Psychological Methods

ISSN: 1082-989X

2022, Vol. 27, No. 5, 804–821
https://doi.org/10.1037/met0000292

Prior Sensitivity of Null Hypothesis Bayesian Testing

Herbert Hoijtink
Department of Methodology and Statistics, Utrecht University

Abstract

Researchers increasingly use Bayes factor for hypotheses evaluation. There are two main applications: null hypothesis Bayesian testing (NHBT) and informative hypothesis Bayesian testing (IHBT). As will be shown in this article, NHBT is sensitive to the specification of the scale parameter of the prior distribution, while IHBT is not. As will also be shown in this article, for NHBT using four different Bayes factors, use of the recommended default values for the scaling parameters results in unpredictable operating characteristics, that is, the Bayes factor will usually be biased against or in favor of the null hypothesis. As will furthermore be shown in this article, this problem can be addressed by choosing the scaling parameter such that the Bayes factor is 19 in favor of the null hypothesis over the alternative hypothesis if the observed effect size is equal to zero, because this renders a Bayes factor with clearly specified operating characteristics. However, this does not solve all problems regarding NHBT. The discussion of this article contains elaborations with respect to: the multiverse of Bayes factors; the choice of "19"; Bayes factor calibration outside the context of the univariate normal linear model; and, reporting the results of NHBT.

Translational Abstract

Researchers increasingly use Bayes factor for hypotheses evaluation. However, Bayes factors do not actually evaluate hypotheses, they evaluate the prior distributions corresponding to these hypotheses. Loosely formulated, prior distributions represent for each hypotheses how each possible value of the parameters appearing in the hypothesis should be weighted. Whereas researchers using the Bayes factor while analyzing their data find it relatively easy to specify hypotheses, it is often not completely clear how the prior distributions should be specified. As will be shown in this article, in the context of the normal linear model, one solution to this problem is obtained if the Bayes factor of the null-hypothesis versus the alternative hypothesis is calibrated such that it is 19 if the observed effect size equals zero. If this calibration is used, the prior distributions corresponding to each hypothesis can uniquely be determined and do no longer have to be specified by the researchers using Bayes factor. With this article come R functions that can be used to apply the approach prosed in combination with the R package bain (https://informative-hypotheses.sites.uu.nl/software/bain/).

Keywords: Bayes factor, informative hypothesis Bayesian testing, null hypothesis Bayesian testing, prior sensitivity

The interest in Bayesian hypothesis evaluation is increasing. There are by now three R packages rendering the Bayes factor: for the evaluation of a null versus the alternative hypothesis `BayesFactor`[1]; and, with the additional option to evaluate informative hypotheses, `bain`[2] and `BFpack`.[3] Parts of the first two packages are also implemented in `JASP`[4] which is an easy to use and versatile statistical package that does not require knowledge of `R`.

Introducing the Bayes Factor

Royal (1997) provides an elaborate discussion of the likelihood ratio (test). Assuming that $y_i \sim \mathcal{N}(\mu, 1)$ for $i = 1, \ldots, N$, a simple instance is

$$LR_{01} = \frac{f(y \mid \mu = \mu_0)}{f(y \mid \mu = \mu_1)}, \tag{1}$$

that is, the likelihood ratio of $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu = \mu_1$, where $f(y \mid \mu = \mu_0) = \mathcal{N}(\bar{y} \mid \mu_0, \frac{1}{N})$ denotes the density of the data

This article was published Online First September 27, 2021.
Herbert Hoijtink https://orcid.org/0000-0001-8509-1973
Herbert Hoijtink is supported by a fellowship from the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS-KNAW) and the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Grant 024.001.003). Shortly before submitting this article, it was posted at the bain website.
Correspondence concerning this article should be addressed to Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, the Netherlands. Email: H.Hoijtink@uu.nl

[1] https://richarddmorey.github.io/BayesFactor/
[2] https://informative-hypotheses.sites.uu.nl/software/bain/
[3] https://cran.r-project.org/web/packages/BFpack/index.html
[4] https://jasp-stats.org/

given the hypothesized value μ_0 with $\bar{y}$ denoting the average of y, that is, the maximum likelihood estimate of y. In this simple example, the likelihood ratio can be interpreted as the relative support in the data for H_0 and H_1, that is, the support for H_0 is LR_{01} times larger than the support for H_1. If $LR_{01} = 5$, the support is five times larger for H_0, if $LR_{01} = .2$, the support is five times larger for H_1, and if $LR_{01} = 1$, H_0 and H_1 are equally supported.

The likelihood ratio cannot always be interpreted in terms of relative support. Consider, for example, H_0: $\mu = 0$ and H_1: $\mu \neq 0$, for which

$$LR_{01} = \frac{f(y \mid \mu = 0)}{f(y \mid \mu = \bar{y})} = \exp\left(-\frac{1}{2}N\bar{y}^2\right). \tag{2}$$

Because the exponent of a negative number is smaller than 1, this likelihood ratio will always express support in favor of H_1 (except when $\bar{y} = 0$ which results in $LR_{01} = 1$). This is due to the fact that H_1 can adapt to the data whereas H_0 cannot and therefore $f(y \mid \mu = \bar{y})$ will always be larger than $f(y \mid \mu = 0)$. The classical manner to deal with this issue is to use $-2\log LR_{01}$ as a test statistic that is evaluated using a chi-square distribution with one degree of freedom to obtain a p-value. However, the interpretation of the likelihood ratio as a measure of relative support remains lost.

The Bayes factor (Hoijtink et al., 2019; Kass & Raftery, 1995) can be seen as a generalization of the likelihood ratio that always retains the interpretation as a relative measure of support. The Bayes factor is the ratio of two marginal likelihoods. Application to H_0: $\mu = 0$ and H_1: $\mu \neq 0$ renders:

$$BF_{01} = \frac{\int_\mu f(y \mid \mu) h_0(\mu) d\mu}{\int_\mu f(y \mid \mu) h_1(\mu) d\mu} = \frac{f(y \mid \mu = 0)}{\int_\mu f(y \mid \mu) h_1(\mu) d\mu}. \tag{3}$$

The density $h_0(\mu)$ denotes the so-called prior distribution of μ under H_0. Because H_0 allows only $\mu = 0$, the integral over this density reduces to $f(y \mid \mu = 0)$. The density $h_1(\mu)$ denotes the prior distribution of μ under H_1, that is, which values of μ with which weights are deemed reasonable under H_1. For the setup at hand a $\mathcal{N}(\mu \mid 0, \tau^2)$ would be a common choice. For the illustration at hand $\tau^2 = 1$ will be used. The Savage-Dickey approach (see, e.g., Wagenmakers et al., 2010) can be used to rewrite Equation 3 into the ratio of the posterior and prior densities under H_1 evaluated in $\mu = 0$:

$$BF_{01} = \frac{\mathcal{N}(\mu = 0 \mid \mu_N, \sigma_N^2)}{\mathcal{N}(\mu = 0 \mid 0, 1)} = \frac{1}{\sqrt{\sigma_N^2}} \exp\left(-\frac{1}{2}\frac{\mu_N^2}{\sigma_N^2}\right), \tag{4}$$

where

$$\mu_N = \frac{N\bar{y}}{1 + N} \tag{5}$$

and

$$\sigma_N^2 = \frac{1}{1 + N}. \tag{6}$$

If $N = 24$ and $\bar{y} = 0$ the support in the data should be larger for H_0 than for H_1. In this case, $BF_{01} = 5 \ \exp\ 0 = 5$, that is, the support for H_0 is five times larger than the support for H_1. If $N = 24$ and $\bar{y} = .5$ the support in the data should be larger for H_1 than for H_0. In this case, $BF_{01} = 5 \exp -\frac{1}{2}\frac{.48^2}{1/25} = 5\exp(-2.88) = .28$, that is, the support for H_1 is about 3.57 times larger than the support for H_0.

As was shown in the previous paragraph, the Bayes factor can express support for and against H_0 and H_1, and can therefore be used as a measure of relative support. If, for example, in null hypothesis Bayesian testing (NHBT, a term introduced by Tendeiro & Kiers, 2019) the Bayes factor for the comparison of H_0: $\mu_1 = \mu_2 = \mu_3$ and H_1: not H_0 equals $BF_{01} = 5$, then the support in the data for the null-hypothesis that the three means are equal is five times stronger than for the alternative hypotheses that the three means are not equal. Alternatively, in informative hypothesis Bayesian testing (Gu et al., 2014; Klugkist, et al., 2005; Mulder, 2014) the informative hypothesis H_i: $\mu_1 > \mu_2 > \mu_3$ can be compared with its complement H_c: not H_i. If $BF_{ic} = .1$, then the support in the data is ten times stronger for the complement of H_i.

An Issue With NHBT: Specification of the Scale of the Prior Distribution

It will now, first of all, be elaborated why the interpretation of the Bayes factor given in the previous paragraph is imprecise. However, subsequently, arguments in favor of the imprecise interpretation will be given. In fact, the Bayes factor quantifies the relative support in the data for two prior distributions. To give another simple example, what is the support for the prior distributions $h(\delta) = 0$ versus $h(\delta) \sim \mathcal{N}(0, .0625)$ (where δ denotes Cohen's d for one mean; Cohen, 1992; and the normal distribution is specified using a mean and a variance), that is, what is the relative support for a hypothesis that specifies that $\delta = 0$ versus an alternative hypothesis that specifies that δ has about a 95% probability of being located in the interval $-.50$ to .50. This specification of hypotheses can be called subjective, because the alternative hypothesis can only be specified if it is meaningful for the researcher at hand: "if δ is nonzero, I do not expect it to be larger than .50." However, the subjective specification of a prior distribution corresponding to the alternative hypothesis is not without difficulties. To name but a few: Should the prior variance be chosen such that the expected effect sizes are covered with a 99%, 95%, or 90% probability? Furthermore, is a user at all able to specify what the expected effect sizes are? And, a fortiori, how to specify subjective prior distributions and answer these questions if multiparameter hypotheses are of interest, and if these parameters are embedded in encompassing models like, for example, the regression coefficients and factor loadings in structural equation models. These questions will not be answered in this article, however, further research in this area would be welcome and valuable.

Most users of the Bayes factor are only concerned with the evaluation of hypotheses and not at all with the subjective specification of the corresponding prior distributions. The latter can be avoided if the user only has to specify the hypotheses of interest "I want to compare H_0 with H_1" and interprets the Bayes factor as the relative support in the data for both hypotheses. However, in order to be able to compute the Bayes factor, this requires the automatic translation of the hypotheses of interest into *completely specified* prior distributions.

The theoretical contribution of this article is to show that in the context of the normal linear model it is possible to obtain completely specified prior distributions and Bayes factors with clear operating characteristics if it is required that BF_{01} = 19 if the observed effect size (e.g., Cohen's d or the proportion of variance explained) equals zero. To increase the accessibility of the article, most of the derivations and statistical elaborations necessary to show this have been placed in Appendixes. However, this article also renders a practical contribution, because R functions and examples are provided that enable researchers to apply the approach proposed when they use `bain` to Bayesianly evaluate hypotheses in the context of the normal linear model. This function and corresponding examples will be implemented in the next version of `bain` but can also be downloaded with this article from the `bain` website[5] under new publications.

In the next section, in the context of a simple statistical model, four Bayes factors will be introduced. As will be shown, all are based on prior distributions with an unknown scaling parameter. Subsequently, the sensitivity of NHBT to the choice of the scaling parameter will be addressed. In the two sections that follow, a choice for the scaling parameter will be proposed and elaborated for the Bayes factor implemented in the R package `bain`. Thereafter, the concepts introduced in this article will, be applied to NHBT concerning multiple regression, ANOVA, ANCOVA, and the Welch test. The article is concluded with a discussion in which, among other things, it will be highlighted that IHBT is not sensitive to the choice of the scaling parameter.

Four Bayes Factors

The elaborations in this and the next three sections are in the context of the following simple statistical model:

$$y_i \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, \ldots, N, \tag{7}$$

where y_i denotes the datum of the ith person, N the sample size, and μ and σ^2 the mean and variance, respectively, of a normal distribution for the data. Note that, in the sequel, $\delta = \mu/\sigma$ will denote Cohen's d in the population of interest and d the value in the sample. In the remainder of this section subsequently two Bayes factors that quantify the relative evidence for H_0: $\mu = 0$ versus H_1: $\mu \neq 0$ and two Bayes factors that quantify the relative evidence for H_0: $\delta = 0$ versus H_1: $\delta \neq 0$ will be introduced.

