Robust Bayes Factor for Independent Two-Sample Comparisons under Imprecise Prior Information

Luisa Ebner L.T.EBNER@STUDENT.VU.NL
Master Student of Artificial Intelligence, Vrije Universiteit Amsterdam, Netherlands

Patrick Schwaferts PATRICK.SCHWAFERTS@STAT.UNI-MUENCHEN.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Institut für Statistik, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

Abstract

This paper proposes the robust Bayes Factor as a direct generalization of the conventional Bayes Factor for a special case of independent two-sample comparisons. Such comparisons are of great importance in psychological research, and more generally wherever the scientific endeavour is to ascertain a potential group effect. The conventional Bayes Factor as the ratio of the marginal likelihoods under two considered hypotheses demands for a precise, subjective specification of the prior distribution for the parameter of interest. Thus, it lacks the possibility of incorporating prior knowledge that is only available partially. Drawing on the theory of Imprecise Probabilities, the *robust* Bayes Factor is presented in view of lifting the restrictions on the specification of the prior distribution as being precise. In practice, the robust Bayes Factor approach enables an analyst to specify hyperparameter *intervals*, whose lengths correspond to the degree of subjective prior uncertainty. Based thereon, a set of (infinitely) many subjective prior distributions is established to substitute one precise prior distribution. Finally, the robust Bayes Factor is defined as an interval, bounded by the minimal and the maximal resultant Bayes Factor values. Latter are obtained by optimizing the conventional Bayes Factor over the predefined set of prior distributions. This explicit incorporation of incomplete prior knowledge increases the feasibility of applying a Bayesian approach to hypothesis comparisons in scientific practice. It reduces error-proneness, enables for an inclusion of multiple perspectives and encourages cautious, more realistic conclusions in hypothesis comparisons.

Keywords: Bayes Factor, Imprecise Probabilities, Robustness, Bayesian Statistics, Prior Specification, Psychological Research, Two-Sample Comparison

1. Introduction

The evaluation of statistical hypotheses is among the main targets of applied sciences, especially in psychological research (see e.g. Liu and Aitkin, 2008). Although being analyzed frequentistically in the past by means of classic hypothesis tests, a Bayesian approach to compare hypotheses is gaining popularity (Van De Schoot et al., 2017). In that, the so called Bayes Factor (BF) is a key quantity for assessing the evidence within the data w.r.t. statistical hypotheses (see e.g. Gönen et al., 2005; Rouder et al., 2009), whose recent developments are located within the field of psychological research, such that a similar perspective is adopted within this paper. A crucial difference between the frequentist and the Bayesian approach is the presence of subjective prior distributions in latter, which on the one hand allows including prior knowledge into the statistical analysis, but on the other hand yields results - especially the Bayes Factor - that might be influenced strongly by the exact specification of the prior distribution, leading to heavy debates about how to specify these priors (see e.g. the debate about extrasensory perception between Bem et al. (2011) and Wagenmakers et al. (2011)).

Conventionally, a Bayesian analysis requires the prior distribution to be precise: There should be a single probability distribution describing the prior knowledge. Yet, this is a very strong requirement as, within a Bayes Factor analysis, the prior distribution formalizes the available knowledge or beliefs about the parameter prior to the scientific investigation, which might be accessible to the applied scientist only vaguely (see e.g. Joyce, 2010; Goldstein, 2006). Furthermore, requiring the researcher to specify a precise and unambiguous probability distribution to represent the available knowledge might be regarded as impossible in a real-world situation. This might be easily realized as the plethora of different "non-informative" priors (found in almost all introductory text books about Bayesian statistics) indicates that there is no agreement on how to formalize non-knowledge even in the simplest contexts. Accordingly, mis-specifying a precise prior distribution might seem unavoidable within an applied Bayes Factor analysis and results might be misleading. A conventional way to cope with this issue is a sensitivity analysis (see e.g. Ríos Insua and Ruggeri, 2012), which assesses how a change in prior distribution would have changed the result. However, the researcher still needs to decide on a certain precise distribution to use, which might be arbitrary, as many pre-

cise prior distributions might be in accordance with the (vaguely) available prior knowledge. In that sense, the most reasonable solution is to use all these reasonable prior distributions in the Bayes Factor analysis, which shall be referred to as robust Bayes Factor (rBF) analysis, leading to a more robust and less arbitrary result.

