with the fact that agreement between the two populations becomes more probable if the samples are large and the difference of the sampling ratios are small; when the ratio is large at $xy' - x'y = 0$, a larger value of the exponent is obviously needed to reduce the product to unity.
> Some numerical values are given by way of illustration. In each case $x = y$, $x' + y' = x + y$, but in general $x' \neq y'$. The table gives $x + y$, the maximum value of the ratio of the posterior probabilities, and that of $x' - y'$ needed to make the ratio equal to unity.

The ratio of the critical value of $x' - y'$ to $(x + y)^{\frac{1}{2}}$ is given in a further column to show how little it varies when the sizes of the samples change by a factor of 2500." (Jeffreys 1935, pp. 205–206; italics added for emphasis)

Later on Jeffreys draws the same conclusion for a test between two means with the standard error known:

> "It is therefore not correct to say that a systematic difference becomes significant when it reaches any constant multiple of its standard error" (Jeffreys 1935, p. 207)

Jeffreys returns to this theme several times throughout the article, for different tests (e.g., correlation, periodicity). The overall impression is that in the 1935 article Jeffreys emphasized the theoretical aspect of the paradox but at the same time downplayed its practical ramifications.

3. The 1936 article *On some criticisms of the theory of probability*

One year later, Jeffreys again raises the key issue:

> "The results show that the probability that such a term is needed is increased or decreased according as the coefficient is more or less than a certain multiple of its standard error; *the multiple needed, however, increases with the number of observations.*" (Jeffreys 1936a, p. 345; italics added for emphasis)

Jeffreys elaborates and discusses the problem of a least-squares fit to a regression equation:

Table 2 Table reproduced from Jeffreys (1936a, p. 352)

$n.$	$b/\sigma_b.$	$n.$	$b/\sigma_b.$	$n.$	$b/\sigma_b.$
5	1.07	200	2.20	10,000	2.96
10	1.36	500	2.40	20,000	3.07
20	1.59	1000	2.54	50,000	3.22
50	1.86	2000	2.67	100,000	3.33
100	2.04	5000	2.84		

Here b/σ_b indicates the ratio of a least squares point estimate b to its standard error σ_b that results in a Bayes factor of 1. This critical ratio increases with n

> "When one unknown is determined at a time by least squares the criterion[5] that the last determined shall be supported by the observations is that
>
> $$\frac{b^2}{\sigma_b^2} > \log_e \frac{2n}{\pi},$$
>
> where n is the number of observations." (Jeffreys 1936a, p. 352)

As b is the least-squares parameter point estimate, and σ_b is the standard error, the equation shows that for support to remain constant as n increases, the multiple of the standard error will need to increase as well. To underscore this point, Jeffreys provides a table, reproduced here as Table 2, which "gives the critical ratios that an unknown found by least squares from n observations shall be supported by the observations." (Jeffreys 1936a, p. 352). For instance, when $n = 10$, we have $b/\sigma_b = \sqrt{\log_e 20/\pi} \approx 1.36$, and for $n = 100$, we have $b/\sigma_b = \sqrt{\log_e 200/\pi} \approx 2.04$.

Jeffreys then explains the consequences of this sample-size induced increase of the critical ratio, and explicitly discusses the paradox:

> "The usual practice has been to regard a departure from a simple law as genuine if it amounts to some constant multiple of the standard error, usually 2 or 3 times. *The ratio given above is not constant, but depends on the number of observations.* If a ratio of 2 or 3 is really needed when the number is small, it expresses a prior belief in the simple law to the extent of saying that the odds in its favour are 6 to 1 or 90 to 1, or else a criterion of convenience that we must not complicate future computations except for specially strong reasons. In either case corresponding, but smaller, increases would be needed throughout the table. *When the number of observations is large the critical ratio exceeds the arbitrary standard, which will thus for* $100,000$ *observations lead to coefficients between 2 and* 3.33 *times their standard errors being accepted as genuine, when in fact the observations render them less probable than before.* Thus there will be mistakes in all cases where there is no real departure and yet the computed departure is between 2 and 3.33 times its standard error. Fortunately the latter event does not occur very

[5] For this result, Jeffreys includes a footnote to Jeffreys (1936b) (relevant pages: 432–440) which was *in press* at the time.

> often; nevertheless it has arisen." (Jeffreys 1936a, pp. 353–354; italics added for emphasis)

Jeffreys concludes the article by demonstrating and explaining the paradox in the field of astronomy with a concrete example.[6] In a regression model for the motion of the node of Venus, there were 12, 319 observations. The Bayes factor is about 6 in favor of $\mathcal{H}_0$. However,

> "On the usual theory the probability of an accidental variation exceeding 3.5 times its standard error is 4×10^{-4}, and the anomaly would have to be taken as real. Such a value will in any case be exceptional, but with the actual number and accuracy of the observations it is more exceptional on the hypothesis that it is real than on the hypothesis that it is due to accidental error." (Jeffreys 1936a, p. 445)

4. The 1936 article *Further significance tests*

In the same year, Jeffreys again stresses the same issue:

> "The results are usually of the form $\alpha n^{\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$, where n is the number of observations and x is the difference found statistically, which may be a difference of two sampling ratios or measurements, a correlation or a harmonic coefficient. σ is the standard error of x as found from the usual statistical theories. α is usually a moderate coefficient. The form of the results can be explained simply in general terms. Suppose that the difference which we are trying to find might have had any value within a range m. It is actually found to be within a certain small range of length τ about x. Then, on the hypothesis that there is a real difference, the probability that the results would be in this range is τ/m. But on the hypothesis that there is no real difference the corresponding probability is $\tau(2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$. Hence by the theorem of inverse probability the probabilities of no real difference and of a real difference are in the ratio $(m/\sigma)(2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2/\sigma^2)$. But if the accuracy of the observations remains constant the standard error of the mean decreases like $n^{-\frac{1}{2}}$; hence the outside factor is of order $n^{\frac{1}{2}}$. (...)
> To put the matter in other words, if an observed difference is found to be of order σ, then on the hypothesis that there is no real difference this is what would be expected; but if there was a real difference that might have been anywhere within a range m it is a remarkable coincidence that it should have happened to be in just this particular stretch near zero. On the other hand if the observed difference is several times its standard error it is very unlikely to have occurred if there was no real difference, but it is as likely as ever to have occurred if there was a real difference. In this case beyond a certain value of x the more remarkable coincidence is for the hypothesis of no real difference, and as we have to decide from the facts as presented we shall accept the difference. The theory merely

[6] This example also features in later papers, discussed below.

> develops these elementary considerations quantitatively and evaluates the factor α. If $P(q \mid \theta h) > \frac{1}{2}$, we shall expect the difference found to persist with more and more accurate observations; if it is less than $\frac{1}{2}$ we shall expect the estimated difference to diminish.
> *The usual statistical method is to evaluate the observed difference and its standard error, and to say that it is not significant if it is less than a certain constant multiple of this error. No explanation of this rule is given, the probability of the observations being found only on the hypothesis that there is no difference, and not compared with that on the alternative hypothesis. The present method provides an explanation; but the multiple found is not constant, depending on the number of observations* and on the ratio of the standard error of one observation to the whole difference possible, but since it involves these numbers only through the square roots of their logarithms the variation in actual cases is not very large. " (Jeffreys 1936b, p. 417; italics added for emphasis)

These quotations show that in 1935 and 1936, Jeffreys had already discovered, understood, published, emphasized, explained, and illustrated the paradox.

5. The 1937 article *The tests for sampling differences and contingency*

In this article, Jeffreys's final paragraph again describes the phenomenon:

> "Attention is called to the fact that in my tests the ratio of the critical value of a difference to the standard error of the latter varies a little with the number of observations. A difference of twice the standard error may be just significant [in the sense of Jeffreys's Bayes factor test – EWAL] when it rests on five observations, but not when it rests on 100. For application of the tests it is therefore necessary to know the number of observations, and in many cases this is not given explicitly in published work and can be disentangled with great difficulty, if at all. In other words a difference of 1.0 ± 0.5 units may be worth considering further if it rests on five observations each with a standard error of 1.2 units; if it rests on 100 observations each with a standard error of 5 units it is not. This comes from pure probability theory and does not allow for the possibility of systematic error of observation, which might be considered at a later stage and would accentuate the effect." (Jeffreys 1937a, p. 494)

6. The 1937 addenda to the first edition of *Scientific inference*

Jeffreys's book *Scientific Inference* first appeared in 1931, before Jeffreys had started to work on Bayes factors in earnest. A 1937 reissue *Scientific Inference*, however, contains addenda that describe the Bayes factor hypothesis test and a description of the reasoning that underpins the paradox (cf. Jeffreys 1936b above):

> "Suppose we consider as a serious possibility that a quantity x may be zero; denote this proposition by q, with prior probability $\frac{1}{2}$. The proposition that x is not zero is denoted by $\sim q$, also with prior probability $\frac{1}{2}$; but if x is not zero

> it may be anywhere in a range of length m. An actual determination from data θ suggests a value of $x_0 \pm \sigma$. Now, if x is really 0, the probability of finding a mean in a range dx_0 about dx_0 is $\frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(-\frac{x_0^2}{2\sigma^2}\right) dx_0$. But if x is not 0, the probability that it would be in such a range is dx_0/m. Given then that x_0 has actually been found in such a range, the posterior probabilities of q and $\sim q$ are in the ratio of these two expressions, namely
>
> $$\frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \frac{m}{\sqrt{(2\pi)}\sigma} \exp\left(-\frac{x_0^2}{2\sigma^2}\right).$$
>
> When x_0 is large compared with σ, this is small, q has a small posterior probability, and we can assert with confidence that x is different from zero. But σ, the standard error of the mean, is proportional to $n^{-\frac{1}{2}}$, where n is the number of observations; hence if n is large the first factor is large of order $\sqrt{n}$, and the ratio will be large if x_0 is less than σ. Thus a discrepancy less than a certain amount increases the probability that the parameter sought is zero; one more than this amount decreases it and indicates that the parameter is needed. In the cases examined the critical value, with ordinary numbers of observations, ranges from about 1.5 to 3 times the standard error, increasing with the number of observations. The larger the number of observations the stronger the support for the simple law $x = 0$ if the empirical value turns out to be within its standard error. To put the argument in words, if x_0 is of order σ, this is what we should expect if x is zero, but if x might be anywhere in a range m it is a remarkable coincidence that it should be in just this one. On the other hand, if x_0 is substantially more than σ, we should not expect it if x is zero, but we should expect it if x is not zero; in both cases we adopt the less remarkable coincidence." (Jeffreys 1937b, pp. 250–251)

A few pages later, Jeffreys provides a concrete example of the paradox (cf. Jeffreys 1936a, p. 445 above):

> "In current statistical practice the word "significance" appears to be used in several different senses, corresponding to different questions, but it is apparently often supposed that they will have the same answers. I have used it in the case where we want to know whether the observations support a new parameter; this is one that regularly occurs, for instance, in astronomy. The multiple of the standard error used to indicate a statistical difference is about the same as my theory gives for ordinary numbers of observations, but it is taken constant. *My fuller theory shows that it should increase somewhat with the number of observations.* I have only once come upon a case where the difference between the criteria would affect the decision, namely the excess motion of the node of Venus, which, if genuine, is inconsistent with Einstein's law of gravitation. It is 3.5 times the standard error, and by the usual rules would have to be taken as real. But the number of observations used is so large that by my rule it is even more likely to be a random error. In fact Sir Arthur Eddington, who does not

> accept the theory of probability, adopted the decision it gives and not that given by his own theory." (Jeffreys 1937b, p. 256, italics added for emphasis)

Jeffreys then draws the explicit comparison to p values:

> "A constant significance limit, in relation to the standard error, would however be equivalent to saying that the prior probability of a zero value varies with the number of observations, which is absurd; or, alternatively, that the chance of a real difference exceeding the standard error is the same no matter how small the standard error is made by increasing the number of observations. Actually, however, my significance limit varies very slowly with the number of observations and with ordinary numbers does not differ much from Fisher's limits based on the arbitrary 5 per cent. and 1 per cent.; in the great majority of actual cases the decisions will be the same. Accordingly it appears that Fisher's practice does not follow from his postulates, but it, or something very like it, follows from mine." (Jeffreys 1937b, p. 259)

It is noteworthy that the two "absurdities" that Jeffreys identifies in this fragment (i.e., as n increases, either lower the probability of $\mathcal{H}_0$ or narrow the prior parameter distribution under $\mathcal{H}_1$) would later be proposed by Robert (1993) (see also Burnham and Anderson 2004) and Bartlett (1957), respectively.

