Research Report:

Prior Sensitivity of Null Hypothesis Bayesian Testing in the Context of Two-level Models

Author: Nikola Sekulovski (6465588) *Supervisor:* prof. dr. Herbert Hoijtink

Department of Methodology and Statistics, Utrecht University

December 2021

This *Research Report* is written as a "mini-thesis," intended to serve as an introduction paper to the Master's Thesis. Preferred journals for publication (of the Master's Thesis) might include (among other): *Psychological Methods* and *British Journal of Mathematical and Statistical Psychology.*

Word count: 2515 (excluding: references, title page, tables and figures).

Correspondence concerning this article should be addressed to n.sekulovski@uu.nl.

Keywords: Bayes Factor, Two-level Models, Null Hypotheses, Prior Sensitivity.

Introduction

Following an increasing wave of criticism directed towards *Null Hypothesis Significance Testing (NHST)* (Cohen, 1994; Wagenmakers, 2007), the Bayesian approach to evaluating hypotheses, epitomized by the *Bayes Factor* (abbreviated as *BF*, Jeffreys, 1935) is gaining momentum. This paradigm, at least compared to NHST, is said to give more intuitive results and, most importantly, does not rely on strict cutoff values such as the often vilified "$\alpha = .05$" (see, for example, Hoijtink, Mulder, Lissa, & Gu, 2019 for further elaborations on the strengths of the BF). However, novel statistical methods should not be taken for granted, and their drawbacks should always be considered and, hopefully, addressed. Hence, the main topic of this paper, which is to *illustrate* the sensitivity of the BF, to the specification of the prior distribution, when evaluating null hypotheses *in the context of (linear) two-level models.*[1]

Multilevel models are useful when the data has a hierarchical structure, for example, when individuals are nested within groups. This allows the researcher to take the within-group dependence into account. In such models, variables can be defined at different levels. *Two-level models* are the most common (and simple) example, with applications ranging from organizational research (e.g., employees nested within companies) to longitudinal studies (e.g., observations nested within individuals). The focus of this paper is on *linear* two-level models i.e., models that have a continuous dependent variable (see, for example, Hox, Moerbeek, & Van de Schoot, 2017).

This brief paper is intended to serve as a predecessor to a more detailed paper that will aim to *address* the sensitivity of the BF in two-level models, based on the work presented in Hoijtink (2021). Throughout the text, the application of the BF in evaluating *null hypotheses* is referred to as *Null Hypotheses Bayesian Testing (NHBT)*, a term also used in Hoijtink (2021), which was first introduced by Tendeiro & Kiers (2019).

The article is divided as follows. In the next section, the mathematical definition of the BF is briefly introduced, followed by the introduction of the *Approximated Adjusted Fractional Bayes Factor* (*AAFBF*, Gu, Mulder, & Hoijtink, 2018). Afterwards, the `R` package `bain` (Gu, Hoijtink, Mulder, & van Lissa, 2021) is presented as the basis for describing a `wrapper function`, specifically

[1] Hypotheses that impose equality and/or about-equality constraints between the parameters of a statistical model.

programmed for the aims of this study. Two additional issues that will be addressed in the successor paper, specific to multilevel models, are briefly mentioned in this section. In the section that follows afterwards, the prior sensitivity of the BF is illustrated using an openly available, two-level, data set and the aforementioned `wrapper function`. This article concludes with a discussion section, setting the stage for the work that will follow. All of the necessary code, to fully reproduce the results, can be obtained from the author's `GitHub` profile.[2]

The Bayes Factor

The *Bayes Factor* (Kass & Raftery, 1995) is defined as the ratio of two marginal likelihoods (see, Equation 1). Tendeiro & Kiers (2019) define the marginal likelihood as: "…weighted average of the likelihood over the observed data, where the weights are provided by the (within) priors."

$$BF_{0,1} = \frac{P(D|H_0)}{P(D|H_1)} = \frac{\int P(D|\theta_{H_0}, H_0)P(\theta|H_0)d\theta_{H_0}}{\int P(D|\theta_{H_1}, H_1)P(\theta|H_1)d\theta_{H_1}}. \tag{1}$$

The definition given in Equation 1 has two important aspects: (1) It defines the marginal likelihood as the denominator of Bayes' rule when used in the context of model (parameter) estimation. (2) It stresses the role of the prior distribution on the marginal likelihood and consequently on the value of the BF itself. This second aspect is the overall reason why the BF is sensitive to the specification of the prior distribution.

