Archive for History of Exact Sciences (2023) 77:443–449
https://doi.org/10.1007/s00407-023-00310-4

EDITORIAL

The Jeffreys–Lindley paradox: an exchange

Jeremy Gray[1] · Joshua L. Cherry[2] · Eric-Jan Wagenmakers[3] · Alexander Ly[4]

Published online: 30 May 2023

Abstract
This Editorial reports an exchange in form of a comment and reply on the article "History and Nature of the Jeffreys–Lindley Paradox" (Arch Hist Exact Sci 77:25, 2023) by Eric-Jan Wagenmakers and Alexander Ly.

1 Comments by J. Gray, Editor-in-Chief

AHES does not normally publish correspondence about an article, and would prefer to see scholarly disagreements dealt with in the form of subsequent articles. We are making an exception in this case because of the significant difficulties involved in the interpretation and mathematical formulation of the Jeffreys–Lindley paradox, and the way they affect historical interpretations.

2 Comment on "History and nature of the Jeffreys–Lindley paradox" by J. L. Cherry

(Joshua L. Cherry is a US government employee his comment cannot be copyrighted)

The Jeffreys–Lindley paradox involves disagreements between classical and Bayesian null hypothesis tests applied to the same observations. The two approaches can lead to opposite conclusions: the classical *p*-value may be low enough that the null hypothesis would be rejected while the Bayesian posterior probability favors the null. The paradox, according to Lindley (1957), is that whatever the prior probability of

✉ Jeremy Gray
j.j.gray@open.ac.uk

1 Faculty of Mathematics and Computer Sciences, The Open University, Milton Keynes MK7 6AA, England, UK

2 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

3 University of Amsterdam, Amsterdam, The Netherlands

4 Centrum Wiskunde and Informatica, University of Amsterdam, Amsterdam, The Netherlands

the null, this disagreement can be arbitrarily strong for sufficiently large sample size. Equivalently, the Bayes factor in favor of the null (BF_{01}) corresponding to any fixed *p*-value grows without bound as the sample size increases. This result holds for any alternative hypothesis satisfying reasonable conditions.

Lindley (1957) discussed the dependence of this result on the assignment of probability mass to a single parameter value by the null hypothesis. Wagenmakers and Ly (2023) recently claimed that a point mass is not necessary for the paradox. They purported to demonstrate that the paradox remains if the point null is replaced by a peri-null (a continuous distribution centered on zero effect size). There are, however, two problems with their argument. First, the Bayes factor in favor of the peri-null approaches a finite limit, rather than infinity as required for Lindley's result. This is not an inconsequential detail: it means that the posterior probability need not favor the null in the limit. Second, their derivation implicitly assumes a point null for the classical test, using the peri-null only for Bayes factor calculation. It is in no sense paradoxical that one approach rejects a null hypothesis while another supports a *different* null hypothesis. If the same peri-null is used for both tests, as is appropriate for comparing them, the Bayes factor behaves quite differently in the limit of interest.

The paradox described by Lindley requires that BF_{01} grow without bound as the sample size increases, but with a peri-null it approaches a finite value. Wagenmakers and Ly state that this leaves the paradox "qualitatively intact". However, the difference between finite and infinite limits, arguably qualitatively important in general, is in this case critical. The paradox, as stated by Lindley (1957), requires that the posterior probability favors the null for any nonzero value of its prior probability. For any value of BF_{01}, the posterior probability will favor the alternative hypothesis if this prior probability is sufficiently low. Thus, unless the Bayes factor goes to infinity, the posterior need not favor the null in the limit. As the Bayesian conclusion is supposed to depend on the prior probability—a feature of the approach often touted as an advantage by its proponents—this is certainly a qualitative difference, and an important one.