7. The 1937 correspondence with Fisher

The sample-size induced discrepancy between Bayes factor and p values was also noted explicitly in a 1937 letter that Jeffreys wrote to Fisher (note that this example was also presented in Jeffreys 1936a, p. 445 and in Jeffreys 1937b, p. 256, as discussed above):

> "A question has just arisen about the excess motion of the node of Venus. It is 3.5 times the standard error, the probability of a random deviation exceeding which is 0.00041. Eddington says that as it is one of 15 it can be accepted as normal. The p. that one of 15 would exceed 3.5σ is 0.006. What I should like to know from you is whether there is another case on record where a statistician has accepted at sight a deviation beyond your 1% limit as random? (The other 14 give a χ^2 of 15).
> By my test the thing is probably random on account of the large number of observations combined, but there's not much to spare, and the situation would be altered if some *specific* systematic error was before the House."
> Sir Harold Jeffreys, in a letter to Sir Ronald Fisher, 1937 (Bennett 1990, p. 161; italics in original)

Later that year, Fisher replied as follows:

> "I should be inclined, naturally, to accept Eddington's judgement on an astronomical point, especially as your own test seems to confirm it. On the other hand, *prima facie*, i.e. on an assumption ordinarily made, the probability 0.006 is amply small enough to claim significance, and would be used for this pur-

> pose with complete confidence, I have no doubt, if anyone had a theory which required such a deviation."
> Sir Ronald Fisher, in a letter to Sir Harold Jeffreys, 1937 (Bennett 1990, p. 162; italics in original)

Fisher's answer is somewhat ambiguous, but it does appear as if he believed a *p* value of 0.006 to be sufficiently compelling for declaring a deviation significant, regardless of sample size. Instead of pushing Fisher on the issue, Jeffreys's response strikes a conciliatory tone:

> "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful. (...)
> I am writing this because there is a tendency about to attribute what I believe to be an entirely exaggerated idea of our disagreement to us, for which we are both possibly partly responsible, and I think an occasional mention of cases where we agree would be for the good of the subject."
> Sir Harold Jeffreys, in a letter to Sir Ronald Fisher, 1937 (Bennett 1990, pp. 162–163; italics in original)[7]

8. The 1937 article *Modern Aristotelianism: Contribution to Discussion*

In this one-page discussion on the role of induction in science, Jeffreys mentions the common elements in the statistical frameworks advocated by Karl Pearson and Ronald Fisher, and then states:

> "I should expect the decisions by my methods to lead to the correct decisions most rapidly, because the method contains more explicit provision for allowing for the whole of the data; but many rules given by Fisher, and others accepted by him, are of exactly the same form as mine [EWAL: point estimates] and would in practice be used in the same way, while in other cases where there are differences [EWAL: Bayes factors vs. *p* values] the actual limits recommended are such that it would be extremely rarely that the decisions would differ in any specific application, and then we should both be doubtful." (Jeffreys 1937c, p. 1004)

As in the 1935 article, Jeffreys downplays the practical ramifications of the paradox—a theme that will recur in the appendix of Jeffreys's book *Theory of Probability*. In later sections, we speculate about Jeffreys's reasons for doing so.

[7] In a 1983 interview with Dennis Lindley, Jeffreys referred to this exchange as follows: "[the correspondence with Fisher] was after I'd said that on most things we should agree and when we disagreed we would both be doubtful. After that, Fisher and I were great friends." ("Transcription of a Conversation between Sir Harold Jeffreys and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys).

Table 3 Table reproduced from Jeffreys (1938a, p. 379)

n (Fisher's $n + 1$)	K	n (Fisher's $n + 1$)	K
5	0.610	9	0.519
6	0.551	10	0.522
7	0.529	20	0.612
8	0.520	30	0.719

9. The 1938 article *The comparison of series of measures on different hypotheses concerning the standard errors*

In this article, Jeffreys (1938a, p. 378) gives the Bayes factor in the case of a t test:

$$\begin{aligned} K &= \left(\frac{2n}{\pi}\right)^{\frac{1}{2}} \left(1 + \frac{\bar{x}^2}{\sigma^2}\right)^{-\frac{1}{2}(n-3)} \\ &= \left(\frac{2n}{\pi}\right)^{\frac{1}{2}} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{1}{2}(n-3)}, \end{aligned}$$

as $t^2 = (n-1)\bar{x}^2/\sigma^2$. This is followed by a table that shows the values of K associated with Fisher's 5% values of t for various sample sizes n, reproduced here as Table 3.

Jeffreys then mentions the paradox:

> "For the first few entries my formula may be appreciably inaccurate, but for $n = 8$ and more it should be fairly good. It appears therefore that the 5% point of the t distribution never corresponds to a value of K less than about 0.5, or to 2 to 1 odds on the need for the new parameter. If we are entitled to interpret this as indicating at what value of K we may consider a new parameter as worth introducing, the value should be about 0.5; but there will then be just about as much confidence in the need for it as in a statement that an estimate of a parameter, whose relevance is not in doubt, is right within its standard error.
> The inequality is reversed at large numbers of observations; thus for $K = 1$ and large n we have the approximation
>
> $$t^2 = \log_e 2n/\pi,$$
>
> whereas the 5% point of the t distribution tends to $t = 1.96$. The properties of the logarithm make the rise very slow; when $n = 100{,}000$, t is still only 3.32. *But if the 5% rule was used habitually there would be cases, with large numbers of observations, when a new parameter is asserted on evidence that is actually against it.* Users of the rule usually advocate it with considerable caution, which would agree with the indications of the present theory up to about 30 observations, but at large numbers it is definitely too lax." (Jeffreys 1938a, pp. 379; italics added for emphasis)

Jeffreys then explains

> “It may be worth while to call attention again to the reason for the increase of t and its analogues in other tests when the number of observations is very large. If we start with the minimum of information about the new parameter, which is quite likely to be zero but might account for most of the outstanding variation until we have actually analysed the data, then as we increase the number of observations the standard error of the estimate steadily falls. If the parameter is not zero, however, it is independent of the number of observations, and will ultimately become several times the standard error of its estimate and asserted to be genuine. If the estimate persists within the order of magnitude of its standard error, our confidence that this is because the parameter is really zero will naturally increase, on the ground that with a large number of observations it is increasingly unlikely that we should have failed to find it if it was there. This of course is a well-known phenomenon in physics, where an estimated difference, always in doubt, strengthens that doubt by diminishing every time the number of observations is increased or the experimental technique improved; and it is represented in the present theory by the increase of the outside factor in K. When the number of observations is small, this factor is not much more than 1, and it is impossible to obtain strong support for q however well the observations may agree with it; and in sampling problems in similar conditions it is also impossible to obtain strong support for $\sim q$. It may be recalled that in the problem of sampling to test an even chance it took an 80:80 sample to give 10 to 1 support for q and a 7:0 one to give 10 to 1 support for $\sim q$. It is in such cases that we say that there is not enough evidence to make a decision, and any definite rule will make a considerable number of mistakes of one kind or the other. Mathematically, the ratio of the estimate to its standard error must increase with the number of observations because it has to counteract this factor to reduce K to any fixed value. In general terms, it must increase because the number of cases where q is still acceptable remains the same, but those where it is untrue and its falsehood still undetected become fewer. (I am not here considering cases where selection of an extreme value, or previous knowledge indicating a restriction on the possible values of a new parameter, needs to be taken into account; they only complicate the matter without altering the general principle.)” (Jeffreys 1938a, pp. 379–380)

10. The 1938 article *Significance tests when several degrees of freedom arise simultaneously*

Here, Jeffreys first describes the Bayes factor and immediately points out its dependence on sample size:

> “If a set of observations are analysed for a new parameter a, which is initially as likely as not to be zero, and the possible range of whose values is s if it is not zero, we can denote the proposition that it is 0 by q, and the proposition that it is not 0 by $\sim q$. [EWAL: Here Jeffreys inserts the following footnote: “My q is always what Fisher (1935) calls a “null hypothesis”.”] Then the prior probabilities of q and $\sim q$ are given by

$$P(q \mid h) = P(\sim q \mid h) = \tfrac{1}{2}, \tag{1}$$

> and the posterior probabilities on data θ are shown, by an approximate argument (Jeffreys 1937b, p. 250 [EWAL: This refers to the fragment from *Scientific Inference* provided earlier]), to be given by
>
> $$K = \frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} \Big/ \frac{P(q \mid h)}{P(\sim q \mid h)} = \frac{s}{\surd(2\pi)\sigma_\alpha} \exp\left(-\frac{\alpha^2}{2\sigma_\alpha^2}\right), \tag{2}$$
>
> where α is the maximum likelihood solution for a and σ_α its standard error. Since s is initially fixed and σ_α decreases like $n^{-\frac{1}{2}}$ when n, the number of observations, increases, the outside factor is proportional to $\surd n$. If K is less than 1, the observations support the introduction of the new parameter; if K is more than 1 they do not. In the cases so far examined the critical value of α/σ_α ranges from about 1.8 to 3 as the number of observations rises from 5 to 5000." (Jeffreys 1938b, p. 161)

Later in the same article, the dependence of the Bayes factor on sample size (as $\surd n$) plays a crucial role. For instance, on p. 164 Jeffreys remarks that "the outside factor in the support for q is of order $n^{\frac{1}{2}}$; this factor would be the support provided if the estimates happened to agree exactly with the predictions made by q." (see also p. 172). However, in this article, Jeffreys does not engage in an explicit comparison between Bayes factors and p values.

11. The 1938 article *Maximum likelihood, inverse probability and the method of moments*

In this article, Jeffreys hints at the paradox but underplays its practical importance:

> "(...) a moderate fraction of the prior probability of a [i.e., a parameter] is concentrated in a particular value a_0. This is the case where a possible value of a is already assigned and the observations are to be used to test whether this value is correct. (...) The result, which I had hardly expected to find, was that if $\alpha - a_0$ is less than a certain multiple of σ_a (varying somewhat with n and the type of problem), the observations increase the probability that a is equal to a_0. This connects up significance tests with the principle of inverse probability, but the results do not differ greatly from those that statisticians have found to work well in practice. The relation to the method of maximum likelihood is that the apparently arbitrary rejection of small differences found by that method is now explained in terms of the general theory." (Jeffreys 1938c, p. 148)

12. The 1938 article *Significance tests for continuous departures from suggested distributions of chance*

This article features a more explicit comparison to p values. Here Jeffreys sets out to test the null hypothesis that a set of frequencies are uniformly distributed. He arrives at the familiar $\sqrt{n}$ form of his test and then engages explicitly with the paradox:

> "Hence (...)
>
> $$K = \frac{P(q \mid \theta h)}{P(\sim q \mid \theta h)} = \sqrt{\left(\frac{n}{2\pi}\right)} c \exp(-\tfrac{1}{2} n a_0^2). \tag{15}$$
>
> The term in $f(t)$ will therefore be supported if a_0 [the MLE—EWAL] is such as to make this less than 1. The standard error of a_0, in this notation, is $n^{-\frac{1}{2}}$, so that the exponential factor has the usual form $\exp(-\frac{1}{2}\chi^2)$.
> The following table, for various values of n, gives K for $a_0 = 0$ for the two values of c, and the values of χ^2 and $a_0 n^{\frac{1}{2}}$ that make $K = 1$. For comparison we may notice that Fisher's (1936, Table III) 5% and 1% limits, for one degree of freedom, are at $\chi^2 = 3.84$ and 6.64; the former would agree in the first case at about 200 observations, the latter at about 4000. In the second case the agreements would come at about 100 and 1700 observations. His test, of course, does not mean quite the same thing; it says when an observed result would be surprising on hypothesis q, whereas mine, for the larger numbers of observations, may admit this and yet say that it would be still more surprising on $\sim q$. In any event cases where the observed a_0 would come in the disputable region would be expected to be rare if *either* of the hypotheses compared was correct, and some third alternative may suggest itself." (Jeffreys 1938d, p. 310; italics in original; table reproduced as Table 4)

13. The 1938 article *Aftershocks and periodicity in earthquakes*

In this article, Jeffreys studies the hypothesis that earthquakes are independent events that do not excite one another. Jeffreys first elaborates on standard practice:

> "The usual statistical procedure, recommended in particular by R. A. Fisher, is to reject the trial hypothesis if the contribution to χ^2 examined is such that the probability of a larger χ^2, if the hypothesis was correct, is less than 0.05; if high confidence is required the trial hypothesis will be rejected (and correspondingly the modified one accepted) if this probability is less than 0.01. The former criterion is somewhat mild, since it would imply the acceptance as genuine of all discrepancies more than twice their standard errors." (Jeffreys 1938e, p. 114)

Jeffreys then describes his own significance test and mentions the paradox:

> "The type of significance test that I have introduced depends on the general theory of probability; the observed values appear in the results only through the

Table 4 Table reproduced from Jeffreys (1938d, p. 310)

n	K		$a_0 n^{\frac{1}{2}}$		χ^2	
5	1.03	1.55	0.25	0.94	0.06	0.88
10	1.46	2.19	0.87	1.25	0.76	1.57
20	2.06	3.09	1.20	1.50	1.45	2.26
50	3.26	4.89	1.54	1.78	2.36	3.17
100	4.61	6.92	1.75	1.97	3.06	3.87
200	6.51	9.76	1.94	2.13	3.75	4.56
500	10.31	15.46	2.16	2.34	4.67	5.48
1000	14.6	21.9	2.32	2.48	5.36	6.17
2000	20.6	30.9	2.46	2.62	6.05	6.86
5000	32.6	48.9	2.46	2.79	6.97	7.78
10,000	46.1	69.2	2.77	2.86	7.66	8.19

Increases in sample size n need to be accompanied by increases in the χ^2 value so that the Bayes factor K remains constant at $K = 1$

> contributions to χ^2 from the degrees of freedom actually considered, so that the tests provide an explanation of the importance of χ^2, which was introduced somewhat arbitrarily by Pearson, though it has properties of symmetry under transformation that would make it commendable by themselves. For a given number of degrees of freedom, the value of χ^2 that makes it more probable than not, on the data, that a new parameter or set of parameters is required, is found to vary somewhat with the number of observations[2]). In the case of a periodicity inferred from observed frequencies, I find that for 200 observations the periodicity is just supported if $\chi^2 = 7.1$; for 500, 8.3; for 1000, 8.9. To establish a 10 to 1 probability that the periodicity is genuine these values must be increased by 5.1. Fisher's 5 per cent. limit for two degrees of freedom is at $\chi^2 = 5.99$, his 1 per cent. one at 9.21. With these numbers of observations his 5 per cent. criterion would therefore sometimes accept a periodicity that mine would reject, though the agreement is good at somewhat smaller values." (Jeffreys 1938e, p. 114)

Jeffreys's footnote 2 lists several of his earlier works in which the paradox is evident:

> "[2]) Scientific Inference, **1937**, 249–252 and 266–9 (General Discussion and Summary of Results). – Proc. Camb. Philos. Soc. **31** (1935) 213–217 (Correlation). – Proc. Camb. Philos. Soc. **32** (1936) 432–445 (Representation of a Series of Measures by Assigned Functions and Tests of Randomness). – Proc. Camb. Philos. Soc. **33** (1937) 35–40 (Comparison of Series of Measures). – Proc. Roy. Soc. London (A) **162** (1937) 479–495 (Contingency and tests for agreement of Sampling Ratios. Improved discussions of some problems treated in earlier papers are given.)" (Jeffreys 1938e, p. 114)