The first is the AAFBF, that is, the approximate (Gu et al., 2018; Hoijtink et al., 2019) adjusted (Mulder, 2014) fractional Bayes factor (O'Hagan, 1995). The fractional prior distribution used for testing H_0 versus H_1 is

$$\mathcal{N}\left(\mu \mid 0, \frac{1}{b}\frac{s^2}{N}\right), \tag{8}$$

where s^2 denotes the unbiased sample variance of y. Note that, the prior distribution has an *adjusted* prior mean of 0 and a variance based on a *fraction* $b = J/N$ of the information in the posterior distribution. For the simple model at hand, the default choice $J = 1$ is implemented in the R package `bain`. Based on Equations 7 and 8 the Bayes factor for testing H_0 versus H_1 can be represented using the Savage-Dickey approach:

$$BF_{\mathcal{N}} = \frac{\mathcal{N}(\mu = 0 \mid \bar{y}, \frac{s^2}{N})}{\mathcal{N}(\mu = 0 \mid 0, \frac{1}{b}\frac{s^2}{N})}, \tag{9}$$

where $\bar{y}$ denotes the sample mean of y. It is the ratio of a normal *approximation* of the posterior distribution of μ (the nominator of Equation 9) evaluated in $\mu = 0$ and a corresponding normal prior distribution (the denominator of Equation 9) evaluated in $\mu = 0$. Further elaborations of the AAFBF can be found later in this article and in Appendix A.

The second is the AFBF, that is, the adjusted (Mulder, 2014; Mulder et al., 2019) fractional Bayes factor (O'Hagan, 1995). The fractional prior distribution used for testing H_0 versus H_1 is

$$\mathcal{T}_{Nb-1}\left(\mu \mid 0, \frac{(N-1)s^2}{N(Nb-1)}\right). \tag{10}$$

Like for the AAFBF $b = J/N$ and the default choice $J = 2$ is implemented in the R package `BFpack`. Note that, this choice renders a $\mathcal{T}$ distribution with one degree of freedom, that is, a scaled Cauchy distribution. Because for $N \to \infty$ this posterior distribution converges to a normal distribution, the main difference between the AAFBF and the AFBF is the use of a normal and $\mathcal{T}$ prior distribution, respectively. Based on Equations 7 and 8 the Bayes factor for testing H_0 versus H_1 can be represented using the Savage-Dickey approach:

$$BF_{\mathcal{T}} = \frac{\mathcal{T}_{N-1}(\mu = 0 \mid \bar{y}, \frac{s^2}{N})}{\mathcal{T}_{Nb-1}\left(\mu = 0 \mid 0, \frac{(N-1)s^2}{N(Nb-1)}\right)}, \tag{11}$$

that is, the posterior (nominator) and a corresponding prior (denominator) densities of μ evaluated in $\mu = 0$.

The third is the scaled information Bayes factor BF_{SI} (there is an online calculator[6] for this Bayes factor) as presented in Rouder et al. (2009). To derive this Bayes factor, the density of the data is rewritten as:

$$y_i \sim \mathcal{N}(\sigma \times \delta, \sigma^2) \text{ for } i = 1, \ldots, N, \tag{12}$$

and combined with a normal prior distribution for δ and Jeffrey's prior for σ^2:

$$h_{SI}(\delta, \sigma^2) = \mathcal{N}(\delta \mid 0, r^2) \times 1/\sigma^2, \tag{13}$$

with the default r = .707. The equation of BF_{SI} can be derived using Equations 12 and 13, the interested reader is referred to Rouder et al. (2009) for the details.

The fourth is the Jeffreys, Zellner, Siow Bayes factor BF_{JZS} as presented in Rouder et al. (2009) which uses the same density of the data as BF_{SI} and a $\mathcal{T}$ prior distribution with 1 degree of freedom, that is, a scaled Cauchy distribution with location parameter 0 and scale factor r:

$$h_{JZS}(\delta, \sigma^2) = \mathcal{T}_1(\delta \mid 0, r^2) \times 1/\sigma^2. \tag{14}$$

[5] https://informative-hypotheses.sites.uu.nl/software/bain/

[6] http://pcl.missouri.edu/bf-one-sample

The default value implemented in the R package BayesFactor is r = .707. The equation of BF_{JZS} can be derived using Equations 12 and 14, the interested reader is referred to Rouder et al. (2009) for the details. The main difference between BF_{SI} and BF_{JZS} is the use of normal and $\mathcal{T}$ prior distributions, respectively.

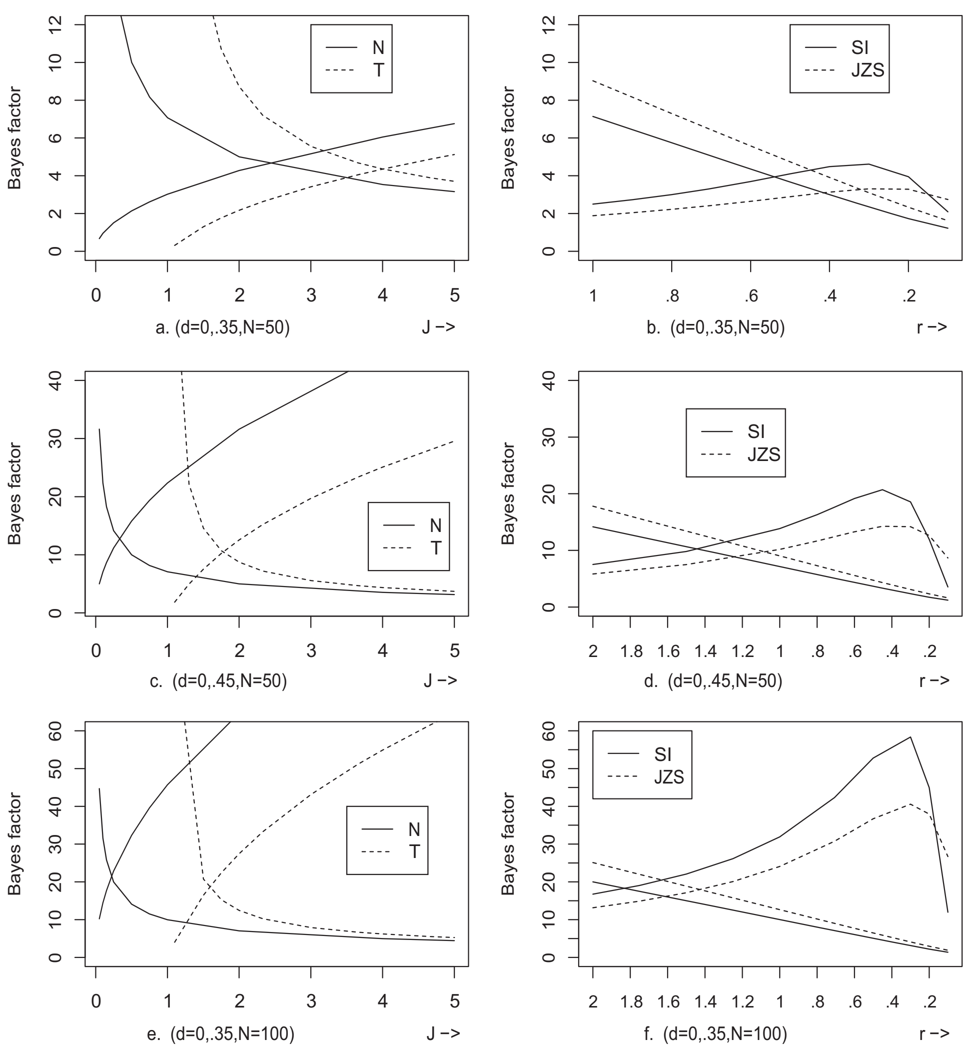
Sensitivity of NHBT to *J* and *r*

In this section the sensitivity of the four Bayes factors to the specification of J and r will be illustrated using their operating characteristics. This sensitivity has previously been discussed in terms of frequentist decision errors by Gu et al. (2016) and Gu et al. (2018). Their proposal to deal with this sensitivity involves integration over a distribution for the effect size (not the prior distribution of the model parameters) that has to be specified by the user. Their approach replaces the sensitivity issue from specification of the prior distribution to specification of the distribution for the effect size and is therefore not a solution of the sensitivity issue. In this article sensitivity will be discussed in terms of operating characteristics in the form of the relative support BF_{01} for H_0 versus H_1 as a function of $d = \bar{y}/s$, N, and J or r. Note that, Hoijtink et al. (2016) used this kind of operating characteristics to discuss the sensitivity of BF_{SI} for the simple model at hand.

Figure 1 displays the operating characteristics of the four Bayes factors introduced in the previous section. In Figure 1a for a data set with N = 50, s^2 = 1 and consequently $d = \bar{y}$ equal to 0 or .35, the dependence of the two fractional Bayes factors on J (which is the main determinant of the prior variance through $b = J/N$, see Equations 9 and 11) is displayed. In Figure 1b the dependence of the SI and JZS Bayes factors on r is displayed. Figures 1c and 1d present the same information, however, here d is equal to 0 and .45. Figures 1e and 1f also present the same information, however, here N is equal to 100 and d is equal to 0 and .30. Note that, when

Figure 1

Four Bayes Factors as a Function J and r, Respectively



Note. In the top two figures N = 50 and d = 0, .35. In the middle two figures N = 50 and d = 0, .45. In the bottom two figures, N = 100 and d = 0, .35. The lines going down correspond to Bayes factors BF_{01} in favor of H_0 when d = 0, and the lines going (mostly) up correspond to Bayes factors BF_{10} in favor of H_1 with d = .35 or d = .45.

$d = 0$ the Bayes factor BF_{01} of H_0 versus H_1 is displayed (thus, the support in favor of H_0), and when $d \neq 0$ the Bayes factor BF_{10} of H_1 versus H_0 is displayed.

Observation 1

As can be seen in each of the six subfigures in Figure 1, the operating characteristics are sensitive to the choice of J and r, that is, if BF_{01} is relatively large for $d = 0$ then BF_{10} is relatively small for $d = .35$, .45, and vice versa. Stated otherwise, J and r can be chosen such that the Bayes factor is biased in favor of H_0 or H_1. The interested reader is referred to Tendeiro and Kiers (2019) where this feature of Bayes factors is also discussed.

Observation 2a

There is no justification for the default values of J and r in terms of the operating characteristics of the corresponding Bayes factors. As can be seen in Figures 1a and 1b, if the defaults values of J and r are used, the Bayes factors are biased in favor of H_0: With $J = 2$, $BF_{\mathcal{T}}$ equals 8.72 and 2.17, when H_0 and H_1 are true, respectively; with $J = 1$, the corresponding numbers for $BF_{\mathcal{N}}$ are 7.07 and 3.02; and, with $r = .707$, BF_{JZS} renders 6.49 and 2.40 and BF_{SI} 5.09 and 3.30. In other words, if $N = 50$ and $d = .35$ it is easier to find support for H_0 than for H_1. However, as can be seen in Figures 1c and 1d, the bias reverses if d equal to 0 and .45 is used, and, as can be seen comparing Figures 1a and 1b, to Figures 1e and 1f, the degree and direction of the bias is also influenced by N.

Observation 2b

When evaluated using the default values of J and r, the Bayes factors in favor of H_0 range from 5.09 to 8.72 in the top two figures, also from 5.09 to 8.72 in the middle two figures, and from 5.99 to 10.37 in the bottom two figures. The Bayes factors in favor of H_1 range from 2.17 to 3.30 in the top two figures, from 12.54 to 22.34 in the middle two figures, and from about 27 to 42 in the bottom two figures. As can be seen, there can be relevant variation in the size of the four Bayes factors if the default values of J and r are used.

Observation 3a

Given N, J and r can be chosen such that equal support for $d = 0$ and $d \neq 0$ is obtained. In Figures 1a and 1b it can be seen that equal support for H_0 and H_1 with $d = .35$ is obtained using $J = 4$ for $BF_{\mathcal{T}}$ and $r = .32$ for BF_{JZS}. As can be seen comparing Figures 1a and 1b with Figures 1c and 1d, respectively, the values of J and r for which equal support is obtained depend on the value of d under H_1. As can be seen comparing Figures 1a and 1b to Figures 1e and 1f, the values of J and r for which equal support is obtained also depend on N.