Even with respect to a weakened version of the paradox that considers only Bayes factors, the derivation of Wagenmakers and Ly is flawed. The essence of the Jeffreys-Lindley paradox is that two statistical approaches support contradictory conclusions about the same null hypothesis. Wagenmakers and Ly demonstrate only different conclusions about different null hypotheses: a peri-null for the Bayesian test but a point null for the classical. Different conclusions about different hypotheses are neither surprising nor indicative of differences between statistical approaches. Bayes factors for point null and peri-null hypotheses can differ in the same way, and necessarily do so in the analogous large-sample limit. An informative comparison would have to use the peri-null for the classical test as well as the Bayesian.

A *p*-value for the peri-null can be calculated using the distribution of the sample mean that it implies. Under the assumptions employed by Wagenmakers and Ly the sample effect size has a normal distribution with mean equal to zero and variance equal to $g_0 + 1/n$, where g_0 is the variance of the peri-null distribution and n is the sample size. As $n \to \infty$, the variance does not go to zero, as with a point null, but instead approaches g_0. Therefore the sample effect size corresponding to a particular *p*-value does not approach zero, but rather a nonzero multiple of $\sqrt{g_0}$, e.g., $\sim 1.96\sqrt{g_0}$ for $p = 0.05$.

The consequences for the corresponding Bayes factor are illustrated in Fig. 1, which can be compared to the left panel of Fig. 2 in Wagenmakers and Ly (2023). In the limit as $n \to \infty$, BF_{01} approaches a value less than one, indicating support for the alternative hypothesis in agreement with the low p-values. Thus, there is no paradox.

The Bayes factor may exceed unity in the limit for any particular p-value if the variance of the peri-null distribution is made sufficiently small. However, it will always be possible to choose a smaller value of p for which the asymptotic Bayes factor is smaller than one. This is illustrated in Fig. 2, where the narrower peri-null (compared to Fig. 1) raises BF_{01} above one for $p = 0.05$, but BF_{01} remains below one for $p = 0.01$. Thus, even if Lindley's statement of the paradox is weakened to require only $\mathrm{BF}_{01} >$