14. The 1939 first edition of *Theory of Probability*

The first edition of Jeffreys's magnum opus *Theory of Probability* describes a scenario similar to that covered in the addenda of the 1937 reissue of *Scientific Inference*. Specifically, Jeffreys introduces the Bayesian hypothesis test by defining the null hypothesis q and the alternative hypothesis $\sim q$. Under $\sim q$, there is a new parameter α. Let m denote the possible range of values for α about 0 within which the prior probability may be taken as uniformly distributed, and let a denote the maximum likelihood estimate and s its standard error. Then, if s is much smaller than m, Jeffreys approximates the Bayes factor K (i.e., BF_{01}) as

$$K = \frac{P(q \mid aH)}{P(\sim q \mid aH)} = \frac{m}{\sqrt{(2\pi)}s} \exp\left(-\frac{a^2}{2s^2}\right),$$

where H indicates background knowledge. Jeffreys then continues:

> "If a is s or less, and s is much less than m, K will be large and the observations support q, that is, they say that the parameter α is probably not needed. But if a is much larger than s, the exponential will be very small and the observations will support the need for the new parameter. There will be a critical value of a/s such that $K = 1$ and no decision is reached.
> In most cases s, being the standard error of a, diminishes with increasing n like $n^{-1/2}$; hence the first factor in K increases like $n^{1/2}$. Thus the larger the number of observations the stronger the support for q will be if $a < s$. This is a satisfactory feature; the more thorough the investigation has been, the more ready we shall be to suppose that if we have failed to find evidence for α it is because α is really 0. But it carries with it the consequence that the critical value of a/s increases with n (though that of a of course diminishes); the increase is very slow, since it depends on $\sqrt{(\log n)}$, but it is appreciable. The test does not draw the line at a fixed value of a/s." (Jeffreys 1939, p. 194; echoed in Jeffreys 1948, pp. 221–222 and Jeffreys 1961, p. 248)

In Appendix I, Jeffreys again explicitly compares the Bayes factor against the p value. Jeffreys concludes:

> "In spite of the difference between the nature of my tests and those based on the P integrals, and the omission of the latter to give the increases of the critical values for large n (dictated essentially by the fact that in testing a small departure found from a large number of observations we are selecting a value out of a long range and should allow for selection), it appears that there is not much difference in the practical recommendations. Users of these tests speak of the 5 per cent. point in much the same way as I should speak of the $K = 10^{-1/2}$ point, and of the 1 per cent. point as I should speak of the $K = 10^{-1}$ point; and for moderate numbers of observations the points are not very different. At large numbers of observations there is a difference, since the tests based on the integral would sometimes assert significance at departures that would actually give $K > 1$. Thus there may be opposite decisions in such cases. But they will be very rare."

> (Jeffreys 1939, pp. 359-360; echoed in Jeffreys 1948, p. 399 and Jeffreys 1961, p. 435)

Appendix I then concludes with four tables associated with different statistical scenarios. Each table shows that a constant level of Bayes factor support requires that larger sample sizes yield a higher multiple of the standard error.

15. The 1940 article *Note on the Behrens–Fisher formula*

In this article, Jeffreys briefly outlines his hypothesis test and adds that the threshold for accepting the alternative hypothesis is not a constant 'as usually defined':

> "A definite limit is then found for z, such that larger values support the need for the new parameter while smaller ones support the null hypothesis, but this limit is not given by any single value of $P(t)$ as usually defined." (Jeffreys 1940, p. 49)

16. The 1942 article *On the significance tests for the introduction of new functions to represent measures*

In this article, Jeffreys once more emphasizes the dependence of the Bayes factor K on sample size. After providing the equation for K in its familiar form, Jeffreys provides a table that shows how K increases with n when t is fixed at 0, and how t^2 increases with n when K is fixed at 1. Jeffreys remarks

> "It is interesting that the values of t^2 for $K = 1$ increase steadily with n, just as the corresponding values of χ^2 do. This of course is the level where the test is quite indecisive." (Jeffreys 1942, p. 260)

17. The 1948 second edition of *Theory of Probability*

Although this second edition is 31 pages longer than the 380-page first edition, the paradox-related content (i.e., pp. 221–222, p. 399) has remained mostly unchanged, except for a small change in notation and for a partly adjusted and expanded set of tables in the appendix.

18. The 1950 article *Bertrand Russell on Probability*

In this article, Jeffreys describes his generic Bayes factor, including the $\sqrt{n}$ term that exposes the paradox:

> "But if we are at liberty to modify a law arbitrarily to any extent we can fit any set of observations exactly, and some of these possibilities would fit any further observation whatever; consequently if there is no limitation on the choice of laws no prediction from observations is possible. (...) [a solution] is given in my *Theory of Probability*, Chapters 5 and 6. This is that where a suggested modification of

> a law involves an increase in the number of adjustable parameters, half the prior probability is concentrated in the old law; in other words, when a modification is suggested it is as likely to be needed as not. This has been shown to lead to satisfactory significance tests in the standard problems of statistics, though there is much more to be done. The results are of the approximate form
>
> $$\frac{\mathrm{P}(q/\theta H)}{\mathrm{P}(q'/\theta H)} = \sqrt{(An)}\mathrm{e}^{-\frac{1}{2}a^2/s_a^2}$$
>
> Here if the new parameter considered is α, it is defined so as to be zero on the old law q, but on the modified law q' it has to be estimated from the observations; H is the previous information and θ the observational evidence. A is a constant usually of order 1, n the number of observations, a the estimate of α by the usual statistical methods, and s_a its standard error. The expression is of order $\sqrt{n}$ if a/s_a is less than 1, but very small if a/s_a is large. Consequently observations support the old law for $a/s_a < 1$ and the new one if it is large. This choice of the prior probability is what I call the simplicity postulate." (Jeffreys 1950, p. 316; italics in original)

19. The 1953 comment on Lindley's article *Statistical inference*

Historically, the 1953 Lindley article *Statistical inference* is particularly relevant, as it can be considered the conceptual forerunner to the 1957 paradox article. Inspired by the work of Abraham Wald, Lindley studied statistical procedures that minimize a weighted sum of Type I and Type II errors.[8] Lindley showed that for consistency to hold regardless of the weight assigned to the errors, the critical value has to increase with sample size: "...the critical value (...) increases with n, although very slowly. In this it agrees with the test proposed by Jeffreys (1948)." (Lindley 1953, p. 60).

In a comment published alongside Lindley's original article, Jeffreys elaborates on the agreement:

> "The appearance of log n [in Lindley's tests – EWAL] is interesting in relation to my significance tests. At first sight the origins of this term look quite different, since in mine *it expresses an allowance for selection; we reasonably discount an exceptional result if we have looked specially hard for one.* In Mr. Lindley's it is an allowance for the cost of installing a new plant when the benefit would be small.
>
> It is easy to see, however, that a similarity might have been expected. If the prior probability distribution for a parameter μ is $P(d\mu \mid H) = f(\mu)\,d\mu$, the likelihood of a set of data θ is $L(\mu, \theta)$, and the benefits of two courses of action, depending on μ, are $K_1(\mu)$, $K_2(\mu)$, the posterior probability distribution of μ is $P(d\mu \mid \theta H) \propto f(\mu)L(\mu, \theta)\,d\mu$, and the expectations of benefit are

[8] At this time, Lindley was still a frequentist, as witness statements such as "...the use of inverse probability solutions as a *general* rule can hardly be considered satisfactory, though in special circumstances they may be adequate." (Lindley 1953, p. 45; see also Fienberg 2003).

> $\int K_1(\mu) f(\mu) L(\mu, \theta)\, d\mu$, $\int K_2(\mu) f(\mu) L(\mu, \theta)\, d\mu$. Thus K enters in combination with f, as Mr. Lindley finds. This might have been expected, since Bayes defined probabilities in terms of ratios of expectations of benefits, and in an economic application K and f will always be combined." (Jeffreys 1953, p. 72; italics added for emphasis)

Lindley then replied to Jeffreys as follows:

> "His connection between the log n term in our two derivations is most interesting, and in conjunction with his statement that, in some circumstances, one should maximize the expected benefit, it makes me realize that my ideas on inference are much closer to Professor Jeffreys' than I had thought." (Lindley 1953, p. 76)

It should not go unmentioned that, in a different comment, Lindley's contribution was evaluated positively by Egon Pearson himself:

> "We see at once the practical "hunch" to which Lindley's approach is here trying to give expression. If we keep α fixed as n increases from 20 to 100 we have a rapidly increasing chance of establishing that a difference is significant when, say, $\mu - \mu_0 = 0.4$. Could we not well afford to sacrifice some of this additional power in order to reduce the risk of rejecting the null hypothesis when it is true, i.e., of making the decision d_1 wrongly? (...)
> Lindley points out that the test proposed by Jeffreys has similar properties to his tests (...). The same practical objective may be attained if desired by the *quite legitimate device of reducing α as n increases*. If the exponents of usually accepted test theory had not thought of this possibility before, it only serves to illustrate the value of looking at a problem of statistical inference from several points of view and making numerical comparisons." (Pearson 1953, p. 69; italics added for emphasis)

In a later section, we will elaborate on the idea that the paradox undercuts only the Fisherian interpretation of a p value as 'evidence against the null hypothesis'; in the Neyman–Pearson paradigm, however, the brunt of the paradox can be avoided by adopting a lower value of α when power is known to be high.

20. The 1955 article *The present position in probability theory*

Here, Jeffreys again presents his generic Bayes factor equation including the $\sqrt{n}$ term:

> "In most cases the results are of very similar form when the number of observations, n, is large. If the straightforward estimate of α_m, apart from the significance question, would be $a_m \pm s_m$, we usually get (θ standing for the data collectively)
>
> $$K = \frac{P(q|\theta H)}{P(q'|\theta H)} \doteqdot \frac{A n^{\frac{1}{2}}}{f(0)} \exp\left(-\frac{a_m^2}{2 s_m^2}\right).$$
>
> A is a constant of order 1. We must have $f(0) > 0$, otherwise the null hypothesis would always be asserted [see also Jeffreys (1961, p. 251—EWAL]. If $f(0) > 0$

> and $|a_m| < s_m$, K is large and q has a high probability. If $|a_m|$ greatly exceeds s_m, K is small and $q^{'}$ has a high probability in comparison with q. In practice s_m^2 usually decreases with n like $1/n$, and $K = 1$ for a moderate value of $|a_m|/s_m$, usually 2 to 4." (Jeffreys 1955, p. 282)

21. The 1957 second edition of *Scientific Inference*

In the second edition of *Scientific Inference*, Jeffreys now presents the generic approximate Bayes factor in the main text (p. 72; as he did in the first and second editions of *Theory of Probability*), where it was previously presented in the addenda of the 1937 reissued first edition. In contrast to that first edition, Jeffreys no longer engages in an explicit comparison between Bayes factors and p values, and only hints a the paradox when he writes:

> "The main point is that the null hypothesis is in general strongly supported if the maximum likelihood estimate of the new parameter is less than its standard error; but the introduction of the new parameter is strongly supported if the estimate is much more than the standard error. *With ordinary numbers of observations* (from 20 to 1000) the transition comes at about 3 times the standard error in most problems." (Jeffreys 1957a, p. 72; italics added for emphasis)

22. The 1957 article *Probability theory in astronomy*

Jeffreys again presents his approximate form:

> "The theory leads to rules of significance for the introduction of new parameters in laws. They are usually approximately of the form
>
> $$K = \frac{P(q|\theta p)}{P(q'|\theta p)} \doteqdot (An)^{1/2} \exp\left(-\frac{a^2}{2s_a^2}\right).$$
>
> Here q is the hypothesis that the new parameter α is zero, that is, that the previous law needs no alteration; q' the hypothesis that α is needed, having a value to be estimated from the observations; a and s_a are the estimate of α and its standard error as given by the method of least squares; n is the number of observations; and A is a constant, usually not far from 1. If $|a| < s_a$, the factor $n^{1/2}$ makes $K > 1$ and the old law is supported; but with ordinary numbers of observations, if $|a| > 2s_a$ or $3s_a$, $K < 1$ and the new law is supported. To apply a test of this sort it is of course of the first importance that the number of observations shall be stated. This is in fact not often done by physicists, but thanks mainly to the work of Fisher (with whom I do not always agree) biologists usually do it, but with different rules." (Jeffreys 1957b, p. 349)

In sum, it appears that at the time of writing, Lindley was unaware of the extent to which Jeffreys had already identified, explained, and explored the paradox. The single reference to the appendix from the 1948 edition of *Theory of Probability* certainly does not do justice to the central position that the paradox occupied in Jeffreys's philosophy; nor is the reference to the 1948 edition historically accurate, as Jeffreys had completed his work related to the paradox already in the second half of the 1930s. The idea that Lindley may not have been fully aware of Jeffreys's prior work on the paradox receives support from the following fragment of Lindley's obituary of Jeffreys:

> "He was one of the finest writers of scientific English, with an accurate, yet almost melodious, style. Like Joyce, he used the language sparingly, condensing many ideas into few words. A paradox that has been much discussed, and erroneously associated with my name, occupies two sentences in the Theory (p. 248)." (Lindley 1989, p. 417)

As outlined above, Jeffreys devoted many more than two sentences to the paradox. The fact that Lindley was only somewhat aware of the extent of Jeffreys's contributions is also consistent with the following remark:

> "Having produced MEU [maximization of expected utility – EWAL] as the constructive device for producing statistical methods, we tried to apply it to standard problems, finding sometimes that it agreed, as in the use of sufficient statistics, but more often finding that it did not, for example in the use of the tail area in a significance test. (Interestingly Jeffreys had pointed this out in 1939 but none of us had fully appreciated what he was saying. This is especially ridiculous in my case since I had attended Jeffreys's lectures in Cambridge in 1947; the only excuse I can offer, apart from my own stupidity, is that he was a bad lecturer. But that is not valid since his book is, at least seen through today's eyes, lucid and still worth reading.)" (Lindley 2000, p. 8)

As outlined above, Jeffreys's pointed out the paradox as early as 1935, returning to the same theme many times prior to the first edition of *Theory of Probability*.