The values of N and d for which the equilibrium is determined will from now on be called N_{ref} and d_{ref}, the value of J or r for which the equilibrium is obtained will be called J_{ref} and r_{ref}, and the value of the Bayes factor in the equilibrium will be called BF_{ref}. Note that, BF_{ref} is obtained both for $d = 0$ and $d = d_{ref}$.

Observation 3b

When evaluated using J_{ref} and r_{ref}, the four BF_{ref}'s range from 3.27 to 4.62 in the top two figures, from 9.37 to 12.57 in the middle two figures, and from about 17 to about 21 in the bottom two figures. This variation is smaller than the variation observed when the default values for J and r are used (compare Observation 2b).

Observation 4

The operating characteristics of the Bayes factor are sensitive to choice of normal or $\mathcal{T}$ prior distributions and the choice of fractional or nonfractional prior distributions. As can be seen in Figure 1 in each of the top, middle, and bottom two figures, in the equilibrium $BF_{\mathcal{N}}$ is slightly larger than $BF_{\mathcal{T}}$ and BF_{SI} is slightly larger than BF_{JZS} (normal priors render more support than $\mathcal{T}$ priors), and that both fractional Bayes factors are larger than the scaled information and JZS Bayes factors.

In the three sections that follow, first of all, Observations 1 through 4 will be used to address Question 1. How to choose between normal and $\mathcal{T}$ and between fractional and nonfractional prior distributions? Subsequently, further elaborations of the AAFBF are presented, followed by a discussion of Question 2. How can N_{ref}, d_{ref}, and J_{ref} be determined, and is there a role for a sensitivity analysis with respect to the choice of J_{ref}?

An Answer to Question 1: How to Choose Between Normal and $\mathcal{T}$ and Between Fractional and Nonfractional Prior Distributions?

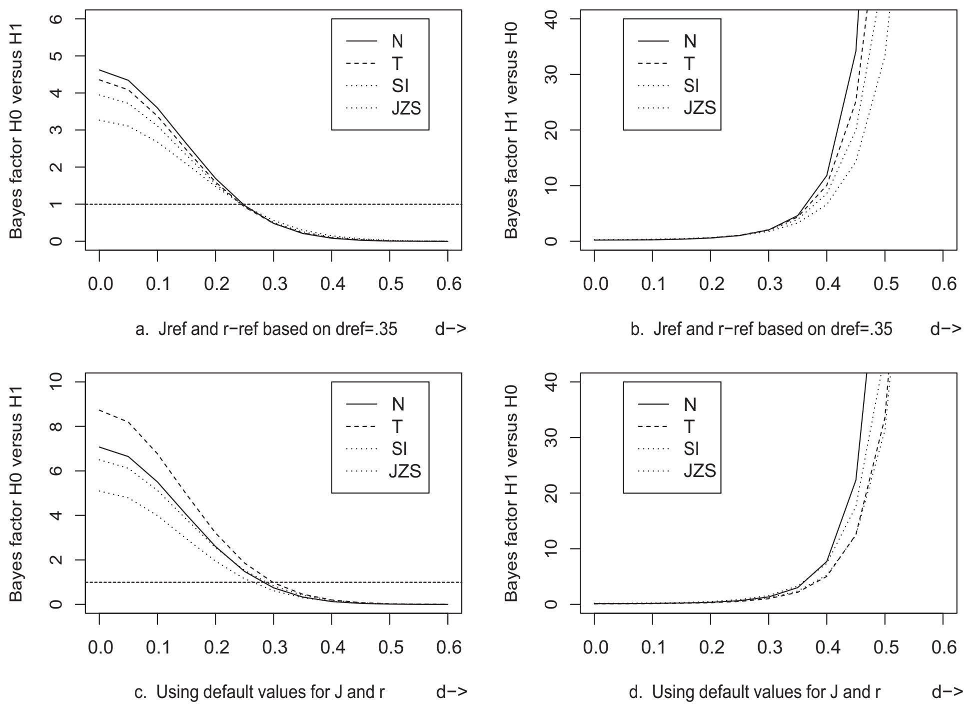
As was presented in Observation 4, with respect to their size in the equilibrium, the four Bayes factors have the same ordering in each of the three conditions presented in Figure 1. It is important to stress that these differences in size are *not* the result of *more or less support in the data* for the hypotheses entertained, it is merely the result of the specifics of the prior distribution used. These differences can therefore not be used to argue in favor of one of the Bayes factors.

However, it can be argued that the current default values of J and r should be replaced by J_{ref} and r_{ref}, because then the operating characteristics of the four Bayes factors become more similar (compare Observations 2b and 3b). This is further illustrated in Figure 2 for $N_{ref} = 50$ and $d_{ref} = .35$: If the observed effect size d is between 0 and (about) .35, the differences between the four Bayes factors computed using J_{ref} and r_{ref} (compare Figure 2a) are smaller than the difference between the four Bayes factors computed using the default values of J and r (compare Figure 2c). For d larger than about .35, the differences between the four Bayes factors may become substantial, irrespective of whether reference or default values for J and r are used (compare Figures 2b and 2d). Furthermore, as can be seen in Figure 2a, the four Bayes factors equal one at about the same effect size d when based on J_{ref} and r_{ref}, while, as can be seen in Figure 2c, the effect size at which each Bayes factor equals one is clearly different if the default values of J and r are used.

In summary, at least for the simple model at hand, the performance of the four Bayes factors becomes more similar when based on J_{ref} and r_{ref}. However, there remain differences, and these differences cannot be used to argue in favor or against one of the four Bayes factors because they reflect differences in the prior distributions used and not differences in the support in the data for the null and alternative hypothesis. Therefore the best answer to Question 1 that is currently available is: Use J_{ref} or r_{ref} and report in research reports explicitly which Bayes factor is used. This issue will be further elaborated in the Discussion section.

Figure 2

The Four Bayes Factors Computed Using N_{ref} = 50 and d_{ref} =.35 as a Function of d



Note. In the top two panels the Bayes factors based on J_{ref} and r_{ref} are displayed. In the bottom two panels the Bayes factors based on the default values of J and r are displayed. In the left hand panels the Bayes factor BF_{01} of H_0 versus H_1 is displayed, in the right hand panels the Bayes factor BF_{10} of H_1 versus H_0. JZS = Jeffreys, Zellner, Siow Bayes factor; SI = scaled information Bayes factor.

The remaining question is how to determine J_{ref} and r_{ref}. This article will only consider J_{ref} for $BF_{\mathcal{N}}$ which will from now on be denoted by BF_{01}. The first reason is that it is the most versatile of the Bayes factors under consideration. It can be used for the evaluation of null, alternative, and informative hypotheses in a wide range of statistical models. The interested reader is referred to the vignette included with the `R` package `bain` for an illustration of many options. The second reason is that the author of this article is one of the authors of this R package and its statistical underpinnings. It would be valuable to obtain and implement derivations analogous to the ones that will be made for $BF_{\mathcal{N}}$ (and other instances of the AAFBF) in the `R` packages and the online calculator that render the other three Bayes factors (also beyond the context of the simple model at hand). However, this is up to the authors of these packages and the corresponding articles.

A Further Elaboration of the AAFBF

The AAFBF was initially derived for IHBT (Gu et al., 2014). Subsequently, it was generalized to also include NHBT (Gu et al., 2018; Hoijtink et al., 2019). In this section, using the simplest form of the AAFBF, that is, $BF_{\mathcal{N}}$, the AAFBF will be further elaborated. Note that, the elaborations that follow generalize straightforwardly from $BF_{\mathcal{N}}$ to the AAFBF. The interested reader is referred to Appendix A for elaborations with respect to the consistency of the AAFBF and updating using the AAFBF.

The simplest form of the AAFBF is:

$$BF_{\mathcal{N}} = \frac{\mathcal{N}(\mu = 0 \mid \bar{y}, \frac{s^2}{N})}{\mathcal{N}(\mu = 0 \mid 0, \frac{s^2}{J})}. \tag{15}$$

The normal approximation of the posterior distribution of μ can be written as

$$\begin{aligned}\mathcal{N}\left(\mu \mid \bar{y}, \frac{s^2}{N}\right) &= \mathcal{N}\left(\mu \mid \bar{y}, \frac{s^2}{N}\right)^{1-b} \times \mathcal{N}\left(\mu \mid \bar{y}, \frac{s^2}{N}\right)^{b} \times C \\ &\approx \mathcal{N}\left(\mu \mid \bar{y}, \frac{s^2}{N}\right)^{1-b} \times \mathcal{N}\left(\mu \mid 0, \frac{1}{b}\frac{s^2}{N}\right),\end{aligned} \tag{16}$$

that is, the information in the posterior distribution is separated into a fraction $1 - b$ representing the information in the density of the data and a fraction b multiplied with a constant prior C which is the fractional prior distribution $h_{\mathcal{N}}(\mu)$ (Gilks, 1995). *Fractional* refers to the fact that the scale of the prior distribution is based on a fraction $b = J/N$ of the information in the density of the data. For the simple model at hand, the default choice $J = 1$ is implemented in the `R` package `bain`.

Going from the first to the second line of Equation 16 the prior mean is *adjusted* to 0, which explains the about equality separating both parts of the equation. This adjustment (which is not only used for $BF_{\mathcal{N}}$ but analogously for the AAFBF in general) is necessary because the unadjusted fractional Bayes factor is inconsistent when the goal is to (possibly jointly with null-hypotheses) evaluate informative hypotheses (Mulder, 2014). If informative hypotheses are specified without using about equality constraints like, for example, $-.2 < \mu < .2$, IHBT using AAFBF does not depend on J (Mulder, 2014). Therefore, in the generalization of Equation 15 that applies to the AAFBF, $J = 0$ can be used, the about equality in the generalization of Equation 16 can be replaced by an exact equality, and the Bayes factor based on the resulting posterior and corresponding fractional prior distribution is an approximate (because of the normal approximation of the posterior distribution) fractional Bayes factor.

However, the outcomes of NHBT using the AAFBF do depend on J, therefore it is an approximate *adjusted* Bayes factor, or, in other words, a consistent (as will be shown in the next subsection) information criterion inspired by the Bayes factor. Returning to Equations 15 and 16, if b is small, that is, J is relatively small compared with N and/or if the observed effect size $d = \bar{y}/s$ is close to zero, the contribution of the prior to the posterior is negligible. Therefore, the approximation in Equation 16 is accurate and the corresponding Bayes factor is an approximate fractional Bayes factor. However, if J is large compared with N and the observed effect size is not close to zero, the about equality in Equation 16 will only hold asymptotically and $BF_{\mathcal{N}}$ (and also the AAFBF in general) is an approximate *adjusted* Bayes factor that is, an information criterion inspired by the Bayes factor. The consequence is that $BF_{\mathcal{N}}$ is only asymptotically the ratio of two marginal likelihoods (Kass & Raftery, 1995) and nonasymptotically the ratio of the posterior density evaluated in $\mu = 0$ (which can be called the fit of an hypothesis, because the larger this density the more the posterior distribution is centered around 0) and the prior density evaluated $\mu = 0$ (which can be called complexity because the smaller this density the larger the precision of H_0 relative to H_1, that is, smaller densities result from prior distributions with larger variances). For a further elaboration of fit and complexity the interested reader is referred to Hoijtink et al. (2019). Another consequence is that posterior model probabilities computed from these Bayes factors become posterior model weights. However, for all practical purposes, the interpretation of the AAFBF will remain the same.

An Answer to Question 2: How Can N_{ref}, d_{ref}, and J_{ref} Be Determined, and Is There a Role for a Sensitivity Analysis With Respect to the Choice of J_{ref}?

The goal of this article is to evaluate H_0 and H_1 (and not the corresponding prior distributions) using the Bayes factor. This implies that the prior distribution has to be constructed such that it is fully specified. As will be shown in this section, this can be achieved by choosing a reference sample size N_{ref} and a reference Bayes factor BF_{ref}. Knowing these two quantities is sufficient to be able to summarize the operating characteristics in terms of d_1, the effect size for which $BF_{01} = 1$, and d_{ref}, the effect size for which $BF_{01} = BF_{ref}$. Both quantities are also sufficient to compute J_{ref}. In the next subsection, first J_{ref} will be derived. Subsequently, it will be shown that knowledge of N_{ref} and one of d_1, d_{ref} and BF_{ref} is sufficient to determine the other two and J_{ref}. In the subsection that follows, based on these quantities, an approach to deal with prior sensitivity will be proposed. This section is concluded with two examples.