Furthermore, the BF can be seen as a multiplicative factor that transforms the prior odds of two hypotheses to the posterior odds, after seeing the data (Equation 2). However, if the prior odds of the hypotheses are set to equal one, by setting the prior probabilities of both hypotheses equal to each other, then the BF will be equal to the posterior odds (Kass & Raftery, 1995).[3]

$$\frac{P(H_0|D)}{P(H_1|D)} = BF_{0,1} * \frac{P(H_0)}{P(H_1)}. \tag{2}$$

[2] See, this link for direct access to the `GitHub` repository.

[3] Not to be confused with prior distributions used in model estimation, which are of main interest in this paper.

Straightforward calculation of the BF, based on its pure mathematical definition, presented in Equation 1, is impossible in most applied (multi-parameter) situations. However, when testing null (and informative) hypotheses, Equation 1 can be written as Equation 3. Thus, translating the BF into a so-called *Approximated Adjusted Fractional Bayes Factor*, which is defined as the ratio of the *fit* and *complexity* of a hypothesis (Gu, Mulder, & Hoijtink, 2018; Hoijtink, Gu, & Mulder, 2019; for a full proof of this BF see, Mulder, 2014).[4] When testing hypotheses containing equality constraints: *fit* (f_0) is the density of the *posterior* distribution supported by the hypothesis at hand; and *complexity* (c_0) is the density of the *prior* distribution supported by the hypothesis at hand (Hoijtink, Mulder, Lissa, & Gu, 2019).[5]

$$AAFBF_{0u} = \frac{f_0}{c_0} = \frac{\int_{\theta \in \Theta_0} \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}) d\theta}{\int_{\theta \in \Theta_0} \mathcal{N}(0, \hat{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}/b) d\theta}. \tag{3}$$

In Equation 3, $\hat{\boldsymbol{\theta}}$ represents a vector of estimated parameters and $\hat{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}$ represents its respective covariance matrix. It is important to stress that the AAFBF is calculated for each hypothesis against the unconstrained hypothesis or its complement (see, Hoijtink, Mulder, Lissa, & Gu, 2019). However, in the case of null hypotheses, these are the same. For example, a $BF_{0u} = 5$ would mean that the data is five times in favour of H_0 compared to the *unconstrained* hypothesis (see, Kass & Raftery (1995) or Hoijtink, Mulder, Lissa, & Gu (2019), for further guidelines).

This BF uses a fraction $b = \frac{J}{N}$ (see, Equation 3) of the information in the data to construct the scaling parameter of the prior distribution, where N represents the sample size and J usually denotes the number of independent constraints in the hypothesis. We will return to the values of J, N and subsequently b in the following sections, since they are directly linked to the sensitivity of this BF. It should be noted that in the denominator of Equation 3, the mean of the normal approximation of the prior distribution, $\mathcal{N}(0, \hat{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}/b)$, is zero *only* in the case when testing hypotheses where all parameters are equal to zero (i.e., $\theta_1 = \theta_2 = 0$). In a situation when testing whether the parameters are equal to each other (i.e., $\theta_1 = \theta_2$) the mean represents the mean value of the differences between the parameters, there are also many other situations (especially with regards to informative hypotheses), and the interested reader is referred to Gu, Mulder, & Hoijtink (2018).

[4] Throughout the remaining parts of this text, the terms BF and AAFBF are used interchangeably.

[5] When testing inequality constrained hypotheses instead of densities, the *fit* and *complexity* represent proportions.

Software

The `R` package `bain` (*Ba*yesian *in*formative hypothesis evaluation), computes the AAFBF of a hypothesis against its complement or the unconstrained hypothesis, using only the estimated parameters and their respective covariance matrix (Gu, Hoijtink, Mulder, & van Lissa, 2021).[6] A so-called `wrapper function`[7] was programmed specifically for the aims of this (and the following) paper, to conveniently use `bain` when testing hypotheses about the *fixed* parameters of *two-level models* built with the `lmer` function from the `R` package `lme4` (Bates, Mächler, Bolker, & Walker, 2015). The function takes an argument `fraction` which specifies the size of the minimal fraction b.