The 1957 contribution from Bartlett

For over 2 decades, Jeffreys had repeatedly pointed out the potential conflict between p values and Bayes factors. However, Jeffreys's work on Bayes factors had been largely ignored. Instead, it was the 1957 article by Lindley that brought the paradox into the limelight. Although Lindley's conclusions were qualitatively correct, he did omit an important term from his equations, an oversight that was quickly corrected by Bartlett (1957):

> "I would agree that he [Lindley – EWAL] establishes the point that one must be cautious when using a fixed significance level for testing a null hypothesis irrespective of the size of sample one is taking. However, there is a slip, in his expression for K under his equation (1), that appears to me, unless corrected, to lead to an overstatement of his point. The prior distribution for θ, given that $\theta \neq \theta_0$, was assumed to be uniform over an interval I, and hence its density function should be $1/I$ in this interval. This leads to the extra factor $1/I$ in the second term in the expression for K.[Here Bartlett adds a footnote: "There is

> also a further dropping of a factor $1/\sigma$ in the last formula on p. 191, but this is a more trivial slip." – EWAL] This expression then becomes consistent with Jeffreys's equation (10), §5.0 in his book (second edition, 1948) [This is the equation for K given above in the section on the 1939 first edition of *Theory of Probability* – EWAL]." (Bartlett 1957, p. 533)

In an editorial note following Bartlett's paper, Sir Maurice Kendall stated that "Mr Lindley agrees and apologizes" for omitting the $1/I$ term from his first equation. However, Kendall points out that this oversight affects neither Lindley's general argument nor his concrete examples.

After including the $1/I$ term omitted by Lindley, Bartlett notes that a uniform prior on the entire real line ("the most natural prior", p. 533) will yield infinite support in favor of the null hypothesis, a "silly answer" (p. 533). Moreover, in order to escape from the paradox, Bartlett argues that in the planning stage of an experiment, sample size may be chosen such that $\sqrt{n}$ is proportional to $1/I$ (i.e., researchers who expect small effects will collect many observations).

Based on our reading, we conclude that both Lindley and Bartlett unwittingly presented a slightly confused version of Jeffreys's earlier work. As far as Lindley is concerned, he indeed omitted the $1/I$ term that is correctly included in Jeffreys's equations (e.g., see above: Jeffreys 1936b, p. 417; 1937b, pp. 250–251; 1938b, p. 161; 1939, p. 194; 1948, pp. 221–222). In addition, Lindley appears to have been unaware of Jeffreys's general approximate $\sqrt{n}$ form of the Bayes factor. Lindley does present this form at a later stage of his paper, but without the $1/I$ term, and preceding it with an attribution to Barnard: "An alternative interpretation of the paradox was suggested to me by Prof. Barnard." (p. 189). Lindley then notes that this $\sqrt{n}$ form shows that "Clearly (...) for fixed significance level the likelihood of the null hypothesis increases indefinitely with the sample size." (p. 189). As mentioned above, the form of this equation and its conclusion were already presented two decades earlier by Jeffreys (1936b, p. 417).

As far as Bartlett is concerned, his conclusion that a uniform (improper) prior leads to the "silly answer" of infinite support for the null hypothesis was anticipated by Jeffreys in 1935:

> "To apply this theory it is therefore necessary that we should have previous knowledge of the range of possible values of y. (...) Since m enters only through its logarithm its effect is in any case not great in practical cases, and it does not need to be determined very accurately (...)
> It may happen, however, that we have no previous information about the range of admissible values of y; then m is effectively infinite, and it appears that no matter how many observations we have we shall never be able to infer a systematic difference." (Jeffreys 1935, p. 207)

Jeffreys also discussed the problem of improper priors for testing in the 1948 second edition of *Theory of Probability*, in the section *Required properties of* $f(\alpha)$:

> "It might appear that on q' the new parameter is regarded as unknown and therefore that we should use the estimation prior probability for it. But this leads

> to an immediate difficulty. Suppose that we are considering whether a location parameter α is 0. The estimation prior probability for it is uniform, and (...) we should have to take $f(\alpha) = 0$, and K would always be infinite. We must instead say that the mere fact that it has been suggested that α is zero corresponds to some presumption that it is fairly small." (Jeffreys 1948, p. 225; Jeffreys 1961, p. 251)

Thus, the popular belief that Bartlett was the first to point out the problem with improper priors for Bayes factor testing (e.g., O'Hagan and Forster 2004, p. 78) is incorrect.

Bartlett also commented on a "more trivial slip" in Lindley's paper, that is, "a further dropping of a factor $1/\sigma$ in the last formula on p. 191". This is the offending equation:

$$\sqrt{\left(\frac{n}{2\pi}\right)} \exp\left\{-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right\}.$$

However, this equation is in fact similar to those presented by Jeffreys. As noted in Cousins (2017), the unit-information prior (e.g., Kass and Wasserman 1995) sets the range m equal to the uncertainty associated with a single observation, meaning that after dividing the m and the $1/\sigma$ terms, only the $\sqrt{n}$ term remains.

Finally, Bartlett suggests to reduce the spread of the prior as $\sqrt{n}$ (see also Andrews 1994; Cox 2006, pp. 106–107, as noted by Cousins 2017). In other words, he assumes that researchers who collect a large sample do so because they expect the effect to be relatively small—the sample size, therefore, provides a clue about the spread of the prior distribution for the test-relevant parameter under $\mathcal{H}_1$. There are several problems with this suggestion. First and foremost, Bartlett's scaling solution makes it impossible for the Bayes factor to produce convincing evidence in favor of the null hypothesis; as n increases, the alternative hypothesis will increasingly resemble the null hypothesis, and consequently the null hypothesis can never reach a compelling level of support. This is a key objection, as a cornerstone of Jeffreys's philosophy of testing is that "An adequate theory of scientific investigation must leave it open for any hypothesis whatever that can be clearly stated to be accepted on a moderate amount of evidence." (Jeffreys 1961, p. 129). This notion harks back to Jeffreys's early work with Dorothy Wrinch, in which they argued that in order for a universal generalization (e.g., propositions such as "all ravens are black") to attain a compelling degree of plausibility it is necessary to adjust Laplace's idea of uniform prior distributions and assign point mass to the general law (i.e., Wrinch and Jeffreys 1921; Ly et al. 2020). Second, Bartlett's solution does not apply to observational studies, where the issue of sample size planning is irrelevant. Third, researchers may collect larger samples for a variety of other reasons including feasibility (e.g., the presence of sufficient funding), ease of data collection (e.g., via online surveys), scientific or societal importance of the topic under study, personality characteristics of the researcher, and so on. Finally, as indicated above, in 1937, Jeffreys already mentioned and rejected Bartlett's 1957 proposal (Jeffreys 1937b, p. 259).

In sum, the arguments presented in Lindley (1957) and Bartlett (1957) were already discussed 2 decades earlier by Jeffreys, in more detail and without errors. The main difference is in the evaluation of the practical ramifications of the paradox; whereas Jeffreys downplays the discrepancy between Bayes factors and p values for practical data analysis ("curiously", according to Cousins 2017, p. 400), Lindley stresses it. In a later article, Lindley doubles down: "There is therefore a serious and systematic difference between the Bayesian and Fisherian calculations, in the sense that a Fisherian approach much more easily casts doubt on the null value than does Bayes. *Perhaps this is why significance tests are so popular with scientists: they make effects appear so easily*. Notice that this result depends on a 'sharp' prior being used, with $p(\theta = 0) > 0$." (Lindley 1986, p. 502, italics added for emphasis). The reason for this difference in perspective is arguably due to the fact that Jeffreys calibrated a $p = 0.05$ result to a Bayes factor of 1 (reasoning that these were the watershed values in the two statistical paradigms), whereas Lindley sought to compare the p value and the posterior probability for the null hypothesis directly.

The root of the paradox: a summary of Jeffreys's argument

Jeffreys generally explained the paradox in two ways. The first way is to note that the p value focuses on the predictions from $\mathcal{H}_0$, whereas the Bayes factor compares the predictions from $\mathcal{H}_0$ against those from a composite $\mathcal{H}_1$. At hand is the scenario where sample size n increases but the multiple of the standard error is constant, such that $\hat{\theta}/\text{se}(\hat{\theta}) = c, \forall n \to \infty$. In this case, the predictive adequacy of $\mathcal{H}_0$ is unaffected—and consequently the p value remains constant also, but the predictive adequacy of $\mathcal{H}_1$ gradually deteriorates. The reason for this deterioration is that, as n increases, an increasingly smaller set of parameter values provides acceptable predictions. An ever increasing part of $\mathcal{H}_1$ is found wanting, and this decreases the average predictive performance across all parameter values under $\mathcal{H}_1$. This phenomenon does not occur if the predictive adequacy of $\mathcal{H}_1$ is based only on the maximum likelihood estimate $\hat{\theta}$; however, this is a cherry-picked value that is in need of a multiplicity correction, for else the null hypothesis could never be supported by any data. The correction for cherry-picking (or *selection*, as Jeffreys called it) is achieved automatically through the prior distribution (see also Cousins 2017, pp. 401–402 and Jaynes 2003, Chapter 20). The "correction for selection" explanation for the deteriorating predictive performance of $\mathcal{H}_1$ was prominently presented in the *Theory of Probability*, for instance in the fragments cited above (i.e., Jeffreys 1938a, pp. 379–380; Jeffreys 1939, pp. 359–360, Jeffreys 1948, pp. 399–400, and Jeffreys 1961, pp. 435–436; see also Jeffreys 1953, p. 72) and also in the following:

> "The possibility of getting actual support for the null hypothesis from the observations really comes from the fact that the value of α indicated by it is unique. q' indicates only a range of possible values, and *if we select the one that happens to fit the observations best we must allow for the fact that it is a selected value.* If $|a|$ is less than s, this is what we should expect on the hypothesis that α is 0, but if α was equally likely to be anywhere in a range of length m it requires that an event with a probability $2s/m$ shall have come off. If $|a|$ is much larger than s, however, a would be a very unlikely value to occur if α was 0, but no more unlikely than any other if α was not 0. In each case we adopt the less remark-

> able coincidence.” (Jeffreys 1961, p. 248, italics added for emphasis; echoed in Jeffreys 1939, pp. 194-195 and Jeffreys 1948, p. 222)

Jeffreys’s second, related explanation for the paradox refers to the need for consistency under $\mathcal{H}_0$. As mentioned in the above fragment, Jeffreys argues that when the estimate is of the order of the standard error, this constitutes increasingly strong evidence in favor of $\mathcal{H}_0$ as sample size grows. The idea is intuitive: for instance, 5 heads out of 10 tosses yields less evidence in favor of the fair coin hypothesis $\theta_0 = 1/2$ than would 500 heads out of 1000 tosses (cf. Berkson 1942, p. 332). This implies, however, that the Bayes factor break-even point $\text{BF}_{01} = 1$ has to be at a multiple of the standard error that increases with n. This effectively creates the paradox (e.g., Wagenmakers, Gronau, Dablander, & Etz, in press).

Two examples by Jack Good

Across several articles, Jack Good attempted to explain why it is problematic to use a significance threshold that is a constant multiple of the standard error. A first example was presented in Good (1980b):

> “Dr. Deborah Mayo raised the following question. How could one convince a very naive student, Simplissimus, that a given tail-area probability (P-value), say 1/100, is weaker evidence against the null hypothesis when the sample is larger? Although this fact is familiar in Bayesian statistics the question is how to argue it without (explicit) reference to Bayesian methods.
> One can achieve this aim, without even referring to power functions, in the following manner.
> Take a very concrete example, say the tossing of a coin, and count the number r of heads (“successes”) in N trials. Ask Simplissimus to specify any simple non-null hypothesis for the probability p of a head. Suppose he gives you a value $p = .5 + \epsilon$. First compute a value of N so that a ϵ value of r approximately equal to N$(.5 + \frac{\epsilon}{7})$ would imply a tail-area probability close to 1/100. Then point out that the fraction $.5 + \frac{\epsilon}{7}$ of successes is much closer to .5 than it is to $.5 + \epsilon$ and therefore must *support* the null hypothesis as against the specific rival hypothesis proposed by Simplissimus. Thus, for any specified simple non-null hypothesis, N can always be made so large that a specified tail-area probability supports the null hypothesis more than the rival one. This should convince Simplissimus, if he had been listening, that the larger is N the smaller the set S of simple non-null hypotheses that can receive support (as compared with p = 1/2) in virtue of a specified P-value. If the tail-area probability, for example 1/100, is held constant, the set S converges upon the point p = 1/2 when N is made larger and larger.” (Good 1980a, pp. 307–308; italics in original)

As elaborated in Ly and Wagenmakers (in press-b):

> “For instance, assume Simplissimus specifies their simple non-null hypothesis as $\theta = 0.57$ with $\epsilon = 0.07$. Then our target value for the number of successes

> s equals $n(0.5 + 0.07/7) = n \times 0.51$. So for a sample proportion of 0.51 we now seek n such that the two-sided tail area probability equals .01. We find that $n = 16700$ –consisting of 8517 heads, for a sample proportion of $s = 8517/16700 = 0.51$, as stipulated– yields a tail-area just below .01. But the sample proportion of 0.51 is much closer to the null hypothesis (i.e., $\theta = 0.50$) than to the non-null hypothesis specified by Simplissimus (i.e., $\theta = 0.57$)."