Derivations

Using $b = J/N$ and $d = \bar{y}/s$, Equation 9 can be rewritten as

$$BF_{01} = \frac{1/\sqrt{2\pi s^2/N}}{1/\sqrt{2\pi s^2/J}} \exp\left(-\frac{1}{2}\frac{\bar{y}^2}{s^2/N}\right)$$

$$= \frac{\sqrt{N}}{\sqrt{J}} \exp\left(-\frac{N}{2}d^2\right). \tag{17}$$

With $N_{ref} = N$, BF_{01} computed for a data set in which $d = 0$ equals BF_{10} computed for a data set in which $d = d_{ref}$ if (compare Equation 17)

$$\frac{\sqrt{N_{ref}}}{\sqrt{J_{ref}}} = \frac{\sqrt{J_{ref}}}{\sqrt{N_{ref}}} \exp\left(\frac{N_{ref}}{2}d_{ref}^2\right). \tag{18}$$

Solving for J_{ref} renders

$$J_{ref} = N_{ref} \exp\left(-\frac{N_{ref}}{2}d_{ref}^2\right), \tag{19}$$

with the value of the Bayes factor in the equilibrium equal to

$$BF_{ref} = \frac{\sqrt{N_{ref}}}{\sqrt{J_{ref}}} = \exp\left(\frac{N_{ref}}{4}d_{ref}^2\right). \tag{20}$$

As can be seen in Equation 20, N_{ref} and BF_{ref} are sufficient to compute d_{ref} and, as can be seen in Equation 19, N_{ref} and d_{ref} are sufficient to compute J_{ref}. If $N_{ref} = N$, that is, the reference sample size is set equal to the realized sample size, Equation 17 simplifies to

$$BF_{01} = \frac{\sqrt{N}}{\sqrt{J_{ref}}} \exp\left(-\frac{N}{2}d^2\right) = \exp\left(\frac{N}{4}(d_{ref}^2 - 2d^2)\right). \tag{21}$$

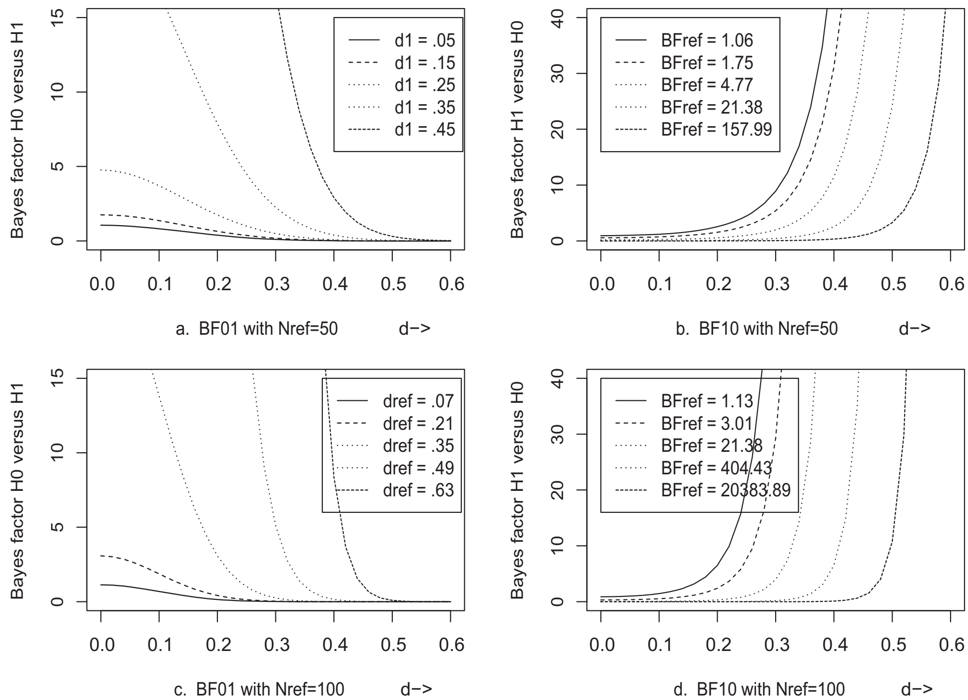
From this equation it can be seen that there is a $d = d_1$ such that $d_{ref}^2 = 2d_1^2$ which results in $BF_{01} = 1$. In fact, d_1 is the effect size at which the size of BF_{01} changes from "larger than 1," that is, expressing support in favor of H_0 to "smaller than 1," that is, expressing support in favor of H_1.

Dealing With Prior Sensitivity

As is already clear from the previous section, in this article N_{ref} is chosen to be equal to the observed sample size N. In Figure 3a and 3b for $N_{ref} = 50$ the operating characteristics of BF_{01} and BF_{10} are displayed. Note that, each line has three labels: BF_{ref}, d_1, and d_{ref}. Using $d_1 = .05$ to specify J_{ref} implies that a researcher is interested in finding support for H_1 if d is larger than .05. However, with $N_{ref} = 50$ this results in bad operating characteristics, that is, BF_{ref} is "only" 1.06 for both $d = 0$ and $d = d_{ref} = .07$. If a researcher chooses $d_1 = .25$ to specify J_{ref}, this results in $BF_{ref} = 4.77$ for both $d = 0$ and $d = d_{ref} = .35$. This is much better in terms of operating characteristics and highlights a limitation of $N_{ref} = 50$, that is, only d larger than .25 will render support in favor of H_1. Increasing the sample size will of course improve the operating characteristics. This can be seen comparing Figures 3a and 3b ($N_{ref} = 50$) with 3c and 3d ($N_{ref} = 100$), respectively.

Figure 3

BF_{01} and BF_{10} Calibrated Using J_{ref} as a Function of N_{ref} = 50 (Figures 3a and 3b) and N_{ref} = 100 (Figures 3c and 3d)



Note. d_{ref} = .07, .21, .35, .49, .63 corresponds to d_1 = .05, .15, .25, .35, .45. The legends in the left hand figures apply to all figures, that is, each line is labeled by both a d_{ref} and a d_1. Additionally, each line is labeled by a $\mathrm{BF}_{\mathrm{ref}}$ which can for both the top and bottom row of figures be found in the legend in the right hand figure.

For each N_{ref} the essence of Figure 3 can be captured using BF_{ref}, d_1, and d_{ref}, and specifying one of these will render J_{ref}. One strategy could be to let the researcher interested in the evaluation of H_0 versus H_1 specify d_1 or d_{ref}. However, this would require the researcher to: (a) do this before the data are collected, (b) preregister the specification, and (c) report the choice made when BF_{01} is presented. A disadvantages of this strategy is that researchers can use specifications that support preconceptions about the truth of H_0 and H_1. They can, for example, use d_1 = .05 if evidence in favor of H_1 is desired, because with d_1 = .05 BF_{10} increases more rapidly with d (see Figure 3b) than for larger values of d_1.

Another strategy is to compute BF_{01} using a value of BF_{ref} that renders adequate operating characteristics and cannot be changed by the user. This approach has the advantage that researchers cannot influence BF_{01} via subjective choice that may favor H_0 or H_1. The remaining question is then which value of BF_{ref} renders a Bayes factor with adequate operating characteristics. In Table 1, for BF_{ref} = 9 and 19, d_1 and d_{ref} are displayed as a function of N_{ref}. Assuming that a priori H_0 and H_1 are equally likely (equal prior model probabilities) this corresponds to posterior model probabilities for H_0 and H_1 of 9/(1 + 9) = .90 and 1/(1 + 9) = .10 for BF_{ref} = 9 and 19/(1 + 19) = .95 and 1/(1 + 19) = .05 for BF_{ref} = 19. Stated otherwise, if d = 0 the Bayesian error associated with a choice in favor of H_0 is.10 and.05, respectively. Although a choice for nine or 19 is to some extend arbitrary, it comes with reasonable error probabilities and adequate operating characteristics. The latter can be illustrated using Table 1. As can be seen, medium effect sizes $d = d_{ref}$ = .50 require about N_{ref} equal to 40 or 50, to obtain BF_{01} equal to 9 and 19, respectively. As can also be seen, small effect sizes $d = d_{ref}$ = .20 require sample size larger than N_{ref} = 100 in order to obtain BF_{01} equal to 9 and 19, respectively.

This article will continue with BF_{ref} = 19, because, if in the data d = 0, there should be substantial support in favor of H_0. This choice will return in the Discussion section. With this choice, an answer to the first part of Question 2 "How can N_{ref}, d_{ref}, and J_{ref} be determined?" has been provided. The second part of Question 2 —"Is there a role for a sensitivity analysis with respect to the choice of J_{ref}?"—should now be rephrased as "Is there a role for a sensitivity analysis with respect to the choice of BF_{ref}?" The answer "no" is in line with the choice for BF_{ref} = 19. This choice reflects that the desired Bayesian error associated with a choice for H_0 if d = 0 is.05. Because there is no uncertainty about this choice, there is no need for a sensitivity analysis.

Example 1

Imagine a researcher who wants to evaluate the hypothesis H_0: $\mu = 0$ versus H_1: $\mu \neq 0$, where $\mu = \mu_1 - \mu_2$, that is, μ is the difference in means between two repeated measures of the same variable (a NHBT counterpart of the paired samples t test). The researcher will collect paired measurement for N = 50 participants in his experiment, that is, N_{ref} = 50. Based on BF_{ref} = 19 (which

Table 1
Operating Characteristics in Terms of d_1 and d_{ref} as a Function of N_{ref} and BF_{ref}

	BF_{ref} = 9[a]		BF_{ref} = 19[b]	
N_{ref}	d_1	d_{ref}	d_1	d_{ref}
10.00	0.66	0.94	0.77	1.09
20.00	0.47	0.66	0.54	0.77
30.00	0.38	0.54	0.44	0.63
40.00	0.33	0.47	0.38	0.54
50.00	0.30	0.42	0.34	0.49
60.00	0.27	0.38	0.31	0.44
70.00	0.25	0.35	0.29	0.41
80.00	0.23	0.33	0.27	0.38
90.00	0.22	0.31	0.26	0.36
100.00	0.21	0.30	0.24	0.34

[a] If BF_{ref} = 9, $b = J_{ref}/N_{ref} = .012$. [b] If BF_{ref} = 19, $b = J_{ref}/N_{ref} = .003$.

renders J_{ref} = .14, d_1 = .34, and d_{ref} = .49) the Bayes factor BF_{01} will be computed.

If in the data d =.10, BF_{01} = 14.80, that is, there is substantial support in favor of H_0. Because d =.10 is a rather small effect size and N is "only" 50, this is a reasonable quantification of the support in the data in favor of H_0. If in the data d =.40, BF_{01} = .35, there is about three times more (but not convincingly more) support for H_1 than for H_0. Because d =.40 is between a small and a medium effect size, and N is "only" 50, this too is a reasonable quantification of the support in the data in favor of H_1.

Example 2

Bayesian updating (Rouder, 2014; Schonbrodt et al., 2017; Schonbrodt & Wagenmakers, 2018) is the process where (a) for an initial batch of N_{min} persons BF_{01} is computed, and (b) after each additional (batch of) person(s) BF_{01} is recomputed, until (c) BF_{01} exceeds a prespecified threshold BF_{thresh} or the maximum sample size N_{max} has been achieved. The interested reader is referred to Appendix A where the statistical underpinnings of Bayesian updating using BF_{01} are elaborated.

Imagine again a researcher who wants to evaluate the hypotheses from Example 1, but now using Bayesian updating. The minimal sample size the researcher will collect is N_{min} = 20, the maximum achievable sample size is N_{max} = 80. The desired threshold is BF_{thresh} = 10. Compared with Example 1, while updating, the additional question is how to specify N_{ref}. When N_{min} and N_{max} are specified a natural choice seems to be $N_{ref} = (N_{min} + N_{max})/2 = 50$. Therefore, BF_{01} will be computed based on N_{ref} = 50 and BF_{ref} = 19, that is, for N = 50 the same operating characteristics as in Example 1 are obtained.

After collecting the data from 20 persons, d = .15 and BF_{01} = 9.59. Because the threshold has not been achieved, the researcher collects the data of another 20 persons. This renders d = .11 (based on all 40 persons) and BF_{01} = 13.34 which is larger than BF_{thresh} = 10, and therefore it is concluded that the data favor H_0 over H_1.

