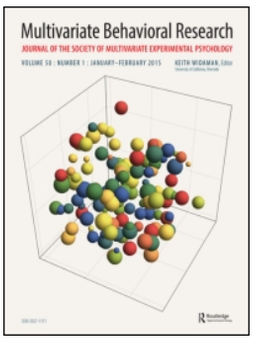

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Bayes Factors Have Frequency Properties—This Should Not Be Ignored: A Rejoinder to Morey, Wagenmakers, and Rouder

Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker

To cite this article: Herbert Hoijtink, Pascal van Kooten & Koenraad Hulsker (2016) Bayes Factors Have Frequency Properties—This Should Not Be Ignored: A Rejoinder to Morey, Wagenmakers, and Rouder, Multivariate Behavioral Research, 51:1, 20-22, DOI: 10.1080/00273171.2015.1071705

To link to this article: https://doi.org/10.1080/00273171.2015.1071705

Published online: 16 Feb 2016.

MULTIVARIATE BEHAVIORAL RESEARCH
2016, VOL. 51, NO. 1, 20–22
http://dx.doi.org/10.1080/00273171.2015.1071705

Routledge
Taylor & Francis Group

Bayes Factors Have Frequency Properties—This Should Not Be Ignored: A Rejoinder to Morey, Wagenmakers, and Rouder

Herbert Hoijtink[a], Pascal van Kooten[b], and Koenraad Hulsker[b]

[a]Department of Methods and Statistics, Utrecht University and CITO Institute for Educational Measurement; [b]Department of Methods and Statistics, Utrecht University

ABSTRACT
Hoijtink, van Kooten, and Hulsker (2016) outline a research agenda for Bayesian psychologists: evaluate and use the frequency properties of Bayes factors. Morey, Wagenmakers, and Rouder (2016) respond that Bayes factors calibrated using frequency properties should not be used. This paper contains the response of Hoijtink, van Kooten, and Hulsker to the criticism of Morey, Wagenmakers, and Rouder (2016).

KEYWORDS
Bayes factor; calibrated Bayes; default prior distribution; frequency calculations; subjective prior distribution.

Introduction

The experiments by Bem (2011) resulted in discussions concerning the statistical approaches used for data analysis. The hypotheses Bem evaluated in his first experiment were $H0 : \delta = 0$ versus $H1 : \delta \neq 0$, where positive values of the effect size δ indicate the existence of psi. Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) present an argument in favor of Bayes factors over p values to evaluate the hypotheses of interest. Specifically they use the Jeffreys, Zellner, and Siow (JZF) Bayes factor, which is closely related to the the scaled information Bayes factor as elaborated in Rouder, Speckman, Sun, Morey, and Iverson (2009). Hoijtink, van Kooten, and Hulsker (2016), subsequently denoted by HKH, determine the frequency properties of the scaled information Bayes factor and conclude (1) that the scale τ of the $N(0, \tau)$ prior distribution of δ under $H1$ influences these frequency properties and that research avenues that determine the scale using these frequency properties should be explored, (2) that the size of the Bayes factor can be interpreted via a translation into frequency-based conditional error probabilities, and (3) that subjective prior distributions may be a viable alternative for the default prior distributions used by Rouder et al. and Wagenmakers et al. Morey, Wagenmakers, and Rouder (2016), subsequently denoted by MWR, respond to HKH and argue that the prior distribution they use is subjective instead of default and that the calibrated Bayes factors proposed by HKH should not be used. In this response to MWR counterarguments will be given.

The prior used by MWR is more default than subjective

In this section the focus will be on the effect size δ that is central in the hypotheses $H0 : \delta = 0$ and $H1 : \delta \neq 0$. An important requirement for a subjective prior distribution for δ under H_1 is that it enables researchers to represent their prior knowledge. One option is a $N(\delta_0, \tau)$ prior with prior mean δ_0 and standard deviation τ.[1] More general distributions (for example, a skewed normal) are conceivable but at least the mean (representing a researcher's best guess) and the standard deviation (representing a researcher's certainty about his best guess) should be specifiable.