In a later article, Good present a second example:

> "In the course of discussion of Good (1980b), Dr. Golde Holtzman suggested that instead of considering a binomial model in which all values of the binomial parameter p are considered, we think of a bag known to contain exactly 1000 balls, some white and some black. The null hypothesis, by definition, is that there are 500 of each. The sampling is to be random, with replacement, with N drawings.
> For definiteness suppose that the outcome is $1/2N + \sqrt{N}$ white and therefore $1/2N - \sqrt{N}$ black balls. (We can suppose N is a perfect square.) Then P, taken as a double tail, is about .05; and the fraction of white balls drawn is $1/2 + N^{-1/2}$. If N is large enough, the closest possible rival to $w = 500$ is $w = 501$, where w is equal to the number of white balls in the bag. If therefore $N^{-1/2}$ is much smaller than $1/1000$, that is, if $N/1,000,000$ is large, the probability of the observed outcome will be much larger assuming the null hypothesis than if any other hypothesis is assumed, even $w = 501$. Thus the tail-area probability of .05 will then support the null hypothesis, and the larger N is (above a certain threshold) the more the support will be if the tail-area probability is the same in each case. Moreover, if we were fairly confident of our model in the first place, the tail-area probability of .05 would not be small enough to cause us to suspect the model." (Good 1983, pp. 312–313)

For simplicity, suppose the bag contains just 10 balls. Drawing $120/200$ white balls yields $\hat{\theta} = 0.60$ and gives $p \approx .006$; Drawing $429/780$ white balls yields $\hat{\theta} = 0.55$ and also gives $p \approx 0.006$; and drawing $9690/19{,}000$ white balls yields $\hat{\theta} = 0.51$ and again gives $p \approx 0.006$. To reject $\mathcal{H}_0$ based on a sample proportion of 0.55 (exactly in between the expected proportion for 5 and 6 white balls out of 10) seems premature, and to do so for a sample proportion of 0.51 seems preposterous, as the data are much more likely under $\mathcal{H}_0 : 5/10$ white balls than under even the most likely of the alternative compositions (i.e., $6/10$ white balls; for similar examples, see, e.g., Freeman 1993; Pericchi and Pereira 2016). The problem becomes even more severe when the bag contains only two balls. In this case, any sample of mixed composition, no matter how lopsided (e.g., 1 white ball and 100 black balls) decisively falsifies $\mathcal{H}_1$ and thereby proves $\mathcal{H}_0$.

Note that for this particular example, a frequentist may argue that the details of the problem necessitate the choice of a different test statistic, such as the likelihood ratio between $\mathcal{H}_0 : \theta = 1/2$ and a specific $\mathcal{H}_1$ (e.g., the one closest to $\hat{\theta}$).

Frequentist considerations

Jeffreys demonstrated that the evidence provided by the data for a point hypothesis $\mathcal{H}_0$ vis-a-vis a composite hypothesis $\mathcal{H}_1$ scales with $\sqrt{n}$; consequently, any evidence threshold cannot be a constant multiple of the standard error. This result undercuts the popular interpretation of the classical p value in terms of a fixed, sample-size independent measure of evidence against $\mathcal{H}_0$. This interpretation was promoted by Fisher himself, who argued explicitly that the interpretation of the p value is independent of sample size:

> "It is not true (...) that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability, we thereby make full allowance for the size of the sample, and should be influenced in our judgment only by the value of probability indicated. The great increase of certainty which accrues from increasing data is reflected in the value of P, if accurate methods are used." (Fisher, 1934, p. 182).

Berkson agreed with Fisher's assessment and stated that "small P's are more or less independent, in the weight of the evidence they afford, of the numbers in the sample." (Berkson 1942, p. 333; cf. Royall 1997, p. 70). Jeffreys's work and the associated paradox cast doubt on this evidential interpretation of the p value.

However, in the Neyman-Pearson paradigm the $\sqrt{n}$ scaling of the evidence can be accommodated by reducing α when n is high. This possibility was already suggested by Jeffreys in 1938:

> "It [the 5% rule – EWAL] would mean drawing the line at such a limit as to give a fixed percentage of what Neyman and E. S. Pearson call errors of the first kind, with respect to the number of cases where q is true; but as the limit is at our disposal we are entitled to take it further out and reduce this percentage still further if there is no special reason to expect values of the new parameter in the range affected. To reject the null hypothesis in any cases at all where it is true is not a desirable action for its own sake. It is an evil that becomes necessary if we are to have any criterion for detecting cases where q is untrue, and we are justified in taking such steps as will reduce its importance to a minimum." (Jeffreys 1938a, p. 379)

A similar remark appears in Appendix I of the first edition of *Theory of Probability*:

> "(...) if we assert a genuine departure whenever P is less than 0.01 we shall expect to be wrong in the long run in 1 per cent. of the cases where q is true. According to my theory we should expect to make fewer mistakes by taking the limit further out; when $K = 1$ lies above $P = 0.01$ there will be a smaller risk of rejecting q wrongly, partly counter-balanced by a slight increase in the risk of missing a small genuine departure." (Jeffreys 1939, p. 360, echoed in Jeffreys 1961, p. 435)

In the main text of *Theory of Probability*, Jeffreys also pointed out that—if the prior distribution for the test-relevant parameter under $\mathcal{H}_1$ is well-calibrated—the total number of errors (i.e., $\alpha + \beta$) is minimized using $\mathrm{BF}_{01} = 1$ as the criterion for accept/reject decisions:

> "It may, however, be interesting to see what would happen if the new parameter is needed as often as not, and if the values when it is needed are uniformly distributed over the possible range. Then the frequencies in the world would be proportional to my assessment of the prior probability. Suppose, then, that the problem is, not knowing in any particular case whether the parameter is 0 or not, to identify the cases so as to have a minimum total number of mistakes of both kinds. (...)
> Hence, with world-frequencies in proportion to the prior probability used to express ignorance, the total number of mistakes will be made a minimum if the line is drawn at the critical value that makes $K = 1$.
> Now I do not say that this proportionality holds; all that I should say myself is that at the outset we should expect to make a minimum number of mistakes in this way, but that accumulation of information may lead to a revision of the prior probabilities for further use and the critical value may be correspondingly somewhat altered. But whatever the frequency law may be (...) K would be altered by a factor independent of the number of observations. *We should therefore get the best result, with any distribution (...), by some form that makes the ratio of the critical value to the standard error increase with n. It appears then that whatever the distribution may be, the use of a fixed P limit cannot be the one that will make the smallest number of mistakes.* The absolute best is of course unknown since we do not know the distribution in question except so far as we can infer it from similar cases." (Jeffreys 1939, pp. 326-328, echoed in Jeffreys 1961, pp. 396-397; italics added for emphasis)

Thus, if the prior distribution is calibrated then the Bayes factor provides an optimal frequentist decision criterion. This also holds when the frequentist purpose is to minimize a weighted sum of errors, $\lambda\alpha + \beta$ (Cornfield 1966). Thus, from a Neyman–Pearson perspective, the conflict with a Bayesian assessment of evidence arises specifically in the common scenario where the researcher fixes the probability α of a Type I error (say to 5%) and then tries to minimize the probability β of a Type II error. However, as pointed out above, in high-n situations, the researcher may prefer to sacrifice some power in order to lower the probability of a Type I error. As indicated above Egon Pearson himself judged this strategy "quite legitimate" (Pearson 1953, p. 69). Applying this strategy substantially reduces the discrepancy between the frequentist and the Bayesian results.[9] For related work, see for instance DeGroot and Schervish (2012, Chapter 9), Good 1992, Kim and Choi 2021, Leamer (1978, Chapter 4), Lehmann (1958), Lindley (1953), Maier and Lakens (in press), Mudge, Baker, Edge, and Houlahan (2012), Pérez and Pericchi (2014), Pericchi and Pereira (2016), Savage et al. (1962, pp. 64–67), and Savage (1964, Section 5).

In sum, the Jeffreys–Lindley paradox may be given a purely frequentist interpretation as a discrepancy between (a) minimizing β for fixed α; versus (b) minimizing

[9] Also note that under this strategy, the frequentist results obey the likelihood principle and the stopping rule principle (e.g., Cornfield 1966; Lindley 1953; Pericchi and Pereira 2016).

the weighted sum of errors, $\lambda\alpha + \beta$.[10] A purely Bayesian version of the paradox will be provided in the next section.

A fully Bayesian version of the paradox

It is well known that the one-sided p value is asymptotically equal to the posterior mass lower than the point of test (e.g., Casella and Berger 1987; Lindley 1965; Pratt 1965; Marsman and Wagenmakers 2017 and references therein); for some problems, the relation is exact. This means that the p value can be given a Bayesian interpretation as the (approximate) probability that the observed effect has the wrong sign. Specifically, the odds form $(1 - p)/p$ is an approximation for BF_{+-}, that is, the Bayes factor for $\mathcal{H}_+ : \delta > 0$ versus $\mathcal{H}_- : \delta < 0$: a Bayesian test for the direction of an effect size δ. Jeffreys considered this a problem of estimation rather than of testing:

> "It should be said that several of the P integrals have a definite place in the present theory, in problems of pure estimation. For the normal law with a known standard error, or for those sampling problems that reduce to it, the total area of the tail represents the probability, given the data, that the estimated difference has the wrong sign-provided that there is no question whether the difference is zero.(...) They give the correct answer if the question is: If there is nothing to require consideration of some special values of the parameter, what is the probability distribution of that parameter given the observations?" (Jeffreys 1961, pp. 387-388; see also Jeffreys 1939, pp. 317-318)

The relation between the one-sided p value and the Bayesian test for direction suggests that the Jeffreys–Lindley paradox can be given a fully Bayesian interpretation. Specifically, data may be constructed which will convince the Bayesian that the population effect is positive rather than negative (i.e., $p(\delta > 0 \mid y, \mathcal{H}_1) \gg p(\delta < 0 \mid y, \mathcal{H}_1)$, whereas this same Bayesian will also be convinced that the population effect is absent rather than present (i.e., $p(\delta = 0 \mid y) \gg p(\delta \neq 0 \mid y)$). Let BF_{+-} denote $p(\delta > 0 \mid y, \mathcal{H}_1) \,/\, p(\delta < 0 \mid y, \mathcal{H}_1)$. Suppose data are constructed such that BF_{+-} is constant. As n increases, the evidence that the effect is absent rather than present will increase without bound, and this ensures that, with sufficiently high n, the Bayesian will believe that the effect is positive rather than negative, and simultaneously believe that it is absent rather than present. This state of knowledge is not incoherent, but it may be counter-intuitive.

A concrete demonstration of the fully Bayesian version of the paradox is given in Fig. 1. Each panel concerns the same Bayesian one-sample t test (Jeffreys 1948) and shows prior and posterior distributions on effect size $\delta = \mu/\sigma$; the prior distribution on δ is a zero-centered Cauchy with scale $1/\sqrt{2}$ (e.g., Gronau et al. 2020; Morey and Rouder 2018). In all three panels, the t values and sample sizes were chosen such that $p(\delta < 0 \mid y, \mathcal{H}_1) = 0.02041783$; thus, $\text{BF}_{+-} = 47.9768$, indicating strong evidence that the population effect is positive rather than negative.

[10] A reviewer suggested that the above considerations are moot in case the alternative hypothesis is composite, as the test-relevant parameter cannot be averaged out in the frequentist paradigm.

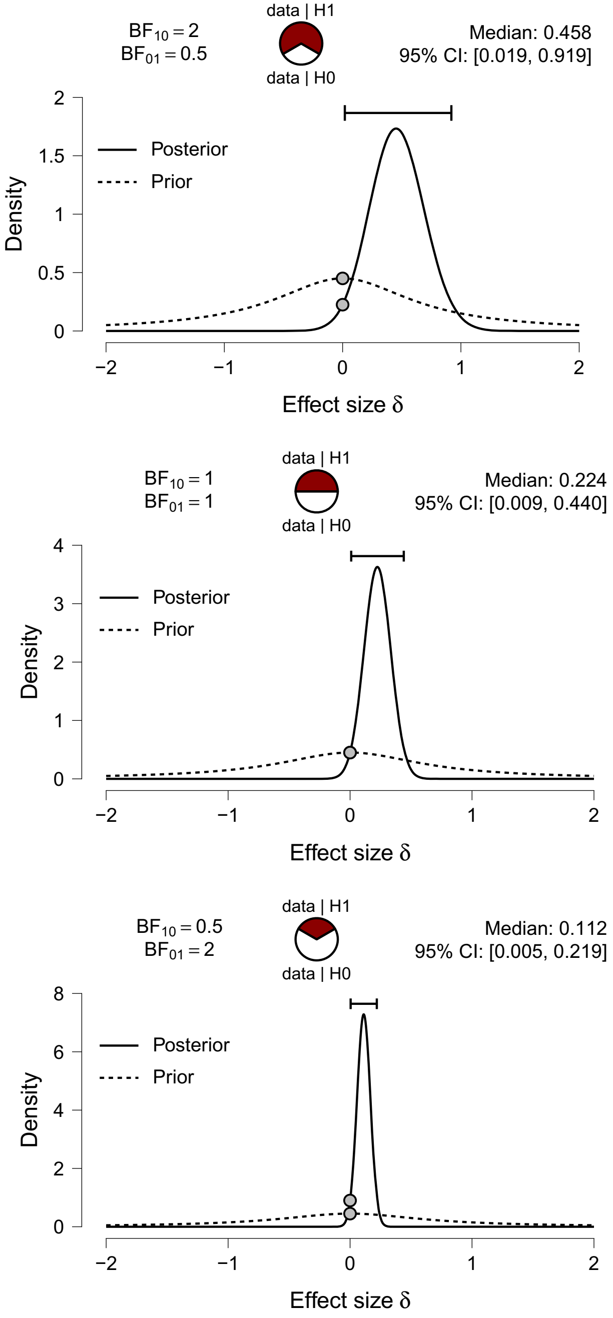

Fig. 1 Fully Bayesian version of the Jeffreys–Lindley paradox, illustrated with the t test. All panels have the same posterior mass on negative effect size: $p(\delta < 0 \mid y, \mathcal{H}_1) = 0.02041783$; thus, $\mathrm{BF}_{+-} = 47.9768$. As sample size n grows, $\mathcal{H}_0$ receives increasing support from the data. Top: $t = 2.321$, $n = 20$. Middle: $t = 2.113$, $n = 82$. Bottom: $t = 2.062$, $n = 332$. See text for details. Figures from JASP (jasp-stats.org)

The top, middle, and bottom panels have 20, 82, and 332 observations, respectively. As sample size increases from top to bottom, the posterior distribution narrows and shifts towards zero. As a result, the Bayes factor increasingly favors $\mathcal{H}_0$ over $\mathcal{H}_1$. In the top panel, $\text{BF}_{10} = 2$ (i.e., weak evidence in favor of the presence of an effect); in the middle panel, $\text{BF}_{10} = 1$ (i.e., complete absence of evidence); and in the bottom panel, $\text{BF}_{10} = 1/2$ (i.e., weak evidence in favor of the absence of an effect).[11]

The pattern shown in Fig. 1 can be appreciated by recourse to the Savage–Dickey density ratio (e.g., Dickey 1971; Verdinelli and Wasserman 1995; Wetzels et al. 2010). Under mild assumptions, this density ratio states that $\text{BF}_{10} = p(\delta = 0 \mid \mathcal{H}_1) \,/\, p(\delta = 0 \mid y, \mathcal{H}_1)$. In other words, the Bayes factor is given by the ratio of prior to posterior ordinate for δ under $\mathcal{H}_1$ at the point of test. The ordinate of the Cauchy prior distribution at $\delta = 0$ equals approximately 0.45. When $\text{BF}_{10} = 2$, this implies that the posterior ordinate equals $0.45/2$. This can be confirmed by a visual inspection of the two grey dots in the top panel from Fig. 1: the data have shifted the posterior distribution away from zero, lowering the ordinate at $\delta = 0$; consequently, the data favor $\mathcal{H}_1$ over $\mathcal{H}_0$. The middle panel shows that the prior ordinate equals the posterior ordinate, for a Bayes factor of 1, whereas the bottom panel shows that the posterior ordinate at $\delta = 0$ is now larger than the prior ordinate, indicating that the data favor $\mathcal{H}_0$ over $\mathcal{H}_1$.