The purpose of this paper is to formally describe the robust Bayes Factor in the context of two independent normally distributed samples with identical variance, which is a commonly employed scenario within psychological research, e.g. to assess gender differences. Therefore, a conventional Bayes Factor analysis for this setting shall be outlined in Section 2 first and its generalization to include sets of prior distributions instead of a single precise prior distribution follows in Section 3.1, concluded by an example (Section 3.2) and a short discussion (Section 4).

2. Bayes Factor

The experimental setup leading to the calculation of this particular Bayes Factor may accord to that of a classical two-sample t-test, whose basic endeavour is to examine a potential group difference. Accordingly, observed data $z := (x, y)$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, x_m)$ may be realizations of independent, normally distributed random variables X_i and Y_j, i.e.

$$X_i \overset{iid.}{\sim} N(\mu, \sigma^2) \quad i \in 1, ..., n \tag{1}$$

$$Y_j \overset{iid.}{\sim} N(\mu + \alpha, \sigma^2) \quad j \in 1, ..., m. \tag{2}$$

Here, μ is the unknown mean of the first sample, σ^2 the unknown variance within each sample and α describes the difference in means between both groups, which may be referred to as the total effect (see e.g. Rouder et al., 2009).

For the purpose of consistent scalability across different scientific contexts and as commonly done in psychological research, latter shall be reparameterized as standardized effect size

$$\delta := \frac{\alpha}{\sigma}. \tag{3}$$

Accordingly, the parameters δ and σ^2 are not independent of each other.

As δ explicitly represents the group difference of interest, the hypothesis set may be outlined conventionally as

$$H_0 : \delta = 0 \quad vs. \quad H_1 : \delta \neq 0. \tag{4}$$

Whereas the null hypothesis H_0 implies strict group mean equality, the alternative H_1 assumes a group effect of yet unspecific extent. The corresponding Bayesian approach is to compare H_0 and H_1 by means of the Bayes Factor as a measure of how well the hypotheses under consideration predict observed sample data relatively.

Naturally, employing a simple null hypothesis, which hypothesizes only one single δ value, is subject to heavy critique (see e.g. Cohen, 1994). A recently promoted Bayesian alternative is to consider a region of practical equivalence (ROPE) around $\delta = 0$ (see e.g. Kruschke, 2018). This, however, was mainly developed using Bayesian estimation rather than Bayesian hypothesis comparison (see e.g. Kruschke, 2015, Chapter 12), yet a few approaches to incorporate these considerations into Bayes Factor analyses do exists (see e.g. Morey and Rouder, 2011). Nevertheless, a simple null hypothesis was chosen within this paper to build on the existing literature about Bayes Factors (see e.g. Gönen et al., 2005; Rouder et al., 2009).

The calculation of the Bayes Factor is based on the idea that the support for a scientific hypothesis depends on how its marginal likelihood matches with an observed sample in comparison to that of the other hypothesis under consideration (see e.g. Morey et al., 2016). As to that, any Bayes Factor calculation presumes the specification of a marginal likelihood under either hypothesis.

Due to the precise assignment of δ under H_0, the corresponding likelihood function is defined as $f(z|\mu, \sigma^2, \delta = 0)$. As μ and σ^2 depict unknown parameters, prior densities $\pi(\mu)$ and $\pi(\sigma^2)$ need to be specified in line with the Bayesian parameter conception. Finally, this yields

$$m_0(z) = \iint f(z|\mu, \sigma^2, \delta = 0)\pi(\sigma^2)\pi(\mu)\, d\mu\, d\sigma^2 \tag{5}$$

as the marginal likelihood under H_0.