Another scenario could be that after collecting the data from 20 persons, d = .5 and BF_{10} = 1.01. Because the threshold has not been achieved, the researcher collects the data of another 20 persons. This renders d = .48 and BF_{10} = 5.90. Given that the threshold has still not been achieved, the researcher collects the data of another 20 persons. This renders d =.49 and BF_{10} = 64.54 which is larger than BF_{thresh} = 10, and it is concluded that the data convincingly favor H_1 over H_0.

Multiple Regression

The concepts and ideas introduced so far in this article generalize seamlessly to hypothesis evaluation in the context of multiple regression (this section) and analysis of variance (the next two sections). The main equation of the multiple regression model is

$$y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_M x_{Mi} + e_i, \tag{22}$$

with

$$e_i \sim \mathcal{N}(0, \sigma^2) \text{ for } i = 1, \ldots, N, \tag{23}$$

where, y_i denotes the score of person $i = 1, \ldots, N$ on the dependent variable, x_{mi} denotes the score of person i on predictor $m = 1, \ldots, M$, α denotes the intercept, β_m the regression coefficient of the mth predictor, e_i the residual of the ith person, and σ^2 the residual variance. The null and alternative hypotheses are given by

$$H_0: \beta_1 = \ldots = \beta_M = 0, \text{ that is, } H_0: R^2 = 0 \tag{24}$$

and

$$H_1: \text{ not } H_0, \text{ that is, } H_1: R^2 > 0, \tag{25}$$

where R^2 denotes the proportion of variance explained.

As is shown in Appendix B, the AAFBF for the comparison of H_0 with H_1 is

$$BF_{01} = \left(\frac{N}{J}\right)^{\frac{M}{2}} \exp\left(-\frac{N-M-1}{2}\frac{R^2}{1-R^2}\right). \tag{26}$$

As is also shown in Appendix B, like for the simple model discussed in the previous sections, with N_{ref} = N and BF_{ref} = 19 it is straightforward to compute R_1^2, R_{ref}^2 (the multiple regression counterparts of d_1 and d_{ref}), and J_{ref}.

An overview of the operating characteristics of the resulting Bayes factor is provided in Table 2. As can be seen, if the sample size is small, N_{ref} = N = 20, the Bayes factor will favor H_1 from R^2 = .25 onward and will reach 19 for R^2 = .40 (for M = 1). The

Table 2
R_1^2 and R_{ref}^2 as a Function of N_{ref}

	M = 1[a]		M = 4[b]	
N_{ref}	R_1^2	R_{ref}^2	R_1^2	R_{ref}^2
20.00	0.25	0.40	0.28	0.44
40.00	0.13	0.24	0.14	0.25
60.00	0.09	0.17	0.10	0.18
80.00	0.07	0.13	0.07	0.14
100.00	0.06	0.11	0.06	0.11
120.00	0.05	0.09	0.05	0.09
140.00	0.04	0.08	0.04	0.08
160.00	0.04	0.07	0.04	0.07
180.00	0.03	0.06	0.03	0.06
200.00	0.03	0.06	0.03	0.06

[a] For M = 1 and BF_{ref} = 19, $J_{ref}/N_{ref} = .003$. [b] For M = 4 and BF_{ref} = 19, $J_{ref}/N_{ref} = .229$.

corresponding numbers for M = 4 are .28 and .44, respectively. Consequently, with a sample as small as N = 20 only substantial proportions of variance explained provide evidence in favor of H_1. With $N_{ref} = N$ = 100 (note that R_1^2 and R_{ref}^2 become independent of M with increasing N) the Bayes factor will favor H_1 from R^2 = .06 onward and will reach 19 for R^2 = .11. These are the consequences in terms of operating characteristics if it is required that BF_{ref} = 19 if the observed R^2 = 0. Note finally, that $J_{ref}/N_{ref} = .003$ if M = 1 and .229 if M = 4. As was elaborated earlier in this article, this implies that the former renders an approximate fractional Bayes factor and the latter an approximate adjusted fractional Bayes factor, that is, an information criterion inspired by the Bayes factor.

Example 3

Imagine a researcher who predicts y from x_1 and x_2 and evaluates H_0: $\beta_1 = \beta_2 = 0$ (M = 2) versus H_1: not H_0. The researcher has collected data from N = 100 persons, that is, N_{ref} = 100. In line with the previous sections, BF_{ref} is chosen to be 19. The resulting J_{ref} = 5.26, $R_1^2 = .06$, and $R_{ref}^2 = .11$. This first of all implies that J_{ref} is relatively small compared to N = 100 which implied that BF_{01} is an approximate fractional Bayes factor. It furthermore implies that from observed $R^2 > .06$ onward, BF_{01} will express support in favor of H_1 and that BF_{01} reaches 19 for R^2 = .11. The observed R^2 turns out to be equal to.10, which results in BF_{01} = .09, that is, BF_{10} = 11.52 which is relevant support in favor of H_1 over H_0.

ANOVA

Consider the ANOVA model with possibly unequal group sizes

$$y_{ig} = \mu_g + \epsilon_{ig},\ \text{with}\ \epsilon_{ig} \sim \mathcal{N}(0, \sigma^2)\ \text{for}\ i = 1, \ldots, N_g\ \text{and}\ g = 1, \ldots, G, \tag{27}$$

where y_{ig} denotes the score of the ith person in *group g*, μ_g denotes the mean in *group g*, ϵ_{ig} the residual of person i in *group g*, and σ^2 the residual variance. With $\beta = [\mu_1 - \mu_2, \ldots, \mu_{G-1} - \mu_G]$,the null and alternative hypotheses are H_0: β = 0, that is, $H_0 : R^2 = 0$ and H_1: not H_0, that is, $H_1 : R^2 > 0$, where R^2 denotes the proportion of variance explained, in the context of ANOVA often referred to as eta-squared.

As is shown in Appendix C, the AAFBF for the comparison of H_0 with H_1 is

$$BF_{01} = \frac{(G * \frac{G^{G-1}}{J})^{\frac{1}{2}}}{|\Omega|^{\frac{1}{2}}} \exp\left(-\frac{N-G}{2}\frac{R^2}{1-R^2}\right), \tag{28}$$

where $N = \sum_g N_g$,

$$\Omega = C' \ \text{diag}[1/N_1, \ldots, 1/N_G]C, \tag{29}$$

and C is a G (the number of means) $\times$ G – 1 (the number of contrasts between means) matrix with $C[g, g]$ = 1 and $C[g,g + 1]$ = – 1 for $g = 1, \ldots, G-1$, and all other entries equal to 0, that is, BF_{01} can be computed using $N_1, \ldots, N_G$, G, J, and R^2. Note furthermore that, as is also shown in Appendix C, using $N_{ref} = N$ and BF_{ref} = 19 it is straightforward to compute R_1^2, R_{ref}^2, and J_{ref}.

ANCOVA

As is elaborated in Appendix C, it is conjectured that the approach for ANOVA generalizes straightforwardly to the ANCOVA model with equal sample sizes per group:

$$y_{ig} = \mu_g + \gamma_1 x_{1ig} + \ldots + \gamma_P x_{Pig} + \epsilon_{ig},\ \text{with}\ \epsilon_{ig} \sim \mathcal{N}(0, \sigma^2)\ \text{for}\ i = 1, \ldots, N_g\ \text{and}\ g = 1, \ldots, G, \tag{30}$$

where y_{ig} denotes the score of the ith person in *group g* on the dependent variable, μ_g denotes the adjusted mean in *group g*, γ_p the regression coefficient of the $p = 1, \ldots, P$ th covariate, x_{pig} the score of person i in *group g* on covariate p, and $N_1 = \ldots = N_G$. The null and alternative hypotheses are $H_0 : \mu_1 = \ldots = \mu_G$, that is, H_0: $R_{partial}^2 = 0$ and H_1: $R_{partial}^2 > 0$, where $R_{partial}^2$ denotes the proportion of variance explained after the effect of the covariates is partialed out, which in the context of ANCOVA is often referred to as partial eta-squared. Equation 28 applies to ANCOVA with equal sample sizes per group if the term $N_{ref} - G$ (the degrees of freedom of the residual sum of squares in an ANOVA) is replaced by $N_{ref} - G - P$ (the degrees of freedom of the residual sum of squares in an ANCOVA) and if R^2, R_1^2 and R_{ref}^2 are replaced by $R_{partial}^2$, $R_{partial,1}^2$ and $R_{partial,ref}^2$, respectively.

As is elaborated in Appendix C, with unequal sample sizes per group, Equation 28 does not apply to ANCOVA. In this case the best option currently available is to determine J_{ref} using $N_1 = \ldots = N_G = N/G$.

Example 4

Imagine a researcher who predicts y using G = 4 groups with sample sizes 25 each and evaluates H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_1: not H_0. In line with the previous sections, $N_{ref} = [25, 25, 25, 25]$ and BF_{ref} = 19 renders J_{ref} = 14.04, $R_1^2 = .06$, and $R_{ref}^2 = .11$. Because J_{ref}/N_{ref} is relatively large, BF_{01} is an approximate adjusted fractional Bayes factor, that is, an information criterion inspired by the Bayes factor. With respect to operating characteristics it can be seen that from $R_1^2 = .06$ onward BF_{01} is increasingly larger than 1 and reaches 19 at $R_1^2 = .11$.

The observed R^2 = .16 renders BF_{01} = .002, that is, BF_{10} = 500, that is, there is overwhelming evidence in favor of H_1. After the addition of two covariates x_1 and x_2, a $R_{partial}^2 = .08$ is observed. This renders BF_{01} = .32, that is, BF_{10} = 3.12, that is, there is still some evidence in favor of H_1, but it is also clear that the differences between the adjusted means are less convincing than the differences between the observed means.

The Welch Test

Consider the statistical model underlying the Welch Test, that is, the counterpart of Student's t test that does not require equal within group variances:

$$y_{ig} = \mu_g + \epsilon_{ig},\ \text{with}\ \epsilon_{ig} \sim \mathcal{N}(0, \sigma_g^2)\ \text{for}\ i = 1, \ldots, N_g\ \text{and}\ g = 1, 2, \tag{31}$$

where y_{ig} and e_{ig} denotes the score on the dependent variable and residual, respectively, of person i in *group g*, σ_g^2 is the residual variance in *group g*, and the null and alternative hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$.

As is shown in Appendix D, the AAFBF for the comparison of H_0 with H_1 is:

$$BF_{01} = \frac{\sqrt{\frac{2}{J} + \frac{2}{J} r}}{\sqrt{\frac{1}{N_1} + \frac{r}{N_2}}} \exp\left(-\frac{1}{2} d^2 \frac{\frac{(N_1-1)+(N_2-1)r}{N_1+N_2-2}}{\frac{1}{N_1} + \frac{r}{N_2}} \right), \tag{32}$$

where

$$d^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}}, \tag{33}$$

and $r = s_2^2/s_1^2$. As is also shown in Appendix D, such as the models discussed in the previous sections, with r equal to the variance ratio in the data set at hand (which with increasing sample sizes converges to the true value), $N_{1,ref}, N_{2,ref}$ and BF_{ref} = 19, it is straightforward to compute d_1, d_{ref}, and J_{ref}.

Example 5

Imagine a researcher who wants to evaluate the $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$, using Welch's Test and Bayesian updating. The minimal sample size the researcher will collect is N_{min} = 20 per group, the maximum achievable sample size is N_{max} = 80 per group, and therefore N_{ref} = 50 per group. The desired threshold is BF_{thresh} = 10. Using BF_{ref} = 19 results in the operating characteristics displayed in the middle line of Table 3. As can be seen, compared with each of the sample sizes J_{ref} is small, that is, BF_{01} is an approximate fractional Bayes factor. It can furthermore be seen that with $N = N_{ref} = 50$ and r = 2.02, BF_{01} = 1 if d = .49 and BF_{01} = 19 if d = .69. Note that these values where recomputed for each observed value or r, however, because the numbers hardly changed, they are only displayed for the reference sample size.

The results of the updating process are also summarized in Table 3. As can be seen, with 20 persons per group, the observed r = 1.90, d = .52, and BF_{01} = 3.11. There is some, but not convincing, evidence in favor of H_0, but the desired threshold of 10 has not been achieved. Increasing the sample size in steps of 10 per group shows that finally, at the maximum sample size of 80 per group, the variance ratio r has converged to 2.00, d to .53 and BF_{01} = .09, that is, BF_{10} = 11.11 that is, the threshold has been achieved and there is substantial evidence in favor of H_1.