The general rule is that when the observations accumulate indefinitely and the posterior distribution for δ becomes more peaked, retaining the same posterior mass on negative values of effect size (i.e., keeping BF_{+-} at a constant value) entails an increase of the posterior ordinate at $\delta = 0$; by the Savage–Dickey density ratio, this means more evidence for $\mathcal{H}_0$ (i.e., BF_{01} grows without bound). In sum, the paradox is also relevant within the framework of Bayesian statistics.

Two attempts to escape from the paradox

The Jeffreys–Lindley paradox inconveniences many statisticians. For frequentist statisticians, the paradox suggests that an epistemic interpretation of a p value requires that sample size is somehow taken into account—with a very large sample, a $p = 0.01$ result may well indicate strong support *in favor* of $\mathcal{H}_0$. For Bayesian statisticians, the paradox suggests that the quantification of evidence hinges on the specification of the test-relevant prior distribution under $\mathcal{H}_1$—this essentially prohibits the use of vague or improper priors.[12] Perhaps for this reason both frequentist and Bayesian statisticians have sought to defang the paradox by questioning Jeffreys's core assumptions. The main objections fall in two categories that will be discussed in turn; the first objection concerns the specification of $\mathcal{H}_0$, whereas the second objection finds fault with the specification of $\mathcal{H}_1$.[13]

[11] Top, middle, and bottom panels have a one-sided p value of .016, .019, and .020, respectively.

[12] As mentioned in the section on Bartlett's article, Jeffreys was well aware of this and suggested that different prior distributions be used for testing vs. estimation (cf. Jeffreys 1935, p. 207; Jeffreys 1948, p. 225).

[13] Robert and Rousseau (2011, p. 42).

Objection 1: "down with point masses!"

In Jeffreys's original development, prior mass 1/2 is assigned to the point-null hypothesis $\mathcal{H}_0$. One attempt to question the relevance of the paradox is to argue that the null hypothesis is never true exactly, and it is unwise to assign separate prior mass to a single point from a continuous distribution (e.g., de Bragança Pereira and Stern 1999, p. 109). For instance, Bernardo (2009) argues that

> "Jeffreys intends to obtain a posterior probability for a precise null hypothesis and, to do this, he is forced to use a mixed prior which puts a lump of probability $p = \Pr(H_0)$ on the null, say $H_0 \equiv \theta = \theta_0$, and distributes the rest with a *proper* prior $p(\theta)$ (he mostly chooses $p = 1/2$). This has a very upsetting consequence, usually known as Lindley's paradox (Lindley, 1957): for any fixed prior probability p independent of the sample sixe [sic] n, the procedure will wrongly accept H_0 whenever the likelihood is concentrated around a true parameter value which lies $O(n^{-\frac{1}{2}})$ from H_0. I find it difficult to accept a procedure which is *known* to produce the wrong answer under specific, but not controllable, circumstances (...)" (Bernardo 2009, p. 174; italics in original)

Moreover, in his paradox paper, Lindley (1957, p. 188) explicitly argues that prior mass needs to be assigned to a point in order for the paradox to arise: "...the phenomenon would persist with almost any prior probability distribution that had a concentration on the null value and no concentrations elsewhere. (...) It is, however, essential that the concentration on the null value exists, and it is this that has to be considered."

The impression that the paradox arises because $\mathcal{H}_0$ has separate prior mass is strengthened by Jeffreys's own work. Indeed, Jeffreys argued that his major conceptual advance over Laplace was the insight that, with moderate sample sizes, a general law can only ever receive compelling evidence when that law is assigned separate mass from the outset (Wrinch and Jeffreys 1921). As summarized by Jeffreys when he was 89 years old:

> "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 φ's 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)

The objection to the role of the point-null consists of two separate arguments, both of which need to hold: (1) the point-null $\mathcal{H}_0$ is never true exactly, and should, therefore, not be assigned separate mass; (2) only when $\mathcal{H}_0$ is assigned separate mass does the paradox manifest itself. With respect to the first argument, Jeffreys argued that assuming the falsity of the null without empirical evidence runs counter to scientific practice: "The onus of proof is always on the advocate of the more complicated hypothesis." (Jeffreys 1939, p. 278; echoed in Jeffreys 1961, p. 343; but see Gelman 2009). In addition, Jeffreys argued that assigning mass to the point-null hypothesis constitutes the best practical way of progress, yields better predictive performance, and prevents the haphazard inclusion of numerous parameters:

> "Some feeling of discomfort seems to attach itself to the assertion of the special value as *right*, since it may be slightly wrong but not sufficiently to be revealed by a test on the data available; but no significance test asserts it as certainly right. We are aiming at the best way of progress, not at the unattainable ideal of immediate certainty. What happens if the null hypothesis is retained after a significance test is that the maximum likelihood solution or a solution given by some other method of estimation is rejected. The question is, When we do this, do we expect thereby to get more or less correct inferences than if we followed the rule of keeping the estimation solution regardless of any question of significance? I maintain that the only possible answer is that we expect to get more. The difference as estimated is interpreted as random error and irrelevant to future observations. In the last resort, if this interpretation is rejected, there is no escape from the admission that a new parameter may be needed for every observation, and then all combination of observations is meaningless, and the only valid presentation of data is a mere catalogue without any summaries at all.(...)
> The distinction between problems of estimation and significance arises in biological applications, though I have naturally tended to speak mainly of physical ones. Suppose that a Mendelian finds in a breeding experiment 459 members of one type, 137 of the other. The expectations on the basis of a $3:1$ ratio would be 447 and 149. The difference would be declared not significant by any test. But the attitude that refuses to attach any meaning to the statement that the simple rule is right must apparently say that if any predictions are to be made from the observations the best that can be done is to make them on the basis of the ratio $459/137$, with allowance for the uncertainty of sampling. I say that the best is to use the $3/1$ rule, considering no uncertainty beyond the sampling errors of the new experiments. In fact the latter is what a geneticist would do. The observed result would be recorded and might possibly be reconsidered at a later stage if there was some question of differences of viability after many more observations had accumulated; but meanwhile it would be regarded as confirmation of the theoretical value. This is a problem of what I call significance." (Jeffreys 1939, pp. 318-320; echoed in Jeffreys 1961, pp. 388-389; italics in original)

With respect to the second argument—that the paradox manifests itself only when $\mathcal{H}_0$ is assigned separate mass, it should first be noted that the paradox may be formulated not on the level of posterior probabilities but on the level of Bayes factors, as

Jeffreys was wont to do. Thus, the paradox can be reformulated to state that data can always be found such that the p value suggests that $\mathcal{H}_0$ should be rejected, whereas the Bayes factor indicates that the same data provide strong support in *favor* of $\mathcal{H}_0$. Since the Bayes factor equals the ratio of marginal likelihoods under $\mathcal{H}_0$ and $\mathcal{H}_1$ it does not depend on the prior model probability that is assigned to $\mathcal{H}_0$ (cf. Pericchi 2011).

The second argument can also be countered directly: as we show below, the paradox does *not* require the presence of a point-null hypothesis. This fact is almost universally overlooked (for an exception see Cousins 2017). Thus, granting that the point-null hypothesis $\mathcal{H}_0 : \delta = 0$ is never true exactly, let us replace $\mathcal{H}_0$ by a peri-null hypothesis, say, $\widetilde{\mathcal{H}}_0 : \delta \sim \mathcal{N}(0, g_0)$ with variance g_0 small to reflect the skeptic's belief that the effect is near zero (e.g., Lindley 2011, Ly and Wagenmakers, in press-a; Morey and Rouder 2011). The peri-null does not include point masses; yet, the Jeffreys–Lindley paradox still applies. For instance, consider the z-test with data normally distributed $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ is known, say, $\sigma = 1$, and normal priors $\delta = \mu/\sigma \sim \mathcal{N}(0, g_k)$ for $k = 0, 1$ with $g_0 < g_1$. The peri-null Bayes factor is then

$$\mathrm{BF}_{\widetilde{0}1}(z, n) = \sqrt{\frac{1 + ng_1}{1 + ng_0}} \exp\left(\frac{(g_0 - g_1)nz^2}{2(1 + ng_0)(1 + ng_1)}\right). \tag{3}$$

Note that for the two-sided test with the α-threshold fixed, we have that $z = \Phi^{-1}(1-\alpha)$ where Φ^{-1} is the quantile function of a standard normal distribution. By definition of the Z-statistic the fixed α threshold can be expressed in terms of the sample mean and yields $\bar{y} = \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$. Observe that with a fixed α threshold, the value of $\bar{y}$ at which the null is rejected goes to zero as n increases. Plugging $z = \Phi^{-1}(1 - \alpha)$ into Eq. (3) shows that $\lim_{n\to\infty} \mathrm{BF}_{\widetilde{0}1}(z, n) = \sqrt{g_1/g_0}$. Since $g_1 > g_0$ this implies that $\mathrm{BF}_{\widetilde{0}1}$ will eventually provide evidence in favor of the peri-null hypothesis, even though $p \leq \alpha$ suggests a rejection of the null. The limit $\sqrt{g_1/g_0}$ is the maximum evidence for the peri-null that can be attained, as for all $\alpha \in (0, 1)$ the peri-null Bayes factor starts at one. Depending on g_0 and g_1, small values of n may result in a value of $\mathrm{BF}_{\widetilde{0}1}(z, n)$ that indicates some evidence for the alternative hypothesis; as n increases, $\mathrm{BF}_{\widetilde{0}1}(z, n)$ will monotonically increase towards $\sqrt{g_1/g_0}$. A specific demonstration is provided in Fig. 2.

The main effect of replacing the point-null hypothesis by a peri-null hypothesis is that for a fixed p value, the evidence in favor of the null no longer grows without bound. However, with $g_0 < g_1$, the peri-null evidence bound $\sqrt{g_1/g_0}$ still favors the null over the alternative for any non-zero α attained.

In sum, the Jeffreys–Lindley paradox does not depend on the presence of a point-null hypothesis, as is usually claimed. For fixed $p = \alpha$, the data will inevitably support a peri-null hypothesis over the alternative hypothesis as sample size grows large. The strength of this support is bounded, but in favor of the peri-null, thus leaving the conflict qualitatively intact. In other words, even when the point-null is replaced by a peri-null hypothesis, "there would be cases, with large numbers of observations, when

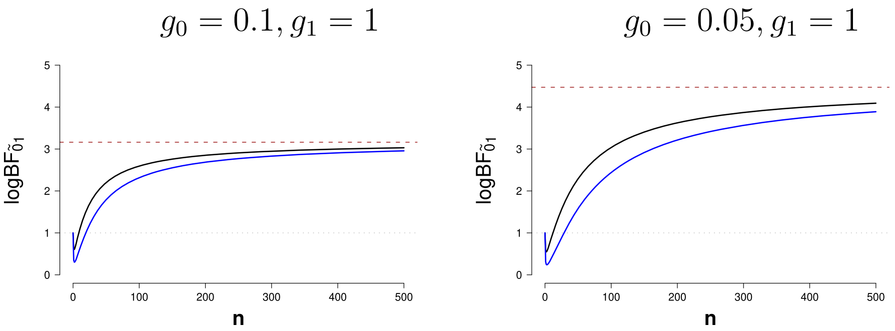

Fig. 2 Replacing the point-null hypothesis by a peri-null hypothesis does not avoid the Jeffreys–Lindley paradox. In the case of the Bayes factor z test, increasing sample size n for a fixed attained value of α inevitably results in positive evidence for the peri-null hypothesis. This evidence converges to an upper bound $\sqrt{g_1/g_0}$ that is indicated by the horizontal dashed brown line. The black and blue curves correspond to data that yield $\alpha = 0.05$ and $\alpha = 0.01$, respectively. Left panel: peri-null hypothesis with $g_0 = 0.1$; right panel: peri-null hypothesis with $g_0 = 0.05$

a new parameter is asserted on evidence that is actually against it.” (Jeffreys 1938a, pp. 379).

Objection 2: the paradox signals that the prior distribution was too wide

Whenever the paradox occurs, a natural objection to the Bayes factor outcome is that the prior distribution for the test-relevant parameter under $\mathcal{H}_1$ was too wide, wasting considerable prior mass on large values of effect size that yield poor predictive performance. Thus, as implied by Bartlett (1957), the paradox reveals a fault in the specification of $\mathcal{H}_1$ rather than $\mathcal{H}_0$.