In the case of H_1, however, the unspecific claim that δ holds any other value but 0 still leaves δ an unknown parameter. Therefore, not only μ and σ^2, but also δ needs to be given a prior distribution under H_1 to obtain the posterior likelihood function. Due to its dependence on σ^2 the prior on δ is conditional and denoted as $\pi(\delta|\sigma^2)$. It assigns varying probability mass to a range of potential δ values in accordance with their plausibility under H_1. This modification transforms H_1 from a general into a specific hypothesis and yields the corresponding *Bayesian* hypotheses set (see e.g. Gönen et al., 2005; Rouder et al., 2009) as

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta|\sigma^2 \sim \pi(\delta|\sigma^2). \tag{6}$$

Finally, the marginal likelihood under H_1 ensues as

$$m_1(z) = \iiint f(z|\mu, \sigma^2, \delta)\pi(\delta|\sigma^2)\pi(\sigma^2)\pi(\mu) d\mu\, d\sigma^2 d\delta. \tag{7}$$

The priors $\pi(\mu)$, $\pi(\sigma^2)$ and $\pi(\delta|\sigma^2)$ need to be specified by the respective analyst according to her/his prior information and beliefs. As stated above, $\pi(\mu)$ and $\pi(\sigma^2)$ enter the posterior likelihood functions under *both* hypotheses. It is argued that this common occurrence largely cancels their effects on the result of a hypothesis comparison (see e.g. Wagenmakers et al., 2010). As to that, μ and σ^2

may be referred to as *common* or *nuisance* parameters. According to an initial proposal by Jeffreys (Jeffreys, 1961), they shall herein be assigned the improper priors (see e.g. Wang and Liu, 2016; Gönen et al., 2005)

$$\pi(\mu) \propto c \quad \text{and} \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2}, \tag{8}$$

where $c > 0$ is a constant value. The prior on μ states that all potential values have equal credibility. The prior on σ^2 states that larger values are less credible than smaller ones and variance values very close to 0 have the highest credibility. This, however, might be questioned in real-world applications so that informative priors for μ and σ^2 might be employed. Yet, for the context within this paper, the choice of nuisance prior distribution does not affect the Bayes Factor value (Wagenmakers et al., 2010). Accordingly, σ^2 might be treated as nuisance parameter, despite δ being dependent on it.

The specification of the prior on δ on the other hand is given an emphasized position within this evaluation process. As it will later on enter the Bayes Factor only through the marginal likelihood under H_1, it considerably affects on its outcome. Thus, $\pi(\delta|\sigma^2)$ may be stated the (only) test-relevant prior (Ly et al., 2016). The choice of a normal distribution for the effect size prior is chiefly promoted in psychological research (see e.g. Berger and Sellke, 1987; Gönen et al., 2005; Rouder et al., 2018), as its shape is most often reasonable to describe prior assumptions regarding an yet unknown effect size. After all, probability mass is hereby spread symmetrically around a certain mean μ_δ that is deemed plausible and this probability mass declines as the distance to the mean increases (see e.g. Rouder et al., 2009; Matthews, 2011). This facilitates reasonable hyperparameter choices and in turn an alternative hypothesis that might have a reasonable counterpart in the real-world. Accordingly, a normal distribution, with parameters independent of σ^2, is chosen within this paper to represent prior knowledge about the value of δ:

$$\delta|\sigma^2 \sim N(\mu_\delta, \sigma_\delta^2). \tag{9}$$

In that, μ_δ and σ_δ^2 are the only hyperparameters to be chosen subjectively by the respective analyst (see e.g. Berger and Sellke, 1987).

Finally, based on equations (7) and (5) the Bayes Factor is commonly defined as the ratio

$$BF = \frac{m_1(z)}{m_0(z)}. \tag{10}$$

The numerator measures the marginal likelihood of z under the assumption of a $\pi(\delta|\sigma^2)$ - distributed effect size. The denominator depicts the counterpart under the assumption of equal group means. As such, the above stated Bayes Factor is typically interpreted as quantifying the statistical evidence the data z hold for the presence of a $\pi(\delta|\sigma^2)$-distributed effect size in comparison to an absence of an effect. Therefore, BF values larger than 1 favor H_1 and BF values smaller than 1 favor H_0.