Bem (2011) states that "The main psi hypothesis was that participants would be able to identify the position of the hidden erotic picture significantly more often than chance" (p. 409), that is, δ is larger than zero. An elicitation procedure supporting the translation of this prior knowledge into a value for δ_0 and τ should clarify what the meanings are of "best guess" and "certainty about the best guess," respectively. In Bem's experiments a small positive effect size is expected; this could be translated into $\delta_0 = .1$. Bem seems to be rather certain about his expectation this could be translated into $\tau = .05$ resulting in a normal prior for δ that places about 95% of the weight on the area 0 to .2. Note that this is only one translation of Bem's hypothesis into a prior distribution. Others may very well propose different vales for δ_0 and τ. One way to deal with these "subjective" differences in opinion is by

CONTACT Herbert Hoijtink H.Hoijtink@uu.nl Department of Methods and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands.
[1]Note that τ is used to denote the prior standard deviation or scale—this in contrast to the use of τ in software like WinBugs and OpenBugs where it denotes the prior precision.

means of consensus priors as advocated by MWR towards the end of their paper.

Rouder et al. (2009) and Wagenmakers et al. (2011) use a $N(0, \tau)$ prior distribution for δ under $H1$ if the the scaled information Bayes factor is used (for the JZF Bayes factor a Cauchy prior centered around zero with scale τ is used). Due to a fixed prior mean of zero, this prior cannot represent Bem's main hypothesis, which states that δ is larger than zero. This is a *non*subjective, that is, default characteristic of this prior.

The prior standard deviation τ can be specified and can be used to represent a researcher's prior knowledge about the effect sizes that are expected. Often $\tau = 1$ has been used (see, for example, Wagenmakers et al., 2011). As noted by MWR for the Cauchy prior this implies that there is a 50% prior probability that $|\delta| > 1$. For the normal prior there is a 95% probability that the effect size is in the range –1.96 to +1.96, that is, there is a 32% prior probability that $|\delta| > 1$. Effect sizes larger than 1 are rarely observed in the behavioral sciences. Cohen (1992) in his "power primer" does not even consider effect sizes larger than .80. Especially in the Bem experiments where at best small effect sizes (say .1) are expected, a prior scale value of 1 is an unrealistic representation of the prior knowledge. It has to be concluded that the subjective nature of a $N(0, 1)$ prior distribution for δ under $H1$ is virtually nonexistent and that the nature of this prior is essentially default.

Evaluating this prior from a frequentist perspective by computing the probabilities of correctly preferring $H0/H1$ if the Bayes factor is larger/smaller than 1 highlights serious problems with the use of this prior. As can be seen in Figure 1 in HKH for a sample size $N = 100$, using $\tau = 1$ implies for smaller effect sizes like .20 (as are expected in Bem's experiments) that it is almost certain that $H0$ is preferred if it is true, but that there is only a probability of about 45% of correctly concluding that H1 is true. As can be seen in Figure 2 of HKH, for $N = 36$ the probability of correctly preferring $H1$ drops to about 25%. There is no denying—if the prior distribution specifies unrealistic ranges of effect sizes, the resulting Bayes factor is biased against $H1$ and has a rather small probability of correctly identifying $H1$.

Both figures also show that for each true effect size both the probabilities of correctly preferring $H0$ and $H1$ can be high. For example, if in the population the effect size is .20, for $N = 100$ choosing τ equal to .125 renders probabilities of about .78 of correctly preferring both $H0$ and $H1$. This led to the idea to choose τ such that the resulting Bayes factors have nice frequency properties. Lines along which such a default approach could be developed have been sketched in HKH. In the next section it will be argued that the criticism of MWR of such an approach is unjustified.

Calibrated Bayes factors

The first criticism of MWR is that "the calibration imposes an arbitrary quantification of the evidence on the data" (p. 13). This statement is unjustified. HKH complete three steps to obtain a clearly defined quantification of evidence:

(1) In the first step researchers have to specify which frequency properties they want their decision procedure to have, that is, what is required of the probabilities of correctly choosing $H0$ if the Bayes factor is larger than 1 and choosing $H1$ if the Bayes factor is smaller than 1. Many choices are conceivable. One choice is specified in Definition 1 from HKH: require both probabilities to be equal.
(2) In the second step the scale τ of the prior distribution has to be chosen such that the resulting Bayes factor has the desired frequency properties. MWR show that such calibrated Bayes factors are inconsistent and that it is not possible for all combinations of sample and effect sizes to find a τ value in agreement with Definition 1. Both their claims will be discussed below.
(3) In the third step frequentist conditional probabilities are computed, that is, what are the probabilities of making the correct decision given the size of the Bayes factor that is observed (see Figure 3 in HKH). This is a *non*arbitrary quantification of the evidence in the data that has a very clear interpretation.