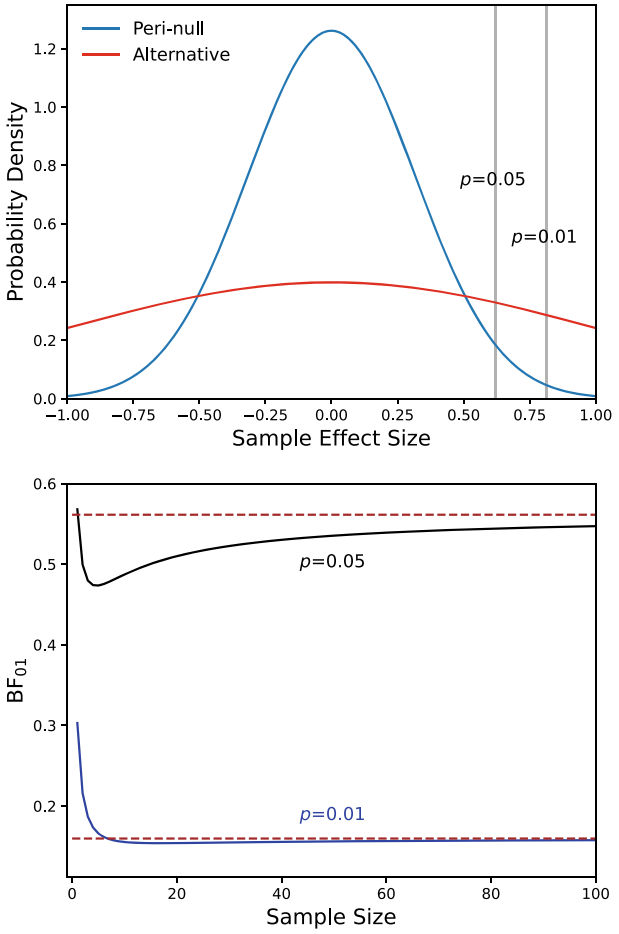

Fig. 1 Comparison of classical and Bayesian analyses when a peri-null hypothesis is used for both. Assumptions are as in Wagenmakers and Ly (2023), with $g_0 = 0.1$, $g_1 = 1$, as in the left panel of their Fig. 2. Top: distributions of the sample effect size under the null and alternative hypotheses in the limit of infinite sample size. Sample effect sizes corresponding to p-values of 0.05 and 0.01 are indicated by vertical lines. Bottom: BF_{01} as a function of sample size for p-values fixed at 0.05 or 0.01

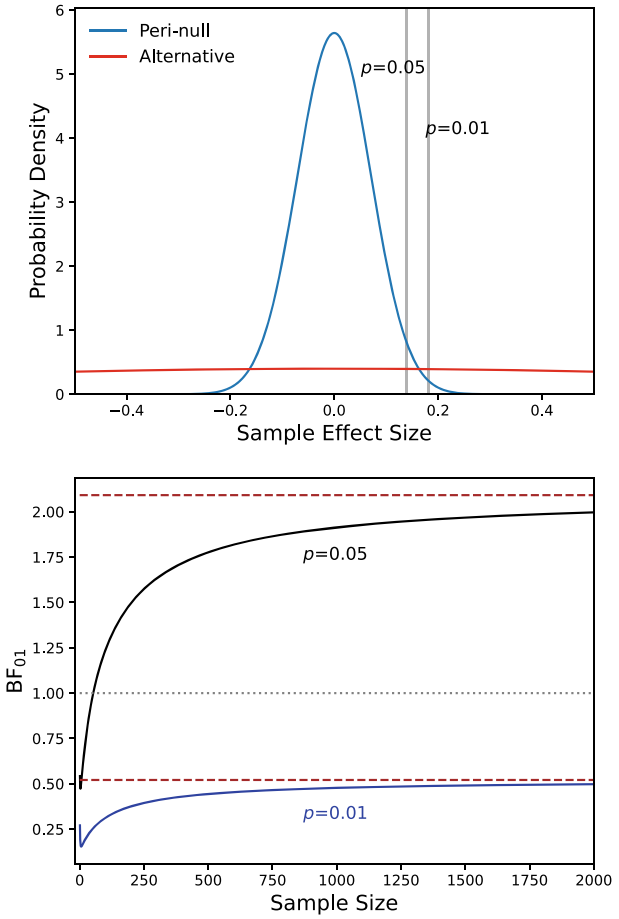

Fig. 2 Comparison of classical and Bayesian analyses with a narrower peri-null distribution. The variance of the peri-null, g_0, is 0.05. Assumptions are otherwise the same as in Fig. 1

1 rather than $\mathrm{BF}_{01} \to \infty$, the requirement that this hold for any nonzero p-value is not met with a peri-null, however narrow it may be.

Even for $p = 0.05$, the peri-null in Fig. 2 fails to produce another feature of the paradox: that it holds for any alternative hypothesis meeting reasonable regularity conditions. This feature was invoked by Wagenmakers and Ly ("the paradox arises under any non-zero prior width") to address an objection to the paradox ("the prior distribution was too wide"). With the peri-null in Fig. 2, the Bayes factor will come to agree with the p-value of 0.05 in supporting the alternative hypothesis if the alternative distribution is made sufficiently narrow, in violation of this aspect of the paradox. This situation resembles Fig. 1, but with the null and alternative distributions made narrower by the same factor.

The Jeffreys–Lindley paradox indeed depends on a null hypothesis that assigns probability mass to a single parameter value. It may be enlightening to consider weaker

results that apply with a peri-null, but their weakness should be appreciated. The paradox is more than the mere possibility of conflicting inferences about the null, which are also possible for Bayesian tests based on different prior probabilities or alternative hypotheses. The paradox entails that the Bayesian conclusion can favor the null for any prior probability, any p-value, and any alternative hypothesis satisfying regularity conditions. None of these aspects of the paradox are met with a peri-null.