For precisely the above stated case, Gönen et al. (2005) reported a closed-form implementation, which allows a Bayes Factor formula that is solely dependent on the pooled-variance two-sample t-statistic t under H_0 and H_1, each. Its concrete implementation applies as

$$BF = \frac{T_\nu(t\,|\,n_\delta^{1/2}\mu_\delta, 1+n_\delta\sigma_\delta^2)}{T_\nu(t\,|\,0,1)}, \tag{11}$$

where $T_\nu(\cdot|a,b)$ is the probability density function of the non-central t-distribution with location a, scale $\sqrt{b}$ and $\nu = n+m-2$ degrees of freedom. Eventually,

$$n_\delta = \left(\frac{1}{n}+\frac{1}{m}\right)^{-1} \tag{12}$$

is typically termed the effective sample size.

In addition to specifying the test-relevant prior $\pi(\delta|\sigma^2)$, a Bayes Factor analysis in a broader sense requires the specification of prior probabilities of the hypotheses themselves: $P(H_1)$ and $P(H_0) = 1 - P(H_1)$. The Bayes Factor value BF is used to update these beliefs in the hypotheses, resulting in the posterior odds

$$\frac{P(H_1|z)}{P(H_0|z)} = BF \cdot \frac{P(H_1)}{P(H_0)}, \tag{13}$$

stating how strongly H_1 is preferred over H_0 after seeing the data z.

Certainly, the prior situation consists of treating both the hypotheses and the parameters as random variables with probability distributions, allowing for Bayesian hierarchical modeling (see e.g. Gelman et al., 2013; Rouder et al., 2018).

In summary, it can be stated that this special case Bayes Factor for independent two-sample comparisons depends on observed data only through their corresponding t-statistic and on (subjective) prior knowledge in terms of the hyperparameters μ_δ and σ_δ^2. This enables for a facile calculation and standardized software implementations – pleasant features that are otherwise unusual in the context of Bayesian analyses. Among others, this granted the Bayes Factor quite some popularity not only in psychological research, as mentioned in the introduction, but also in a number of other research domains (see e.g. Rouder et al., 2018; Van De Schoot et al., 2017). Among its preferable properties are the possibility to include data-external information, its interpretation as evidence statement and its foundation following the likelihood principle (Berger and Wolpert, 1988) as well as the law of likelihood (Hacking, 1965). In line with latter, the analysis is conditional on the data and therefore sequential experimental designs are argued to be no problem (Rouder, 2014), which allow increasing the sample size if the evidence within the data is not sufficient enough (see e.g. Schönbrodt et al., 2017).

The basic cause, for which the Bayes Factor is groundedly criticized and backed away from, is mostly down to the strict demand for a precise, test-relevant prior $\pi(\delta|\sigma^2)$. Finally, this is the motivation for a generalizing robust Bayes Factor, dedicated to loosen the Bayes Factors' flawed demand for prior precision.

3. Robust Bayes Factor

3.1. Theory

As outlined in the previous section, a common approach to a Bayes Factor analysis is to assume a normal prior for δ (see e.g. Berger and Sellke, 1987; Gönen et al., 2005; Rouder et al., 2018). Accordingly, a first attempt to generalize the Bayes Factor to allow sets of prior distributions is by considering a set of normal distributions. In that, all normal distributions with parameter values

$$\mu_\delta \in [\underline{\mu_\delta}, \overline{\mu_\delta}] \tag{14}$$

$$\sigma_\delta^2 \in [\underline{\sigma_\delta^2}, \overline{\sigma_\delta^2}] \tag{15}$$

shall be considered, where the intervals specify the parameter values that are considered as being in accordance with the (potentially vague) prior knowledge about the respective parameter values, given the alternative hypothesis H_1 is true and this prior knowledge is truly expressible as normal distribution. Therefore, in consequent generalization of equation (9), the set

$$\mathscr{M} := \{N(\mu_\delta, \sigma_\delta^2) | \mu_\delta \in [\underline{\mu_\delta}, \overline{\mu_\delta}], \sigma_\delta^2 \in [\underline{\sigma_\delta^2}, \overline{\sigma_\delta^2}]\} \tag{16}$$

represents the test-relevant prior, such that the hypotheses might be formulated as