This objection is valid in the sense that—as n increases and the p value remains constant—the increasingly poor predictive performance of $\mathcal{H}_1$ is indeed due to the fact that an increasing proportion of prior mass is inconsistent with the data, and this is the root of the paradox. For instance, the paradox would not arise if the predictions of $\mathcal{H}_1$ were evaluated under the maximum likelihood estimator $\hat{\theta}$ (Cousins 2017). However, as mentioned above, $\hat{\theta}$ is a cherry-picked value, and using it would favor $\mathcal{H}_1$ over $\mathcal{H}_0$ regardless of the data.

In general terms, the critique that the prior was too wide is made post hoc; after observing a near-zero effect size, one may always argue that, in hindsight, the prior was too wide—if such reasoning were allowed then the data could never undercut $\mathcal{H}_1$ and support $\mathcal{H}_0$. As long as the prior width does not shrink as a function of sample size, the paradox arises under any non-zero prior width.

Concluding comments

In this paper, we examined the history and nature of the Jeffreys–Lindley paradox. Our main conclusions are as follows:

1. Contrary to what the current literature suggest (e.g., Bernardo and Smith 2000, p. 394; O'Hagan and Forster 2004, p. 78), the Jeffreys–Lindley paradox was central to Harold Jeffreys's philosophy of Bayesian testing; in Jeffreys's tests, the critical threshold is not based on a constant multiple of the standard error but instead involves a $\sqrt{n}$ term.[14]
2. From 1935 to 1936, Jeffreys had discovered, understood, published, emphasized, explained, and illustrated the paradox. It remained a recurring theme throughout his later articles and books.
3. The articles by Lindley (1957) and Bartlett (1957) echo earlier work by Jeffreys. This is acknowledged by both authors, but they do not seem fully aware of the extent to which Jeffreys had already studied the issue. The two 1957 articles also introduced some mathematical errors and conceptual misunderstandings.[15]
4. The paradox is caused by the fact that as n increases and p remains constant, an ever increasing set of parameter values under $\mathcal{H}_1$ is inconsistent with the observed data, decreasing $\mathcal{H}_1$'s average predictive performance (i.e., the marginal likelihood).
5. A fully frequentist version of the paradox contrasts the inductive behavior of two frequentists, one who fixes α and minimizes β, the other who minimizes a weighted linear sum of α and β (e.g., Cornfield 1966; Lindley 1953; Lehmann 1958). As n grows large, the same data that prompt the former frequentist to reject $\mathcal{H}_0$ will prompt the latter frequentist to retain $\mathcal{H}_0$. The behavior of the latter frequentist is qualitatively consistent with the tests proposed by Jeffreys.
6. A fully Bayesian version of the paradox contrasts the beliefs of two Bayesians, one who tests $\mathcal{H}_+ : \delta > 0$ versus $\mathcal{H}_- : \delta < 0$ (i.e., the direction of the effect), the other who tests $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_1 : \delta \neq 0$ (i.e., the presence of the effect). As n grows large, the same data that prompt the former Bayesian to conclude that the data offer strong support for the hypothesis that the effect is positive will prompt the latter Bayesian to conclude that the data offer strong support for the hypothesis that the effect is absent.
7. Contrary to what the current literature suggest, the root of the paradox is not in the assignment of prior mass to a point hypothesis $\mathcal{H}_0$; the paradox is also present when the point-null hypothesis is replaced by a peri-null hypothesis (i.e., a relatively peaked continuous distribution).
8. The Jeffreys–Lindley paradox is relatively robust: it holds whether or not $\mathcal{H}_0$ is a point-null or a peri-null hypothesis, and it holds regardless of the width of the prior distribution for the test-relevant parameter under $\mathcal{H}_1$—as long as the width is larger than that of the prior distribution under the peri-null hypothesis, and as long as it does not shrink with sample size.
9. The Jeffreys–Lindley paradox results from the discrepancy between two modes of inference: (1) evaluating a single model (e.g., fixed-α decision making); (2) contrasting two models, one of which is relatively simple (e.g., the skeptic's $\mathcal{H}_0$) and

[14] For what it's worth, a Google search for "Lindley paradox" or "Lindley's paradox" yields about 6,190 results, whereas "Jeffreys–Lindley paradox" yields about 3.530 results; the phrase "Jeffreys's paradox" or "Jeffreys paradox" or "Jeffreys' paradox" yields 964 results (July 5th, 2021).

[15] In his later work, Jeffreys never cited the 1957 articles, perhaps because he felt these did not offer novel insights.

one which is more complex (e.g., the proponent's $\mathcal{H}_1$). In other words, the traditional frequentist test is absolute, whereas Jeffreys's Bayes factor test is relative.

We wish to emphasize that, when discussion the paradox, Lindley himself was always careful to credit Jeffreys (e.g., Robert 2013, p. 119: "Dennis systematically refereed [sic] to Jeffreys for stating the paradox, both in his paper and his personal communications."). However, it appears that Lindley did not fully appreciate the degree to which Jeffreys had worked on the paradox in the 1930s already. This may appear surprising, since Lindley had taken classes from Jeffreys; indeed, Lindley may be considered one of only a handful of statisticians who were keenly aware of Jeffreys's statistical methodology. A hint at the reason for this blind spot is given by Lindley himself, in a festschrift in honor of Jeffreys:

> "There have been several occasions on which one of us statisticians has asked Jeffreys about some point, and his answer has been "I dealt with that in the *Theory*" and he would go on to point out where. The questioner would then return to his room, take the book down from his shelf and sure enough, after some thinking, he would realize that the point was discussed there and that the discussion went some, if not the whole, way to answering the original question. In that last sentence I say "after some thinking" because Jeffreys's style does not give immediate comprehension. It is necessary to work at it. In my experience illumination usually appears and one wonders why it was so difficult to see at first. That is one reason why the book, although widely bought, has not been read or cited as much as it ought." (Lindley 1980, p. 119; italics in original)

We share Lindley's experience. In fact, we have studied Jeffreys's work for many years, and we have reread *Theory of Probability* several times over. Only recently did it dawn on us that the paradox was a central element of Jeffreys's statistical philosophy on hypothesis testing. We cannot offer a compelling explanation for why this was so difficult for us to see at first.

It is certainly the case that Jeffreys underplayed the differences between p values and Bayes factors from a pragmatic point of view. For instance, Jeffreys stated that "The rule that a difference becomes significant at about two or three times its standard error is therefore about right for ordinary numbers of observations." (Jeffreys 1935, p. 213) and "Thus even though P tests sometimes theoretically assert $\sim q$ when the number of observations is large and my tests support q, the occasions will be extremely rare." (Jeffreys 1939, p. 360, echoed in Jeffreys 1961, p. 435). Moreover, Jeffreys felt that in such cases the data often indicate model misspecification, in the sense that both the null hypothesis and the alternative hypothesis are found wanting, and a closer consideration of the data may suggest a third alternative (e.g., Jeffreys 1938d, p. 310; Jeffreys 1961, p. 436).

Jeffreys's assessment of the p value as "about right for ordinary numbers of observations" conflicts with the assessment of later Bayesians (e.g., Berger and Delampady 1987; Edwards 1965; Edwards et al. 1963; Sellke et al. 2001), who have argued that p values just below 0.05 do not constitute compelling evidence against $\mathcal{H}_0$. Jeffreys's relatively mild assessment is due to the fact that he calibrated $p = 0.05$ to $\text{BF}_{10} = 1$. However, it may be argued that in order to "reject" the null hypothesis we need strong evidence, or at least not evidence that is “hardly worth mentioning” (when $1 < \mathrm{BF}_{10} < 3$; Jeffreys 1939, p. 357). In addition, few researchers will consider a p value of exactly 0.05 as the point where they believe the data to be entirely uninformative. Would Jeffreys have endorsed the recent proposal to reduce the significance level for new discoveries from $\alpha = 0.05$ to $\alpha = 0.005$ (Benjamin et al. 2018)? We believe he would have had reservations. Although the proposal was motivated in part by Bayesian insights that originate from Jeffreys himself, the stricter α level still entails a threshold that is a constant multiple of the standard error and omits the crucial $\sqrt{n}$ term. Moreover, endorsing the $\alpha = 0.005$ proposal would mean an implicit admission that his repeated reassurances concerning the use of $\alpha = 0.05$ as “about right” were in fact wrong.

A thorough understanding of the Jeffreys–Lindley paradox remains critically important for the assessment of statistical methodology, both old and new. Ultimately, the paradox may even bring about some reconciliation between the Bayesian and the frequentist frameworks—in particular, the paradox may motivate frequentists to explore procedures that minimize the weighted sum of α and β, which ought to yield conclusions similar to those obtained with Jeffreys’s Bayesian tests (cf. Lindley 1953; Pericchi and Pereira 2016). We believe this manuscript provides some new historical and conceptual background to the Jeffreys–Lindley paradox, and we hope that this will be useful for statistical theory as well as statistical practice.

Acknowledgements The authors declare that they coordinate the development of the open-source software package JASP (https://jasp-stats.org), a non-commercial, publicly-funded effort to make Bayesian and non-Bayesian statistics accessible to a broader group of researchers and students.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix: Jeffreys discusses the paradox post 1957

As far as the paradox-related material in Jeffreys’s books is concerned, the 1961 third edition of *Theory of Probability* does not add anything to the 1948 second edition (cf. Jeffreys 1948, pp. 221–222, p. 399 to Jeffreys 1961, p. 248, p. 435), which itself did not add much to the 1939 first edition (Jeffreys 1939, p. 194, pp. 359–360). Likewise, the 1973 third edition of *Scientific Inference* repeats the short relevant fragment from the 1957 second edition provided in the main text (cf. Jeffreys 1973, pp. 74–75 to Jeffreys 1957a, pp. 71–72).

Jeffreys does touch on the paradox in three papers published after 1957. First, in the 1974 article *Fisher and inverse probability*, Jeffreys hints at the paradox when he writes:

> “I think that astronomers had found much earlier that discrepancies up to twice the standard error usually disappeared when more information became available, but those over three times usually persisted. In fact, *with ordinary numbers of observations*, say 10 to 500, these rough rules are usually not far from the 95 per cent and 99 per cent rules or from the more detailed ones that I derive in my *Theory of Probability*.” (Jeffreys 1974, p. 2; first italics added for emphasis)

Later, the 1977 article *Probability theory in geophysics* contains a relevant fragment that is highly similar to Jeffreys (1957b, p. 349) cited above:

> “The theory leads to rules of significance for changes in laws, involving the introduction of new parameters in laws. They are usually approximately of the form
>
> $$K = \frac{P(q \mid \theta p)}{P(q^{'} \mid \theta p)} \doteqdot (An)^{\frac{1}{2}} \exp\left(-\frac{a^2}{2s_a^2}\right).$$
>
> Here q is the hypothesis that the new parameter α is zero, that is, that the previous law needs no alteration; q' the hypothesis that α is needed, having a value to be estimated from the observations; a and s_a are the estimate of α and its standard error as given by the method of least squares; n is the number of observations; and A is a constant, usually not far from 1. If $a < s_a$, the factor $n^{\frac{1}{2}}$ makes $K > 1$ and the old law is supported; but with ordinary numbers of observations, if $a > 2s_a$ or $3s_a$, $K < 1$ and the new law is supported. To apply a test of this sort it is of course of the first importance that the number of observations shall be stated. This is in fact not often done by physicists, but thanks mainly to the work of Fisher (with whom I do not always agree) biologists usually do it, but with different rules. I once remarked to Fisher that in nearly all practical applications we should agree, and that when we differed we should both be doubtful.” (Jeffreys 1977, p. 89)

Finally, in 1980 Jeffreys published the chapter *Some general points in probability theory* in the book *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys*. Jeffreys summarizes his contributions and concludes as follows:

> “Many complications have been dealt with. The usual form, if y is used for the observational data, is approximately
>
> $$K = \frac{P(\mathcal{H}_0 \mid y)}{P(\mathcal{H}_1 \mid y)} = An^{1/2} \exp\left\{-\frac{(a-\alpha_0)^2}{2s_a^2}\right\},$$
>
> where A is of order 1, n the number of observations, a and s_a the estimates by maximum likelihood of the new parameter and its standard error. If $a < s_a$ and n is large we get strong confirmation that no change in α is needed; if $a - \alpha_0$ is several times s_a there is strong support for a change. For n from about 10 to 500 the usual result is that $K = 1$ when $(a-\alpha_0)/s_a$ is about 2, $10^{-1/2}$ when it is about 2.7, 10^{-1} about 3.2, and $10^{-}2$ about 4. These are not far from the rough rule long known to astronomers, i.e., that differences up to twice the standard error

> usually disappear when more or better observations become available, and that those of three or more time usually persist. They are also not far from the 0.05, 0.01 and so on limits for the usual P. 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. As an argument astronomer's experience is far better. P has a definite place when we already know what parameters are relevant, and we want to know their amounts; this is what I call a problem of estimation. A problem of significance is one where we are considering a change in the form of the law itself." (Jeffreys 1980, p. 453)

References

Aitkin, M. 1991. Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)* 53: 111–142.

Andrews, D.W.K. 1994. The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica* 62: 1207–1232.

Bartlett, M.S. 1957. A comment on D. V. Lindley's statistical paradox. *Biometrika* 44: 533–534.

Bayarri, M.J., J.O. Berger, A. Forte, and G. García-Donato. 2012. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics* 40: 1550–1577.

Benjamin, D.J., J.O. Berger, M. Johannesson, B.A. Nosek, E.-J. Wagenmakers, R. Berk, and V.E. Johnson. 2018. Redefine statistical significance. *Nature Human Behaviour* 2: 6–10.

Bennett, J.H., ed. 1990. *Statistical inference and analysis: Selected correspondence of R. A. Fisher*. Oxford: Clarendon Press.

Berger, J.O., and M. Delampady. 1987. Testing precise hypotheses. *Statistical Science* 2: 317–352.

Berkson, J. 1942. Tests of significance considered as evidence. *Journal of the American Statistical Association* 37: 325–335.