The second criticism of MWR is that calibrated Bayes factors are inconsistent. MWR are correct that using a procedure straightforwardly based on Figures 1 and 2 from HKH renders inconsistent Bayes factors. However, Figures 1 and 2 only present the frequentist information upon which calibrated Bayes factors could be based. HKH sketch three options to use this frequentist information to choose the prior scale τ: the subjective option, the rational option, and data-based procedures. HKH write, "We want to provide a research agenda … not execute [it] and "providing less add hoc choices … should be added to the research agenda of Bayesian psychologists" (p. 6). Researchers having a more positive attitude than MWR with respect to calibrated Bayes factors would have asked themselves "how can consistent calibrated Bayes factors be obtained?" As is implied by Figure 4 from MWR a quick answer to this question is: limit the range of allowed values for τ to the interval $(0, 1]$. For smaller sample or effect sizes this would render a τ value still aiming at equal probabilities of correctly choosing $H0$ and $H1$. For larger sample or effect sizes the probabilities of correctly choosing $H0$ and $H1$ will be so high that it is no longer necessary to aim for equal probabilities. However, note that it is *not* claimed that this is *the* answer to obtain consistent

calibrated Bayes factors. More refined calibrated Bayes factors have to be developed. Furthermore, not only the consistency of the Bayes factor itself has to be considered, but the probabilities of correctly deciding that $H0/H1$ is true should go to 1 as the sample size increases.

The third criticism of MWR is that a τ value that will yield equal error probabilities can not be obtained for all combinations of effect sizes and sample sizes. As can be seen in Figure 2, for example, as τ goes to zero for $\delta = .20$ and $N = 36$ the probabilities of correctly deciding that $H0$ and $H1$ are true go to about .70 and .55, respectively, which is as equal as can be obtained for the chosen effect and sample size. The same situation occurs when a power analysis is executed for a t test for two independent means. If the interest is in $\delta = .2$, a Type I error of .05, and the sample size is $N = 36$ per group, the required power of .80 cannot be obtained (Cohen, 1992). In both situations (aiming for equal error probabilities and aiming for a power of .80) larger sample sizes are needed to obtain a procedure with the required frequency properties. The third criticism of MWR is not a criticism but a fact of life: "You can't always get what you want" (Jagger & Richards, 1969). MWR comment that in this situation "the only τ that meets HKH's calibration definition (Definition 1) is $\tau = 0$" (p. 15). However, this would render $H1$ identical to $H0$, which does not make sense and is nowhere suggested by HKH. It is clear that τ should always be strictly larger than zero.

Conclusion

In the last decade there has been a lot of attention on the evaluation of p values from a Bayesian perspective (see, for example, Wagenmakers, 2007). This has clarified properties of p values, increased attention on Bayes factors, and in general, increased understanding of what can be achieved using p values. HKH provide the start of an evaluation of Bayes factors from a frequentist perspective. This has clarified that the frequency properties of the scaled information Bayes factor depend on the prior scale τ and that "common" choices of τ result in questionable frequency properties. Stated otherwise, looking at Bayes factors through frequentist glasses may increase understanding of what can be achieved with Bayes factors and identify areas where matters have to be reconsidered and further research is needed.

Note that evaluation of Bayes factors from a frequentist perspective should not be limited to the scaled information Bayes factor as was done by HKH when outlining a research agenda for Bayesian psychologists. It may increase understanding and lead to changes in each situation (different statistical models, different sets of hypotheses) for which Bayes factors are developed and applied. And then maybe "You can't always get what you want, but if you try sometime, you just might find, you get what you need," (Jagger & Richards, 1969) both from a frequentist and a Bayesian perspective.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: No funding.

Acknowledgements The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi:10.1037/a0021524

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037//0033-2909.112.1.155

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they used the Bayes factor. *Multivariate Behavioral Research*, *51*, 1–9. doi:10.1080/00273171.2014.969364.

Morey, R. D., Wagenmakers, E. -J., & Rouder, J. N. (2016). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*, 10–17. doi:10.1080/00273171.2015.1052710.

Jagger, M., & Richards, K. (1969). You can't always get what you want. *Let It Bleed*. London, UK: ABKCO Music.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. doi:103758/PBR.16.2.225

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*, 779–804. doi:10.3758/bf03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790