While programming this function, two questions specific to multilevel models, in addition to the sensitivity issue arose: (1) When testing equality constrained hypotheses of the form: $\theta_1 = \theta_2$ i.e., when comparing whether the parameters are equal to each other, it only makes sense to test the *standardized* coefficients. However, standardizing the (fixed) parameters of a multilevel model is not straightforward. First, we cannot use the standardized regression coefficients (like the $\beta's$ in multiple regression) directly, since there exists no way of obtaining their respective *standardized* covariance matrix. Secondly, when standardizing the data beforehand (which directly results in standardized coefficients), there remains an open question of whether to use, so-called, *overall standardization* or *within-group standardization* (see, Schuurman, Ferrer, Boer-Sonnenschein, & Hamaker, 2016). (2) Since the sample size (N) is required to derive the fraction b (Equation 3), it is unclear whether to use the level one observations, the level-two observations or something "in-between," like effective sample size (Hox, Moerbeek, & Van de Schoot, 2017, p. 5).

Since this paper only aims to *illustrate* the sensitivity of the AAFBF to the specification of the prior distribution, it was decided that overall standardization of the data and a sample size equal to the level two observations will be used to compute the BF's.

[6] Obtained by using standard statistical software packages, which usually apply some form of Maximum Likelihood Estimation.

[7] See, this link for a definition of `wrapper functions`.

Sensitivity Analysis

The `tutorial` data, which is openly available within the `R` package `R2MLwiN` (Zhang, Parker, Charlton, Leckie, & Browne, 2016), represents a subset derived from a larger data set of examination results from six London school boards. The data contains observations on 10 variables from 4059 students nested within 65 schools. In this paper, the variable *Exam score* serves as the outcome variable and the variables *LRT score* and *Average LRT* score are used as predictors, where the latter represents a level-two predictor. All of the variables were standardized by the authors of the data set (Table 1).[8]

Table 1: Descriptive Statistics

	M	SD	min	max
Exam score	0	1	-3.67	3.67
LRT score	0	1	-2.94	3.02
Avg. LRT score	0	0.31	-0.76	0.64

M = Mean
SD = Standard Deviation
min = minimum value
max = maximum value

A two-level model, with the `lmer` function from the package `lme4` using *Full Maximum Likelihood Estimation* (Bates, Mächler, Bolker, & Walker, 2015), is fitted to the data:

$$Exam\ score_{ij} = \gamma_{00} + \gamma_{10} LRT\ score_{ij} + \gamma_{01} AvgLRT_j + u_{1j} LRT\ score_{ij} + u_{0j} + e_{ij}, \tag{4}$$

with,

$$\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ e_{ij} \sim \mathcal{N}(0, \sigma_e^2).$$

Where,

$$\boldsymbol{U} = \{u_{1j}, u_{0j}\},\ \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{u1} & \sigma^2_{u1,u0} \\ \sigma^2_{u1,u0} & \sigma^2_{u0} \end{pmatrix}.$$

Here γ_{00}, γ_{10} and γ_{01} represent the fixed intercept, fixed slope and level 2 coefficient, respectively. The terms u_{1j}, u_{0j} represent the random effects for the slope and the intercept, respectively, having

[8] This would be the same as applying *overall standardization* which is currently available within the wrapper function.

a bivariate normal distribution with a mean vector ($\boldsymbol{\mu}$) and a covariance matrix ($\boldsymbol{\Sigma}$). Where σ^2_{u1} represents the variance of the slope, σ^2_{u0} represents the variance of the intercept, and $\sigma^2_{u1,u0}$ represents their covariance. Finally, e_{ij} is the residual term, normally distributed with a mean of zero and variance σ^2_e.

The results from fitting the model specified in Equation 4 are summarized in Table 2. The fixed coefficient for the first level predictor, *LRT score*, is estimated to be 0.55 and for *Average LRT score* the estimated, level-2 coefficient, is 0.29. Going back to Equation 3, this would mean that $\hat{\boldsymbol{\theta}}$ is a vector containing two parameter estimates.[9] These two parameters are used in constructing the hypotheses for the sensitivity analysis.