Table 3
An Application of Welch's Test

N_1	N_2	r	J_{ref}	d_1	d_{ref}	d	BF_{01}
20.00	20.00	1.90				0.52	3.11
30.00	30.00	2.10				0.49	2.43
40.00	40.00	2.05				0.50	1.39
50.00	50.00	2.02	0.28	0.49	0.69	0.51	0.74
60.00	60.00	2.01				0.52	0.36
70.00	70.00	2.00				0.53	0.16
80.00	80.00	2.00				0.53	0.09

Frequency Properties of the Bayes Factor Calibrated Using BF_{ref} = 9, 19, and 99

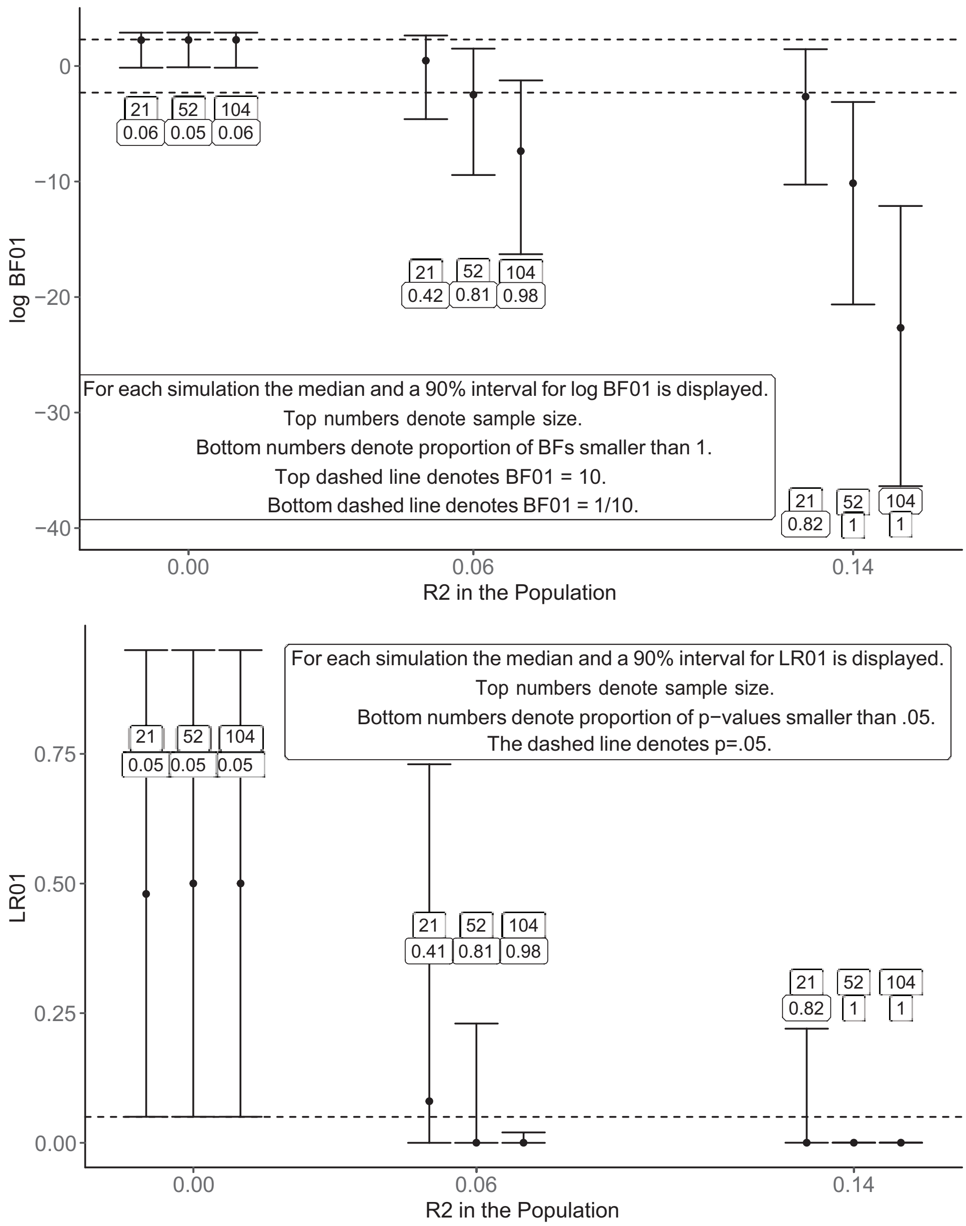
This article has shown that the Bayes factor can be calibrated such that BF_{01} = 19 if the observed effect size equals zero, that is, the support in the data D for H_0 is 19 times stronger than the support for H_1. The choice for BF_{ref} = 19 was inspired by the fact that, using equal prior model probabilities $P(H_0) = P(H_1) = .5$, the posterior model probabilities are $P(H_0 \mid D) = .95$ and $P(H_1 \mid D) = .05$, respectively. This means that numerically $P(H_1 \mid D)$ is equal to the usual α level of.05. However, note that, the interpretation of $P(H_1 \mid D)$ and α are quite different. The former is the Bayesian error, that is, the probability of incorrectly rejecting H_0 *given the information in the data set at hand*, while the latter is the probability of incorrectly rejecting H_0 *if data are repeatedly sampled from a population in which H_0 is true*. To clarify this further, in this section, the frequency properties of BF_{01} calibrated using $BF_{ref} = 9, 19, 99$ will be explored and compared to those of the likelihood ratio test.

Nine populations consisting of G = 3 groups and one dependent variable (an ANOVA setup) were constructed. Two factors were manipulated: $R^2 = 0, .06, .14$ which are populations were, consecutively, H_0 is true, and, H_1 is true with (according to Cohen, 1992) medium and large effect sizes; and N = 21, 52, 104 per group (according to Cohen, 1992; with α = .05 a large effect size can be detected with a power of .80 if N = 21 and a medium effect size if N = 52). From each population 10,000 data sets were simulated and for each BF_{01} calibrated using $BF_{ref} = 9, 19, 99$ and LR_{01} were computed. Frequentist error probabilities and power were computed using the proportion of BF_{01} smaller than 1 (which denotes evidence in favor of H_1) and the proportion of LR_{01} smaller than .10, .05, and .01, respectively. The results for BF_{ref} = 19 and α = .05 are displayed in Figure 4.

A number of features can be observed. As can be seen in the top panel of Figure 4, the Bayes factor is a measure of support that can express support in favor of H_0 (values larger than 1) and H_1 (values smaller than 1). As can be seen in the bottom panel of Figure 4 the likelihood ratio is not a measure of support because it always expresses support in favor of H_1 (all values are smaller than 1). As can furthermore be seen in the top panel of Figure 4, choosing BF_{ref} = 19 implies that BF_{01} equals 19 if the observed value of R^2 = 0. However, if data sets are repeatedly sampled from a population in which R^2 = 0, the observed values of R^2 are usually somewhat larger than zero resulting in Bayes factor values that range from slightly smaller than one to almost 19 with a median value of about 10.

As can be seen in Figure 4 the frequency properties of the Bayes factor and the likelihood ratio test are about equal. When the null-hypothesis is true, both the Bayes factor and the likelihood ratio test have a nominal α level of about .05. Furthermore, when 52 persons per group are used, a medium effect size is detected with a power of about .80, and when 21 persons per group are used, a large effect size is detected with a power of about .80. This figure suggests (but as will be elaborated below, incorrectly) that the frequency properties of the Bayes factor and the likelihood ratio test are similar. This is further highlighted in Table 4 which compares the frequency properties of the Bayes factor (for $BF_{ref} = 99, 19, 9$, that is, $P(H_1 \mid D) = .01, .05, .10$ if the observed effect size equals zero) with those of the likelihood ratio test with α levels of .01, .05, and

Figure 4
Frequency Properties of BF_{01} Calibrated Using BF_{ref} = 19 and LR_{01} Evaluated Using α = .05



.10, respectively. However, this similarity is a coincidence as will be elaborated in the next paragraph.

In Table 5 the results of a simulation study based on a multiple regression model using one and three predictors, respectively, are displayed (both the dependent variable and the predictors are standardized and the correlation among the three predictors are set to .3). Again R^2 and N are based on Cohen (1992). As can be seen in both the top and bottom panel of Table 5 the correspondence between the frequency properties of the Bayes factor and the likelihood ratio test that was observed in Table 4 is lost. Stated otherwise, if the observed effect size equals zero, for the Bayes factor there is no longer a close correspondency between $P(H_1 \mid D)$ and the nominal error probabilities, whereas for the likelihood ratio test, the α level and the nominal error probabilities are rather close and asymptotically the same. Also, for the likelihood ratio test, the power is close to the nominal level according to Cohen (1992; see, e.g., the .82 and .80 in the fourth line of numbers in the top panel of Table 5). For the Bayes factor the power levels can be higher or lower, depending on whether the nominal α level is higher or lower, respectively.

As this section has highlighted, the strength of the Bayes factor is its interpretation as a measure of relative support which can, assuming equal prior model probabilities, also be expressed in terms of posterior model probabilities. However, both with respect to their interpretation and their numerical value, posterior model probabilities are *not* frequentist error probabilities.

Table 4
ANOVA: Frequency Properties of the Bayes Factor Calibrated Using BF_{ref}

ANOVA with G = 3 groups									
	$R^2 = 0$			$R^2 = .06$			$R^2 = .14$		
N per group	21	52	104	21	52	104	21	52	104
$BF_{ref} = 99$	.01	.01	.01	.21	.60	.93	.62	.98	1
$\alpha = .01$	.01	.01	.01	.20	.60	.93	.60	.98	1
$BF_{ref} = 19$	.06	.05	.06	.42	.81	.98	.82	1	1
$\alpha = .05$	.05	.05	.05	.41	.81	.98	.82	1	1
$BF_{ref} = 9$	.12	.12	.11	.55	.89	.99	.90	1	1
$\alpha = .10$	.11	.11	.10	.54	.88	.99	.89	1	1

Note. The table contains the proportion of BF_{01} smaller 1 and proportion of p-values smaller than α obtained for 10,000 simulated data sets.

On the other hand, the likelihood ratio cannot be interpreted as a measure of support, but the likelihood ratio test has excellent frequency properties. Therefore, researchers wanting to test H_0 versus H_1 face a principled decision: Do they prefer to evaluate these hypotheses in terms of support or in terms of frequentist error probabilities? Readers interested in exploring the Bayesian point of view are referred to Wagenmakers (2007) who discusses classical and Bayesian inference when the focus is on hypothesis evaluation.

Discussion

In my experience, users of Bayes factors almost exclusively want to determine the support in the data for the null and alternative hypotheses. Stated otherwise, they do not want to determine the support for and specify the prior distributions corresponding to the null an alternative hypotheses. To be able to do this, fully specified prior distributions that do not require user input are required. As was illustrated using four Bayes factors for the evaluation of H_0: $\mu = 0$ versus H_1: $\mu \neq 0$ this can currently be achieved using default values for the unknown scale parameters that are proposed by the developers of these Bayes factors. However, as was also shown, in terms of the operating characteristics of the resulting Bayes factors, there is no justification for these default values.

In this article it is proposed to use reference scale parameters (J_{ref} for the AAFBF) chosen such that BF_{01} = 19 for a data set in which the effect size (d or R^2) is equal to zero. This renders a clearly defined Bayes factor with clear and adequate operating characteristics that can be summarized using N_{ref}, and d_1 or R_1^2 and d_{ref} or R^2_{ref}. This has been elaborated for the AAFBF applied to NHBT for the one group model, multiple regression, AN(C)OVA, and Welch's Test. Annotated `R` functions and examples showing how to determine J_{ref} and how to use it for NHBT using `bain` can be downloaded from the bain website.