Acknowledgements

This work was supported by the intramural research program of the National Library of Medicine, National Institutes of Health.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Lindley, D.V. 1957. A statistical paradox. *Biometrika* 44: 187–192. https://doi.org/10.1093/biomet/44.1-2.187.

Wagenmakers, Eric-Jan., and Alexander Ly. 2023. History and nature of the Jeffreys–Lindley paradox. *Archive for History of Exact Sciences*. 77:25–72. https://doi.org/10.1007/s00407-022-00298-3.

3 Reply to the Comment by J. L. Cherry

Authors: Eric-Jan Wagenmakers and Alexander Ly

The Jeffreys–Lindley paradox may be given different interpretations:

- As sample size increases indefinitely, there will be an inevitable conflict between any fixed positive p-value and the posterior probability for the null hypothesis $\mathcal{H}_0$. For instance, for the same data y^n as $n \to \infty$, we will simultaneously find $p = \epsilon$ ("confidently reject the null") and $p(\mathcal{H}_0 \mid y^n) = 1 - \epsilon$ ("the null is by far the more plausible model"). We term this interpretation the *strong form* of the Jeffreys–Lindley paradox (e.g., Lindley 1957, p. 187).
- As sample size increases indefinitely, there will be an inevitable conflict between any fixed positive p-value and the statistical evidence as quantified by the Bayes factor. For instance, for the same data y^n as $n \to \infty$, we will simultaneously find $p = \epsilon$ ("confidently reject the null") and $\text{BF}_{01} > 1$ ("the data are more likely under the null than under the alternative"). We term this interpretation the *original form* of the Jeffreys–Lindley paradox (e.g., Jeffreys 1938, p. 379).
- As sample size increases, any fixed positive p-value will correspond to a different degree of (Bayesian) conviction: "5% in to-day's small sample does not mean the same as 5% in to-morrow's large one." (Lindley 1957, p. 189). We term this interpretation the *weak form* of the Jeffreys–Lindley paradox.

In Wagenmakers and Ly (2023) we stated that the Jeffreys–Lindley paradox remains "qualitatively intact" when the Bayesian analysis replaces the point-null hypothesis with a peri-null hypothesis. More precisely, a Bayesian analysis with a peri-null hypothesis still produces the *original form* of the Jeffreys–Lindley paradox outlined above. In our opinion, this form of the paradox packs sufficient punch: presumably,

few researchers would wish to reject the null hypothesis when that hypothesis actually predicted the observed data better than the alternative hypothesis.

Cherry (2023) argues that the peri-null Bayes factor ought to be compared to a frequentist peri-null $\widetilde{p}$-value. Although at first glance this may seem fair, in our opinion this comparison is irrelevant in the context of the paradox. What is at stake is the epistemic status of the point-null p-value. Does a fixed point-null p-value mean the same in today's small sample as in tomorrow's large sample? The original form of the Jeffreys–Lindley paradox reveals an inevitable conflict between this p-value and the Bayes factor as sample size increases; can we perhaps pin the cause for this conflict on the use of a point-null hypothesis in the Bayesian analysis? In other words, could we argue that the conflict arises because the Bayesian analysis makes an indefensible assumption, namely that the null-hypothesis can be exactly true? Does the paradox then arise because the Bayesian analysis assigns prior mass to a point, which is possibly anomalous and at odds with Bayesian thinking? This is a popular line of argumentation, and we demonstrated that it is false.