Table 2: Estimates from fitting the two-level model with ‘lmer‘

	Fixed effects		Random effects	
	est	SE	var	SD
γ_{00}	-0.001	0.036	0.074	0.273
γ_{10}	0.552	0.020	0.015	.122
γ_{01}	0.295	0.105	/	/

est = (Full) Maximum Likelihood Estimate
SE = Standard Error
var = Variance of the random effects
SD = Standard Deviation of the random effects (i.e., $\sqrt{var}$)

The *null* hypotheses tested to illustrate the sensitivity issue are specified as

$$H_{0_1} : \gamma_{10} = \gamma_{01} = 0 \; and \; H_{0_2} : \gamma_{10} = \gamma_{01}.$$

Additionally, one informative hypothesis of the form

$$H_i : \gamma_{10} > 0; \; \gamma_{01} > 0,$$

is added to the analysis to illustrate that *only* null hypotheses are sensitive to the specification of the prior distribution (Hoijtink, 2011).

First, using the wrapper function, we test all three hypotheses (at once) against the unconstrained hypothesis, applying $J = 2$, $N = 65$ (number of level 2 observations), which results in $b = 0.03$,

[9] The intercept and the random effects are treated as nuisance parameters.

setting the argument `fraction = 1`. Afterwards, the number in `fraction` is iteratively changed, taking on the values 2, 3 and 4, respectively. The results presented in Table 3 summarize the sensitivity issue in terms of the *complexity (c)* and $BF's_{.u}$, for the values of b.[10] First, for the minimal value of b, the resulting $BF_{0_1u} = 0$ indicates that the evidence in the data is completely in favour of the unconstrained hypothesis H_u, in other words, there is no evidence in the data for H_{0_1}. Furthermore, $BF_{0_2u} = 0.35$ indicates that the data is about 2.8 times in favour of H_u compared to H_{0_2}. Lastly, $BF_{iu} = 4.21$ suggests that the data are 4 times in favour of H_i against H_u. Moreover, it can be seen that the values for the complexity (see, Equation 3) for H_{0_1} and H_{0_2} increase as the value for b changes. This is not the case with H_i, where the complexity remains constant, clearly illustrating that only *null* hypotheses are sensitive to the specification of the fraction b. Consequently, only the resulting BF_{0_2u} changes from 0.35 when using 1 * b to 0.17 when using 4 * b. It should be noted that this is also the case with BF_{0_1u}, however, because the value of the *fit* for H_{0_1} happens to be a very small number there are not enough decimals to display this difference. The results of the sensitivity analysis are visually summarized in Figure 1, where, (a) BF_{0_2u} and (b) BF_{iu}, are plotted against the different values for b.

Table 3: Sensitivity analysis in terms of complexity and $BF_{.u}$ for H_{0_1}; H_{0_2} and H_i

	$1 * b$		$2 * b$		$3 * b$		$4 * b$	
H	c	$BF_{.u}$	c	$BF_{.u}$	c	$BF_{.u}$	c	$BF_{.u}$
H_{0_1}	2.31	0.00	4.61	0.00	6.92	0.00	9.23	0.00
H_{0_2}	0.64	0.35	0.91	0.25	1.11	0.20	1.28	0.17
H_i	0.23	4.21	0.23	4.21	0.23	4.21	0.23	4.21

c = complexity
$BF_{.u}$ = BF of the hypothesis at hand against the unconstrained hypothesis
$H_{0_1} : \gamma_{10} = \gamma_{01} = 0,\ H_{0_2} : \gamma_{10} = \gamma_{01},\ H_i : \gamma_{10} > 0; \gamma_{01} > 0$

Discussion

The sensitivity described in this paper highlights the instability of the AAFBF to the specification of the values for J and N and, consequently, to the value of b (since, $b = \frac{J}{N}$), in the context of two-level models. This has already been discussed for other statistical models (for example, Hoijtink, Mulder, Lissa, & Gu, 2019), and represents an undesirable characteristic of the BF.

[10] The values for the *fit* are not presented, since they do not depend on the prior (see, Equation 3).