Bernardo, J.M. 1980. A Bayesian analysis of classical hypothesis testing (with discussion). *Trabajos de Estadistica y de Investigacion Operativa* 31: 605–647.

Bernardo, J.M. 2009. [Harold Jeffreys's theory of probability revisited]: Comment. *Statistical Science* 24: 173–175.

Bernardo, J.M. 2011. Integrated objective Bayesian estimation and hypothesis testing. In *Bayesian statistics*, vol. 9, ed. J.M. Bernardo, et al., 1–68. Oxford: Oxford University Press.

Bernardo, J.M., and A.F.M. Smith. 2000. *Bayesian theory*. Chichester: Wiley.

Berrar, D., and W. Dubitzky. 2017. On the Jeffreys-Lindley paradox and the looming reproducibility crisis in machine learning. In 2017 IEEE international conference on data science and advanced analytics (DSAA) (pp. 334–340).

Burnham, K.P., and D.R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* 33: 261–304.

Casella, G., and R.L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association* 82: 106–111.

Colquhoun, D. 2019. The false positive risk: A proposal concerning what to do about p-values. *The American Statistician* 73: 192–201.

Consonni, G., D. Fouskakis, B. Liseo, and I. Ntzoufras. 2018. Prior distributions for objective Bayesian analysis. *Bayesian Analysis* 13: 627–679.

Cornfield, J. 1966. Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician* 20: 18–23.

Cousins, R.D. 2017. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese* 194: 395–432.

Cox, D.R. 2006. *Principles of statistical inference*. Cambridge: Cambridge University Press.

de Bragança Pereira, C.A., and J.M. Stern. 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy* 1: 99–110.

de Bragança Pereira, C.A., J.M. Stern, and S. Wechsler. 2008. Can a significance test be genuinely bayesian. *Bayesian Analysis* 3: 79–100.

DeGroot, M.H., and M.J. Schervish. 2012. *Probability and statistics*, 4th ed. New York: Addison-Wesley.

Dickey, J.M. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics* 42: 204–223.

Edwards, W. 1965. Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin* 63: 400–402.

Edwards, W., H. Lindman, and L.J. Savage. 1963. Bayesian statistical inference for psychological research. *Psychological Review* 70: 193–242.

Etz, A., and E.-J. Wagenmakers. 2017. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science* 32 (2): 313–329.

Fienberg, S.E. 2003. When did Bayesian inference become "Bayesian? *Bayesian Analysis* 1: 1–41.

Fisher, R.A. 1934. *Statistical methods for research workers*, 5th ed. London: Oliver and Boyd.

Fisher, R.A. 1935. *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R.A. 1936. *Statistical methods for research workers*, 6th ed. London: Oliver and Boyd.

Freeman, P.R. 1993. The role of p-values in analysing trial results. *Statistics in Medicine* 12: 1443–1452.

Gelman, A. 2009. Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science* 24: 176–178.

Good, I.J. 1980. The contributions of Jeffreys to Bayesian statistics. In *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys*, ed. A. Zellner, 21–34. Amsterdam: North-Holland Publishing Company.

Good, I.J. 1980. The diminishing significance of a p-value as the sample size increases. *Journal of Statistical Computation and Simulation* 11: 307–313.

Good, I.J. 1983. The diminishing significance of a fixed p-value as the sample size increases: A discrete model. *Journal of Statistical Computation and Simulation* 16: 312–313.

Good, I.J. 1992. The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association* 87: 597–606.

Gronau, Q.F., A. Ly, and E.-J. Wagenmakers. 2020. Informed Bayesian t-tests. *The American Statistician* 74: 137–143.

Howie, D. 2002. *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Jaynes, E.T. 2003. *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jefferys, W.H. 1990. Bayesian analysis of random event generator data. *Journal of Scientific Exploration* 4: 153–169.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society* 31: 203–222.

Jeffreys, H. 1936a. On some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 22: 337–359.

Jeffreys, H. 1936b. Further significance tests. *Mathematical Proceedings of the Cambridge Philosophical Society* 32: 416–445.

Jeffreys, H. 1937a. The tests for sampling differences and contingency. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 162: 479–495.

Jeffreys, H. 1937b. *Scientific inference*, 1st ed. Cambridge: Cambridge University Press.

Jeffreys, H. 1937c. Modern Aristotelianism: Contribution to discussion. *Nature* 139: 1004.

Jeffreys, H. 1938a. The comparison of series of measures on different hypotheses concerning the standard errors. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 167: 367–384.

Jeffreys, H. 1938b. Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 165: 161–198.

Jeffreys, H. 1938c. Maximum likelihood, inverse probability and the method of moments. *Annals of Eugenics* 8: 146–151.

Jeffreys, H. 1938d. Significance tests for continuous departures from suggested distributions of chance. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 164: 307–315.

Jeffreys, H. 1938e. Aftershocks and periodicity in earthquakes. *Gerlands Beiträge zur Geophysik* 53: 111–139.

Jeffreys, H. 1939. *Theory of probability*, 1st ed. Oxford: Oxford University Press.

Jeffreys, H. 1940. Note on the Behrens-Fisher formula. *Annals of Eugenics* 10: 48–51.

Jeffreys, H. 1942. On the significance tests for the introduction of new functions to represent measures. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 180: 256–268.

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

Jeffreys, H. 1950. Bertrand russell on probability. *Mind: A Quarterly Review of Psychology and Philosophy* 59: 313–319.

Jeffreys, H. 1953. Comment on "statistical inference" by Dennis Lindley. *Journal of the Royal Statistical Society Series B (Methodological)* 15: 72.

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

Jeffreys, H. 1957. *Scientific inference*, 2nd ed. Cambridge: Cambridge University Press.

Jeffreys, H. 1957. probability theory in astronomy. *Monthly Notices of the Royal Astronomical Society* 117: 347–355.

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

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

Jeffreys, H. 1974. Fisher and inverse probability. *International Statistical Review* 42: 1–3.

Jeffreys, H. 1977. Probability theory in geophysics. *Journal of the Institute of Mathematics and its Applications* 19: 87–96.

Jeffreys, H. 1980. Some general points in probability theory. In *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys*, ed. A. Zellner, 451–453. Amsterdam: North-Holland Publishing Company.

Kamary, K., K. Mengersen, C.P. Robert, and J. Rousseau. 2014. Testing hypotheses via a mixture estimation model. arXiv:1412.2044.

Kass, R.E., and A.E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795.

Kass, R.E., and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90: 928–934.

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

Kim, J.H., and I. Choi. 2021. Choosing the level of significance: A decision-theoretic approach. *Abacus* 57: 27–71.

Leamer, E. 1978. *Specification searches: ad hoc inference with nonexperimental data*. New York: Wiley.

Lehmann, E.L. 1958. Significance level and power. *The Annals of Mathematical Statistics* 29: 1167–1176.

Lindley, D.V. 1953. Statistical inference. *Journal of the Royal Statistical Society Series B (Methodological)* 15: 30–76.

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

Lindley, D.V. 1965. *Introduction to probability and statistics from a Bayesian viewpoint. Part 2. Inference*. Cambridge: Cambridge University Press.

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

Lindley, D.V. 1986. Comment on "tests of significance in theory and practice" by D. J. Johnstone. *Journal of the Royal Statistical Society. Series D (The Statistician)* 35: 502–504.

Lindley, D.V. 1989. Obituary: Harold Jeffreys, 1891–1989. *Journal of the Royal Statistical Society Series A* 152: 417–419.

Lindley, D.V. 2000. What is a Bayesian? *The ISBA Bulletin* 7: 7–9.

Lindley, D.V. 2011. Comment on "integrated objective Bayesian estimation and hypothesis testing" by J. M. Bernardo. In *Bayesian statistics*, vol. 9, ed. J.M. Bernardo, et al., 37–38. Oxford: Oxford University Press.

Ly, A., A. Stefan, J. van Doorn, F. Dablander, D. van den Bergh, A. Sarafoglou, and E.-J. Wagenmakers. 2020. The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the *p*-value hypothesis test. *Computational Brain and Behavior* 3: 153–161.

Ly, A., A.J. Verhagen, and E.-J. Wagenmakers. 2016a. Harold Jeffreys's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. *Journal of Mathematical Psychology* 72: 19–32.

Ly, A., A.J. Verhagen, and E.-J. Wagenmakers. 2016b. An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology* 72: 43–55.

Ly, A., and E.-J. Wagenmakers. (in press-a). Bayes factors for peri-null hypotheses. TEST. arXiv:2102.07162.

Ly, A., and E.-J. Wagenmakers. (in press-b). A critical evaluation of the FBST *ev* for Bayesian hypothesis testing. Computational Brain and Behavior. https://psyarxiv.com/x9t6n/.

Maier, M., and D. Lakens. (in press). Justify your alpha: a primer on two practical approaches. Advances in Methods and Practices in Psychological Science. https://psyarxiv.com/ts4r6.

Marsman, M., and E.-J. Wagenmakers. 2017. Three insights from a Bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement* 77: 529–539.

Morey, R.D., and J.N. Rouder. 2011. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods* 16: 406–419.

Morey, R.D., and J.N. Rouder. 2018. BayesFactor 0.9.124.2. Comprehensive R Archive Network. http://cran.r-project.org/web/packages/BayesFactor/index.html.

Mudge, J.F., L.F. Baker, C.B. Edge, and J.E. Houlahan. 2012. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One* 7: e32734.

Nasir, M.A., A.M. Soliman, M. Shahbaz, et al. 2020. Operational aspect of the policy coordination for financial stability: role of Jeffreys-Lindley's paradox in operations research. *Annals of Operations Research* 20: 1–25.

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

Ormerod, J.T., M. Stewart, W. Yu, and S.E. Romanes. 2017. Bayesian hypothesis tests with diffuse priors: can we have our cake and eat it too? Manuscript submitted for publication. https://arxiv.org/pdf/1710.09146.pdf.

Pearson, E.S. 1953. Comment on "statistical inference" by Dennis Lindley. *Journal of the Royal Statistical Society Series B (Methodological)* 15: 68–69.

Pérez, M.-E., and L.R. Pericchi. 2014. Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics and Probability Letters* 85: 20–24.

Pericchi, L.R. 2011. Comment on "integrated objective Bayesian estimation and hypothesis testing" by J. M. Bernardo. In *Bayesian statistics*, vol. 9, ed. J.M. Bernardo, et al., 25–29. Oxford: Oxford University Press.

Pericchi, L.R., and C. Pereira. 2016. Adaptative significance levels using optimal decision rules: balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics* 30: 70–90.

Pratt, J.W. 1965. Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society B* 27: 169–203.

Robert, C.P. 1993. A note on Jeffreys–Lindley paradox. *Statistica Sinica* 3: 601–608.

Robert, C.P. 2013. On the Lindley–Jeffreys paradox. In *A book for Dennis*, ed. A. O'Hagan, 118–122. San Francisco: Blurb.

Robert, C.P. 2014. On the Lindley–Jeffreys paradox. *Philosophy of Science* 81: 216–232.

Robert, C.P., N. Chopin, and J. Rousseau. 2009. Harold Jeffreys's theory of probability revisited. *Statistical Science* 24: 141–172.

Robert, C.P., and J. Rousseau. 2011. Comment on "integrated objective Bayesian estimation and hypothesis testing" by J. M. Bernardo. In *Bayesian statistics*, vol. 9, ed. J.M. Bernardo, et al., 41–44. Oxford: Oxford University Press.

Royall, R. 1986. The effect of sample size on the meaning of significance tests. *The American Statistician* 40: 313–315.

Royall, R.M. 1997. *Statistical evidence: a likelihood paradigm*. London: Chapman and Hall.

Savage, L.J. 1964. The foundations of statistics reconsidered. In *Studies in subjective probability*, ed. H.E. Kyburg and H.E. Smokler, 173–188. New York: Wiley.

Savage, L.J., M.S. Bartlett, G.A. Barnard, D.R. Cox, E.S. Pearson, C.A.B. Smith, and C.B. Winsten. 1962. *The foundations of statistical inference*. London: Methuen.

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

Senn, S. 2001. Two cheers for P-values? *Journal of Epidemiology and Biostatistics* 6: 193–204.

Shafer, G. 1982. Lindley's paradox. *Journal of the American Statistical Association* 77: 325–351.

Spanos, A. 2013. Who should be afraid of the Jeffreys–Lindley paradox? *Philosophy of Science* 80 (1): 73–93.

Sprenger, J. 2013. Testing a precise null hypothesis: The case of Lindley's paradox. *Philosophy of Science* 80 (5): 733–744.

Szabó, B., and A. van der Vaart. 2019. *Bayesian statistics [lecture notes]*. Leiden: Leiden University.

Vehtari, A., A. Gelman, and J. Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27: 1413–1432.

Verdinelli, I., and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* 90: 614–618.

Villa, C., and S. Walker. 2017. On the mathematics of the Jeffreys–Lindley paradox. *Communications in Statistics Theory and Methods* 46: 12290–12298.

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

Wagenmakers, E.-J., Q.F. Gronau, F. Dablander, and A. Etz. (in press). The support interval. Erkenntnis. https://psyarxiv.com/zwnxb/.

Wagenmakers, E.-J., A.J. Verhagen, A. Ly, D. Matzke, H. Steingroever, J.N. Rouder, and R.D. Morey. 2017. The need for Bayesian hypothesis testing in psychological science. In *Psychological science under scrutiny: recent challenges and proposed solutions*, ed. S.O. Lilienfeld and I. Waldman, 123–138. New York: Wiley.

Wasserstein, R.L., and N.A. Lazar. 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70 (2): 129–133.

Wetzels, R., R.P.P.P. Grasman, and E.-J. Wagenmakers. 2010. An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis* 54: 2094–2102.

Wrinch, D., and H. Jeffreys. 1919. On some aspects of the theory of probability. *Philosophical Magazine* 38: 715–731.

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

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

Yin, G., and H. Shi. 2020. Demystify Lindley's paradox by interpreting p-value as posterior probability. arXiv:2002.10883 (arXiv preprint).

Zellner, A. 1971/1996. An introduction to Bayesian inference in econometrics. New York: Wiley.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.