Beyond the context sketched in the previous two paragraphs, there still remain issues with respect to NHBT deserving further attention that will be addressed in this section. In contrast to NHBT, the operating characteristics of IHBT do not depend on the prior scale parameter; that is, the prior distributions follow directly from the hypotheses under consideration (Mulder, 2014). In line with Cohen

Table 5
Multiple Regression: Frequency Properties of the Bayes Factor Calibrated Using BF_{ref}

Multiple regression with M = 1 predictors									
	$R^2 = 0$			$R^2 = .13$			$R^2 = .26$		
N	26	58	116	26	58	116	26	58	116
$BF_{ref} = 99$	.01	.00	.00	.17	.46	.86	.47	.90	1
$\alpha = .01$	.02	.01	.01	.26	.61	.94	.60	.95	1
$BF_{ref} = 19$	.02	.02	.02	.33	.68	.95	.67	.96	1
$\alpha = .05$	.06	.06	.05	.49	.82	.99	.80	.99	1
$BF_{ref} = 9$	.05	.04	.04	.44	.78	.98	.77	.98	1
$\alpha = .10$	.12	.11	.11	.61	.89	.99	.88	1	1
Multiple regression with M = 3 predictors									
	$R^2 = 0$			$R^2 = .13$			$R^2 = .26$		
N	34	76	152	34	76	152	34	76	152
$BF_{ref} = 99$	.04	.04	.03	.35	.73	.97	.73	.98	1
$\alpha = .01$	.01	.01	.23	.59	.94	.93	.60	.96	1
$BF_{ref} = 19$	.14	.13	.12	.59	.89	.99	.89	1	1
$\alpha = .05$	.07	.06	.05	.45	.80	.98	.80	.99	1
$BF_{ref} = 9$	.25	.23	.22	.72	.94	1	.94	1	1
$\alpha = .10$	.13	.12	.10	.57	.88	.99	.88	1	1

Note. The table contains the proportion of BF_{01} smaller 1 and proportion of p-values smaller than α obtained for 10,000 simulated data sets.

(1994) and Royal (1997, pp. 79–81) it is therefore argued that null-hypotheses should only be considered if they present a plausible description of the state of affairs in the population of interest. If they do not, use IHBT which is straightforward. If they do, see Wainer (1999) for examples where null hypotheses are plausible. Realize that NHBT requires a reference scale value, which, however reasonable, is still a choice that needs to be communicated.

Issue 1: A Multiverse of Bayes Factors

In the beginning of this article, four Bayes factors were introduced. However, there may very well be more Bayes factors that can be used to evaluate H_0: $\mu = 0$ versus H_1: $\mu \neq 0$. Stated otherwise, there is a multiverse of Bayes factors that render a different quantification of the support in the data, not only for the one group model, but also for other statistical models. There is no objective reason to prefer one of these Bayes factors over the other, because, if the goal is to evaluate H_0 versus H_1 and not the corresponding prior distributions, the only objective information with respect to H_0 and H_1 is contained in the data. Further research is needed to address the consequences of this multiverse and to show how to deal with it when evaluating null hypotheses.

It is conjectured that this problem is more or less irrelevant for IHBT. Momentarily, IHBT in the context of the normal linear model can be executed with `bain` (based on a normal approximation of the posterior distribution) and `BFpack` (based on a $\mathcal{T}$ posterior distribution). For N not too small, the $\mathcal{T}$ posterior distribution converges to a normal and both Bayes factors will render the same results independent of whether normal or $\mathcal{T}$ prior distributions are used.

Issue 2: Nevertheless a Subjective Choice

However reasonable it is to choose J_{ref} such that $BF_{01} = 19$ if the observed effect size equals 0 and $N_{ref} = N$ (which renders a Bayesian error of .05 if H_0 is preferred over H_1), it remains annoying that for NHBT the results depend on this subjective choice. Further discussion among Bayes factor developers and users should address whether "19" is the number we can agree on. As argued in this article, the number should not be user-specified, because it would allow the user to calibrate the Bayes factor such that it becomes less (or more) supportive of H_0. An exploration of both this and the previous issue will benefit from an implementation of the approach proposed in this article for the Bayes factors contained in the `R` packages `BFpack` and `BayesFactors`.

Issue 3: Beyond the Models Discussed in This Article

This article developed a calibration of the AAFBF for null hypotheses evaluated for the one group model, multiple regression, AN(C)OVA, and Welch's Test. It is up to unexplored territory how to calibrate the AAFBF when used for NHBT in other models like logistic regression and structural equation modeling and also for the ANCOVA with unequal group sizes. This is an area for further research that will be explored in the future.

Issue 4: Reporting the Results of NHBT

As was discussed in in this article, for the AAFBF the outcomes of IHBT do not depend on J and it is therefore an approximate fractional Bayes factor. Furthermore, for N not too small the normal approximation of the posterior distribution will be very accurate and the AAFBF is a fractional Bayes factor. However, for the AAFBF the outcomes of NHBT do depend on J, therefore it is an approximate *adjusted* Bayes factor, or, in other words, a (as was shown in Appendix A) consistent information criterion inspired by the Bayes factor. If J is relatively small compared with N and/or if the observed effect size is close to zero, the AAFBF behaves like an approximate fractional Bayes factor. However, if J is large compared with N and the observed effect size is not close to zero, the AAFBF is an information criterion. When using the AAFBF for NHBT and the approach proposed in this article it is therefore important that users report: (a) that the AAFBF was used; (b) it was calibrated choosing J_{ref} such that $BF_{01} = 19$ if the observed effect size equals 0 and $N_{ref} = N$; and (c) whether J_{ref}/N_{ref} is small (say, smaller than .05) which allows an interpretation of the outcomes as approximate Bayes factors, or large (say, larger than .05) in which case the outcomes should be interpreted as Bayes factor inspired information criteria. Note that, with the availability of a multiverse of Bayes factors, reporting of (a) which Bayes factor was used, and (b) which scale parameter was used, is always necessary. This is an issue because currently this information is not reported by many users of Bayes factors.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round, ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Gilks, W. R. (1995). Discussion of fractional Bayes factors for model comparison (by O'Hagan). *Journal of the Royal Statistical Society Series B*, *57*(1), 118–120. http://www.jstor.org/stable/2346088

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Gu, X., Mulder, J., Dekovic, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511–527. https://doi.org/10.1037/met0000017

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hoijtink, H., Gu, X., & Mulder, J. (2019). Bain, multiple group Bayesian evaluation of informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 219–243. https://doi.org/10.1111/bmsp.12145

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*, *51*(1), 2–10. https://doi.org/10.1080/00273171.2014.969364

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Boing-Messing, F., Hoijtink, H., Meijerink, M., Williams, D. R., Menke, J., Fox, J.-P., Roseel, Y., Wagenmakers, E.-J., & van Lissa, C. (2019). *Flexible Bayes factor testing of scientific theories in R*. https://arxiv.org/abs/1911.07728

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society. Series B*, *57*(1), 99–138. http://www.jstor.org/stable/2346088

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Royal, R. (1997). *Statistical evidence. A likelihood paradigm*. Chapman and Hall/CRC.

Schonbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Schonbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Tendeiro, J., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *15*, 779–804. https://doi.org/10.3758/BF03194105

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*(2), 212–213. https://doi.org/10.1037/1082-989X.4.2.212

Appendix A

Consistency of and Udating Using the AAFBF

Consistenty of the AAFBF

Note that, Equation 15 can be written as:

$$BF_{\mathcal{N}} = \frac{\mathcal{N}(\mu = 0 \mid \bar{y}, \frac{s^2}{N})}{\mathcal{N}(\mu = 0 \mid 0, \frac{s^2}{J})} = \frac{\sqrt{N}}{\sqrt{J}} \exp\left(-\frac{N}{2}\frac{\bar{y}^2}{s^2}\right) = \frac{\sqrt{N}}{\sqrt{J}} \exp\left(-\frac{N}{2} d^2\right). \tag{34}$$

If H_0 is true, then $d \rightarrow 0$ for $N \rightarrow \infty$, and the remaining term $\frac{\sqrt{N}}{\sqrt{J}} \rightarrow \infty$ because J is fixed. If H_1 is true, as can be seen from the second term of Equation 34, then for $N \rightarrow \infty$ $s^2 \rightarrow \sigma^2$, that is, the denominator becomes a constant, while in the nominator $\bar{y} \rightarrow \mu$ and the variance $s^2/N \rightarrow 0$, that is, the nominator goes to zero and, consequently, $BF_{\mathcal{N}} \rightarrow 0$. This implies that $BF_{\mathcal{N}}$ is consistent.

Updating Using $BF_{\mathcal{N}}$

Note that, in this paragraph, the subscripts for $\bar{y}$ and s^2 denote upon which sample size each is based. Using this notation, Equation 16 becomes:

$$\mathcal{N}\left(\mu \mid \bar{y}_N, \frac{s_N^2}{N}\right) \approx \mathcal{N}\left(\mu \mid \bar{y}_N, \frac{s_N^2}{N}\right)^{1-\frac{J_{ref}}{N}} \times \mathcal{N}\left(\mu \mid 0, \frac{s_N^2}{J_{ref}}\right). \tag{35}$$

The updated counterpart after the addition of N_{add} additional observations is:

$$\mathcal{N}\left(\mu \mid \bar{y}_{N+N_{add}}, \frac{s^2_{N+N_{add}}}{N+N_{add}}\right) \approx \mathcal{N}\left(\mu \mid \bar{y}_{N+N_{add}}, \frac{s^2_{N+N_{add}}}{N+N_{add}}\right)^{1-\frac{J_{ref}}{N+N_{add}}} \times \mathcal{N}\left(\mu \mid 0, \frac{s^2_{N+N_{adj}}}{J_{ref}}\right), \tag{36}$$

which for $N_{add} \rightarrow \infty$ reduces to

$$\approx \mathcal{N}\left(\mu \mid \bar{y}_N, \frac{s_N^2}{N}\right) \times \mathcal{N}\left(\mu \mid \bar{y}_{N_{add}}, \frac{s^2_{N_{add}}}{N_{add}}\right) \times \mathcal{N}\left(\mu \mid 0, \frac{\sigma^2}{J_{ref}}\right), \tag{37}$$

because $J_{ref}/(N + N_{add}) \rightarrow 0$ and $s^2_{N+N_{add}} \rightarrow \sigma^2$. Asymptotically, updating using the posterior and prior distributions upon which $BF_{\mathcal{N}}$ is based, is similar to the classical "update a fixed prior into a posterior using additional batches of data."

It is important to stress that, while updating, it is not an option to make J_{ref} a function of N instead of N_{ref} because the resulting Bayes factor is inconsistent. Inserting Equation 19 with N_{ref} replaced by N into the last term of Equation 34 renders:

(*Appendices continue*)

$$BF_{01} = \frac{\sqrt{N}}{\sqrt{J_{ref}}}\exp\left(-\frac{N}{2}d^2\right)$$
$$= \frac{\sqrt{N}}{\sqrt{N\exp\left(-\frac{N}{2}d_{ref}^2\right)}}\exp\left(-\frac{N}{2}d^2\right)$$
$$= \exp\left(-\frac{N}{2}\left(d^2 - \frac{1}{2}d_{ref}^2\right)\right). \tag{38}$$

If H_1 is true and $N \to \infty$, then $d \to \delta$. If $0 < \delta^2 < \frac{1}{2}d_{ref}^2$ then $BF_{01} \to \infty$, which is inconsistent.

Appendix B

Derivations for the Multiple Regression Model

The approximate adjusted fractional Bayes factor for the comparison of H_0 with H_1 is:

$$BF_{01} = \frac{\mathcal{N}(\beta_1 = 0 \quad \ldots \quad \beta_M = 0 \mid \hat{\beta}, \Sigma_\beta)}{\mathcal{N}(\beta_1 = 0 \quad \ldots \quad \beta_M = 0 \mid 0, \frac{N}{J}\Sigma_\beta)}$$
$$= \left(\frac{N}{J}\right)^{\frac{M}{2}}\exp\left(-\frac{1}{2}\beta'\Sigma_\beta^{-1}\beta\right) \tag{39}$$

where $\hat{\beta}$ denotes the estimates of the regression coefficients and Σ^β their covariance matrix. This can be simplified to

$$BF_{01} = \left(\frac{N}{J}\right)^{\frac{M}{2}}\exp\left(-\frac{1}{2}\frac{\beta'X'X\beta}{s^2}\right) \tag{40}$$

because $\Sigma_{\alpha,\beta} = s^2(Z'Z)^{-1}$, where Z is a $(1 + M) \times N$ matrix containing a column of 1's and the scores on all M predictors. Σ_β is obtained by deleting the first row and column of $\Sigma_{\alpha,\beta}$, and $\Sigma_\beta^{-1} = X'X/s^2$ where X is a $M \times N$ matrix containing the scores on the centered predictors for $m = 1, \ldots, M$.