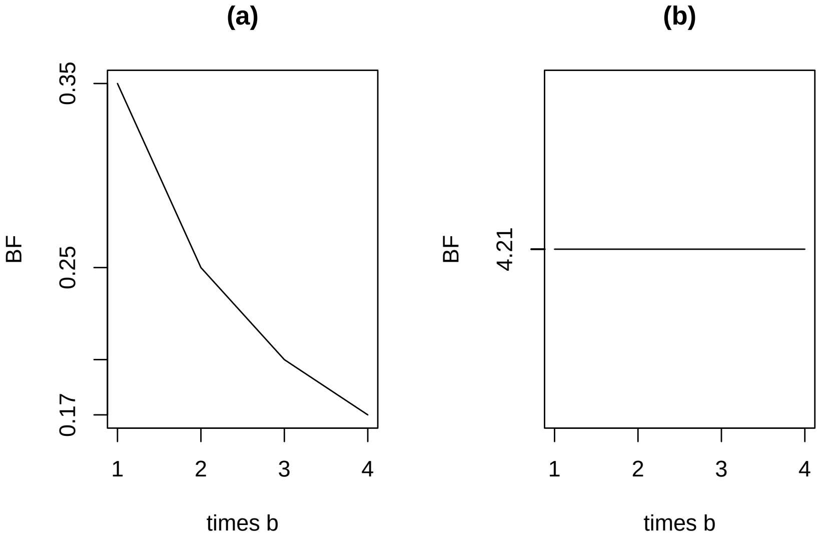

Figure 1: (a) $BF_{0_2 u}$ and (b) BF_{iu}, for different values of the fraction b (i.e., $1 * b, 2 * b, 3 * b$ and $4 * b$)

Hoijtink (2021) has addressed this issue in the context of Multiple Linear Regression, AN(C)OVA and the Welch test by choosing a so-called *reference* value for J which results in a BF with a completely specified prior distribution that doesn't suffer from the aforementioned sensitivity issue. This value is derived by setting the BF to be equal to 19 in favour of the null hypothesis when the effect size in the data is zero. The motivation behind the (subjective) choice of the number 19 is inspired by the fact that, when using equal prior model probabilities (see, Equation 2), the posterior model probabilities are $P(H_0|D) = .95$ and $P(H_1|D) = .05$ which, *numerically*, mimics the conventional $\alpha = .05$ in NHST.

In the successor paper: (1) Two different methods of standardization will be used and their impact on the resulting BF's will be discussed. (2) The question regarding the sample size will be addressed in detail, along with the introduction of a new method for calculating the *effective sample size* in two-level models containing random slopes. This method is inspired by the concept of *multiple imputation* of missing data (Van Buuren, 2018). By using Bayesian estimation of a two-level model, and treating the (posterior) estimates for the random and fixed effects from each i^{th} (MCMC) sampled vector as imputed missing values coming from an i^{th} imputed data set and applying the formulas given in Chapter 2.3 by Van Buuren (2018), the aim is to obtain an estimate for the effective sample size in models containing random slopes. (3) Most importantly, the approach given by Hoijtink (2021), will be implemented to calculate the new (*reference*) values for J and b using the sample size and the number of (fixed) coefficients. The derived *reference* value for b will be used to calculate BF's for *null* hypotheses with `bain`.

Even though, the AAFBF can straightforwardly be used when testing inequality constrained (informative) hypotheses, which in general provide a better description of relations between parameters, there still exists the need to use null hypotheses in some specific situations (see, for example, Wainer, 1999). Also, with NHBT, one can easily quantify the support in the data *in favour* of the null hypothesis (which is not possible with NHST). Thus, it is fair to argue that addressing the prior sensitivity of NHBT, for different statistical models, represents a worthwhile task for contemporary statisticians. Leaving this problem unsolved can easily open the gate to questionable research practices which would, undoubtedly, undermine the Bayesian approach to Hypothesis Testing.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997. doi: dx.doi.org/10.1037/0033-2909.112.1.155

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2021). *Bain: Bayes factors for informative hypotheses.* Retrieved from https://CRAN.R-project.org/package=bain

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. doi: 10.1111/bmsp.12110

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists.* CRC Press.

Hoijtink, H. (2021). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods.* doi: 10.1037/met0000292

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

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

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203–222. Cambridge University Press. doi: doi.org/10.1017/S030500410001330X

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

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

Schuurman, N. K., Ferrer, E., Boer-Sonnenschein, M. de, & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*, *21*(2), 206. doi: 10.1037/met0000062

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, *24*(6), 774. doi: 10.1037/met0000221

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/bf03194105

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

Zhang, Z., Parker, R. M. A., Charlton, C. M. J., Leckie, G., & Browne, W. J. (2016). R2MLwiN: A package to run MLwiN from within R. *Journal of Statistical Software*, *72*(10), 1–43. doi: 10.18637/jss.v072.i10