Leaving aside the issue of relevance, we have two specific concerns about the peri-null $\widetilde{p}$-value. First, by adopting a peri-null $\widetilde{p}$-value, Cherry (2023) effectively changes the data-generating process. In the common z-test the two hypotheses that are being compared are

$$\mathcal{H}_0 : \bar{X} \sim \mathcal{N}(0, 1/n) \text{ and } \mathcal{H}_1 : \bar{X} \sim \mathcal{N}(\mu, 1/n) \text{ with } \mu \in \mathbb{R}. \tag{1}$$

Cherry suggests to compute the p-value under a different data-generating process than $\mathcal{H}_0$, namely, under $\mathcal{H}_{\widetilde{0}} : \bar{X} \sim \mathcal{N}(0, 1/n + g_0)$, where g_0 corresponds to the prior width of the peri-null prior $\mu \sim \mathcal{N}(0, g_0)$. From a frequentist point of view, $\mathcal{H}_{\widetilde{0}}$ is not satisfactory, as it implies that across repeated experiments, as $n \to \infty$ the sample mean becomes only a random draw from $\mathcal{N}(0, g_0)$. In other words, the frequentist interpretation of the peri-null distribution is unclear. It could not reflect the relative plausibility of the different non-zero values for the parameter (as this would be a Bayesian notion), and we do not see how it results from a repeated sampling argument.

Second, Cherry correctly notes that under $\mathcal{H}_{\widetilde{0}}$ with g_0 sufficiently small, that the peri-null Bayes factor $\text{BF}_{\widetilde{0}1} > 1$, where $\text{BF}_{\widetilde{0}1}$ is constructed from $\mu \sim \mathcal{N}(0, g_0)$ and $\mu \sim \mathcal{N}(0, g_1)$, $0 < g_0 < g_1$, under the peri-null and alternative hypothesis, respectively. To align the "frequentist" and Bayesian conclusions, Cherry recommends to lower the threshold $\widetilde{p} < \alpha$. A direct computation shows that under $\mathcal{H}_{\widetilde{0}}$ the peri-null Bayes factor $\text{BF}_{\widetilde{0}1}$ remains less than one, if $\bar{x} = (\frac{1}{n} + g_0)^{1/2} z_\alpha$ with $z_\alpha < \left(\frac{g_1 \log(g_1/g_0)}{g_1 - g_0}\right)^{1/2}$, where z_α represents the (two-tailed) quantile of a standard normal distribution, that is, $\mathbb{P}(Z \geq z_\alpha) = \alpha/2$. A true frequentist would be reluctant to change their α level based on the choice of both g_0 and g_1. If they would find this adjustment acceptable, then they might as well directly base their inference on the peri-null Bayes factor (even though this procedure is inconsistent, e.g., Ly and Wagenmakers 2023).

In our opinion the resolutions suggested by Cherry are irrelevant, impractical, and philosophically unsatisfactory to frequentists and Bayesians alike. We stand behind our conclusion that the original form of the Jeffreys–Lindley paradox cannot be attributed

to the use of a supposedly anomalous point-null hypothesis in the Bayesian analysis. For any fixed p-value, no matter how low, the data will asymptotically support the skeptic's peri-null hypothesis over the proponent's alternative hypothesis, under the trivial condition that the skeptic's hypothesis is more narrow than the proponent's hypothesis. This general result should inconvenience not only frequentists but also Bayesians who use the Jeffreys–Lindley paradox as an argument against Bayes factor hypothesis testing.

Acknowledgements

This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to EJW (016.Vici.170.083). Centrum Wiskunde and Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands.

Declarations

Conflict of interest 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.

References

Cherry, J. L. 2023. Comment on: History and nature of the Jeffreys–Lindley paradox. Archive for History of Exact Sciences. In press.

Jeffreys, H. 1938. 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.

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

Ly, A., and E.-J. Wagenmakers. 2023. Bayes factors for peri-null hypotheses. TEST. Retrieved from https://arxiv.org/abs/2102.07162. In press.

Wagenmakers, E.-J., and A. Ly. 2023. History and nature of the Jeffreys–Lindley paradox. *Archive for History of Exact Sciences* 77: 25–72.