Because $\beta'X'X\beta$ is the explained sum of squares, and $(N - M - 1)s^2$ is the unexplained sum of squares

$$BF_{01} = \left(\frac{N}{J}\right)^{\frac{M}{2}}\exp\left(-\frac{N-M-1}{2}\frac{E}{U}\right)$$
$$= \left(\frac{N}{J}\right)^{\frac{M}{2}}\exp\left(-\frac{N-M-1}{2}\frac{R^2}{1-R^2}\right), \tag{41}$$

where E denotes the explained sum of squares, U the unexplained sum of squares, and R^2 the proportion of variance explained.

With sample size N_{ref}, BF_{01} computed for a data set in which $R^2 = 0$ is equal to BF_{10} computed for a data set in which $R^2 = R^2_{ref}$ if

$$\left(\frac{N_{ref}}{J_{ref}}\right)^{\frac{M}{2}} = \left(\frac{N_{ref}}{J_{ref}}\right)^{-\frac{M}{2}}\exp\left(\frac{N_{ref}-M-1}{2}\frac{R^2_{ref}}{1-R^2_{ref}}\right) \tag{42}$$

which results in

$$J_{ref} = N_{ref}\exp\left(-\frac{N_{ref}-M-1}{2M}\frac{R^2_{ref}}{1-R^2_{ref}}\right). \tag{43}$$

The Bayes factor value in the equilibrium is equal to

$$BF_{ref} = (\frac{N_{ref}}{J_{ref}})^{\frac{M}{2}} = \exp\left(\frac{N_{ref}-M-1}{4}\frac{R^2_{ref}}{1-R^2_{ref}}\right). \tag{44}$$

Using $N = N_{ref}$ and $J = J_{ref}$, Equation 41 reduces to

$$BF_{01} = \exp\left(\frac{N_{ref}-M-1}{4}\left(\frac{R^2_{ref}}{1-R^2_{ref}} - 2\frac{R^2}{1-R^2}\right)\right), \tag{45}$$

which implies that $BF_{01} = 1$ if

$$\frac{R^2_{ref}}{1-R^2_{ref}} = 2\frac{R^2}{1-R^2}, \tag{46}$$

that is, $R_1^2 = x/(1+x)$, with

$$x = \frac{1}{2}\frac{R^2_{ref}}{1-R^2_{ref}}. \tag{47}$$

Like for the simple model discussed in the first half of the article, given $N_{ref} = N$, one of BF_{ref}, R^2_{ref} and R_1^2 is sufficient to specify the other two.

(*Appendices continue*)

Appendix C

Derivation of J_{ref} for ANOVA With Unequal Group Sizes

The approximate adjusted fractional Bayes factor for the comparison of H_0 with H_1 is:

$$BF_{01} = \frac{\mathcal{N}(\beta = 0 \mid \hat{\beta}, \Sigma_\beta)}{\mathcal{N}(\beta = 0 \mid 0, \Phi_\beta)} = \frac{|\Sigma_\beta|^{-\frac{1}{2}} \exp(-\frac{1}{2}\hat{\beta}'\Sigma_\beta^{-1}\hat{\beta})}{|\Phi_\beta|^{-\frac{1}{2}}}, \tag{48}$$

where the mean vector of the posterior distribution $\hat{\beta} = [\overline{y}_1 - \overline{y}_2, \ldots, \overline{y}_{G-1} - \overline{y}_G]$, the covariance matrix of the posterior distribution

$$\Sigma_\beta = C'\Sigma_\mu C, \tag{49}$$

where, $\Sigma_\mu = \ \mathrm{diag}[s^2/N_1, \ldots, s^2/N_G] = s^2\Omega$ and C is a G (the number of means) $\times$ $G - 1$ (the number of contrasts between means) matrix with $C[g, g] = 1$ and $C[g, g + 1] = -1$ for $g = 1, \ldots, G - 1$, and all other entries equal to 0, and the covariance matrix of the prior distribution

$$\Phi_\beta = C'\Phi_\mu C, \tag{50}$$

where $\Phi_\mu = \ \mathrm{diag}[\frac{G}{J}s^2, \ldots, \frac{G}{J}s^2] = s^2\frac{G}{J}C'C$. In the same manner as in Appendix B, Equation 48 can be written as

$$BF_{01} = \frac{|\Sigma_\beta|^{-\frac{1}{2}} \exp(-\frac{N-G}{2}\frac{R^2}{1-R^2})}{|\Sigma_\Phi|^{-\frac{1}{2}}}, \tag{51}$$

where $N = \sum_g N_g$.

Equal support for $R^2 = 0$ and $R^2 = R^2_{ref}$ is obtained if J is chosen such that

$$\frac{|\Sigma_\beta|^{-\frac{1}{2}}}{|\Phi_\beta|^{-\frac{1}{2}}} = \frac{|\Phi_\beta|^{-\frac{1}{2}}}{|\Sigma_\beta|^{-\frac{1}{2}} \exp\left(-\frac{N_{ref}-G}{2}\frac{R^2_{ref}}{1-R^2_{ref}}\right)}, \tag{52}$$

that is,

$$|C'\Omega_{ref}C|^{-1} \exp\left(-\frac{N_{reg} - G}{2}\frac{R^2_{ref}}{1 - R^2_{ref}}\right) = \left|\frac{G}{J}C'C\right|^{-1}. \tag{53}$$

Solving for J renders:

$$J_{ref} = G * \left(\frac{|C'\Omega_{ref}C|}{G}\right)^{G-1} \exp\left(-\frac{N_{ref} - G}{2(G-1)}\frac{R^2_{ref}}{1 - R^2_{ref}}\right), \tag{54}$$

where the subscript *ref* in Ω_{ref} denotes that this matrix is based on the reference sample sizes in each group. Combining Equation 51 with $R^2 = 0$ and 54 renders

$$BF_{ref} = \exp\left(\frac{N_{ref} - G}{4}\frac{R^2_{ref}}{1 - R^2_{ref}}\right). \tag{55}$$

Using $N = N_{ref}$ and $J = J_{ref}$ Equation 51 can be rewritten as

$$BF_{01} = \exp\left(\frac{N_{ref} - G}{4}\left(\frac{R^2_{ref}}{1 - R^2_{ref}} - 2\frac{R^2}{1 - R^2}\right)\right). \tag{56}$$

which implies that $BF_{01} = 1$ if

$$\frac{R^2_{ref}}{1 - R^2_{ref}} = 2\frac{R^2}{1 - R^2}, \tag{57}$$

that is, $R^2_1 = x/(1 + x)$, with

$$x = \frac{1}{2}\frac{R^2_{ref}}{1 - R^2_{ref}}. \tag{58}$$

Like for the simple model discussed in the first half of the paper, given $N_{ref} = N$, one of BF_{ref}, R^2_{ref} and R^2_1 is enough to specify the other two. Note that, for $N_1 = \ldots = N_G$, the equations presented in this section are equal to those presented in the previous section.

ANCOVA

It is conjectured that Equations 51 through 58 also apply to ANCOVA with equal sample sizes per group if the terms $N - G$ and $N_{ref} - G$ are replaced by $N - G - P$ and $N_{ref} - G - P$, respectively, and if R^2, R^2_1, and R^2_{ref} are replaced by the corresponding partial proportions of variance, that is, what is the proportion of variance explained by the grouping variable, if the effects of the covariates are partialed out. This conjecture is supported by experiments with the following setup: (a) create a data set containing a dependent variable, a grouping variable (with equal sample sizes per group) and covariates; (b) compute the partial proportion of variance; (c) compute BF_{01} using Equation 28; and (d) compute BF_{01} using Equation 48 which is implemented in the `R` package `bain`. In all experiments the Bayes factors resulting from (c) and (d) where exactly the same. The same does not hold for ANCOVA with unequal sample sizes per group. This is shown by one experiment in which (a) through (d) were repeated for a data set with unequal sample sizes per group. The Bayes factors resulting from (c) and (d) were not the same.

(*Appendices continue*)

Appendix D

Derivation of J_{ref} and BF_{ref} for Welch's Test

The approximate adjusted fractional Bayes factor for the comparison of H_0 with H_1 is:

$$BF_{01} = \frac{\mathcal{N}(\beta = 0 \mid \hat{\beta}, \Sigma_\beta)}{\mathcal{N}(\beta = 0 \mid 0, \Phi_\beta)} = \frac{\sqrt{\Phi_\beta}}{\sqrt{\Sigma_\beta}} \exp\left(-\frac{1}{2}\frac{\hat{\beta}^2}{\Sigma_\beta}\right) \tag{59}$$

where the mean of the posterior distribution $\hat{\beta} = \overline{y}_1 - \overline{y}_2$ the variance of the posterior distribution $\Sigma_\beta = s_1^2/N_1 + s_2^2/N_2$, and the variance of the prior distribution $\Phi_\beta = 2/J * s_1^2 + 2/J * s_2^2$. Rewriting in terms of Cohen's d:

$$d^2 = \frac{(\overline{y}_1 - \overline{y}_2)^2}{\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}} \tag{60}$$

renders

$$\begin{aligned} BF_{01} &= \frac{\sqrt{\frac{2}{J}s_1^2 + \frac{2}{J}s_2^2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \exp\left(-\frac{1}{2}d^2 \frac{\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}\right) \\ &= \frac{\sqrt{\frac{2}{J} + \frac{2}{J}r}}{\sqrt{\frac{1}{N_1} + \frac{r}{N_2}}} \exp\left(-\frac{1}{2}d^2 \frac{\frac{(N_1-1) + (N_2-1)r}{N_1+N_2-2}}{\frac{1}{N_1} + \frac{r}{N_2}}\right) \end{aligned} \tag{61}$$

where $r = s_2^2/s_1^2$.

Equal support for $d = 0$ and $d = d_{ref}$ is obtained if J is chosen such that

$$\frac{\sqrt{\frac{2}{J} + \frac{2}{J}r}}{\sqrt{\frac{1}{N_1} + \frac{r}{N_2}}} = \frac{\sqrt{\frac{1}{N_1} + \frac{r}{N_2}}}{\sqrt{\frac{2}{J} + \frac{2}{J}r}} \exp\left(\frac{1}{2}d^2 \frac{\frac{(N_1-1) + (N_2-1)r}{N_1+N_2-2}}{\frac{1}{N_1} + \frac{r}{N_2}}\right), \tag{62}$$

that is,

$$\frac{1}{J}\frac{2+2r}{\frac{1}{N_1} + \frac{r}{N_2}} = \exp\left(\frac{1}{2}d^2 \frac{\frac{(N_1-1)+(N_2-1)r}{N_1+N_2-2}}{\frac{1}{N_1} + \frac{r}{N_2}}\right), \tag{63}$$

Solving for J with $N_{1,ref} = N_1$ and $N_{2,ref} = N_2$ and the observed value of r (which for $N_1, N_2 \to \infty$ converges to the true value) renders:

$$J_{ref} = \frac{2+2r}{\frac{1}{N_{1,ref}} + \frac{1}{N_{2,ref}}r} \exp\left(-\frac{1}{2}d_{ref}^2 \times \frac{\frac{(N_{1,ref}-1)+(N_{2,ref}-1)r}{N_{1,ref}+N_{2,ref}-2}}{\frac{1}{N_{1,ref}} + \frac{1}{N_{2,ref}}r}\right), \tag{64}$$

and also:

$$BF_{ref} = \exp\left(\frac{1}{4}d_{ref}^2 \times \frac{\frac{(N_{1,ref}-1)+(N_{2,ref}-1)r}{N_{1,ref}+N_{2,ref}-2}}{\frac{1}{N_{1,ref}} + \frac{1}{N_{2,ref}}r}\right). \tag{65}$$

If $N = N_{ref}$ and $J = J_{ref}$, Equation 61 can be rewritten as:

$$BF_{01} = \exp\left(\frac{1}{4}(d_{ref}^2 - 2d^2) \times \frac{\frac{(N_{1,ref}-1) + (N_{2,ref}-1)}{N_{1,ref}+N_{2,ref}-2}}{\frac{1}{N_{1,ref}} + \frac{1}{N_{2,ref}}r}\right), \tag{66}$$

that is, $d_1 = \sqrt{.5d_{ref}^2}$.

Received April 22, 2020
Revision received March 4, 2021
Accepted March 8, 2021 ■