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta|\sigma^2 \sim \mathscr{M}\,, \tag{17}$$

with priors for the nuisance parameters as in equation (8). Within this formulation, "$\delta|\sigma^2 \sim \mathscr{M}$" is analogue to the alternative hypothesis in equation (6), in which a distribution of $\delta|\sigma^2$ is provided. Within the framework of the robust Bayes factor, however, the set $\mathscr{M}$ of prior distributions is employed instead of a single prior distribution[1]. Therefore, the alternative hypothesis states that δ is distributed in accordance with the (vaguely available) knowledge about δ, mathematically expressed by the set $\mathscr{M}$. This set – or its convex hull – shall be considered as an entity of its own (c.p. Walley, 1991). Accordingly, the alternative hypothesis H_1 is allowed to contain all available information without being overly precise.

For every precise distribution within $\mathscr{M}$, it is possible to calculate the corresponding precise Bayes Factor, leading to a range of different Bayes Factor values, which shall be referred to as robust Bayes Factor

$$rBF = [\underline{BF}, \overline{BF}]\,, \tag{18}$$

where

$$\underline{BF} = \min_{\substack{\mu_\delta \in [\underline{\mu_\delta}, \overline{\mu_\delta}] \\ \sigma_\delta^2 \in [\underline{\sigma_\delta^2}, \overline{\sigma_\delta^2}]}} BF \tag{19}$$

$$\overline{BF} = \max_{\substack{\mu_\delta \in [\underline{\mu_\delta}, \overline{\mu_\delta}] \\ \sigma_\delta^2 \in [\underline{\sigma_\delta^2}, \overline{\sigma_\delta^2}]}} BF\,. \tag{20}$$

Analogue to the precise case, prior probabilities of the hypotheses ($P(H_1)$ and $P(H_0)$) might be updated by the robust Bayes Factor, leading to a range of posterior odds

$$\left[\underline{BF} \cdot \frac{P(H_1)}{P(H_0)}, \overline{BF} \cdot \frac{P(H_1)}{P(H_0)}\right]. \tag{21}$$

Although not addressed within this paper, it might be possible to also specify the prior probabilities of the hypotheses interval-valued (c.p. Schwaferts and Augustin, 2019).

In this case of a normal test-relevant prior, the robust Bayes Factor and the corresponding posterior odds are intervals, as the Bayes Factor is continuous in the parameters μ_δ and σ_δ^2. As illustrated within the following example, this allows the interpretation of the resulting robust Bayes Factor to be straight forward.

3.2. Example

A fictitious example with simulated data (reproducible with the R code in the electronic appendix) shall be given to illustrate the methodology of the robust Bayes Factor, which is based on a study by van Loo et al. (2017). The occurrence of major depression (MD) is about twice as high in women than in men, however, once diagnosed potential gender differences are less investigated. In that, it might be assessed, if there is a gender difference in the recurrence of MD, as some previous studies reported similar recurrence rates and others reported higher recurrence rates for women than for men (a summary of these studies is found in van Loo et al., 2017). The risk of recurrence might be captured by a score, which can calculated by a number of different risk predictors (see van Loo et al., 2017). Within this example, it is simply assumed that the score might be modeled by a normal distribution and that both women (Y) and men (X) have an equal variance in score values (as in equations (1) and (2)).

With Jeffreys priors for the nuisance parameters (see equation (8)) and the standardized difference in score means δ being hypothesized to be 0 (H_0) or normally distributed $N(\mu_\delta, \sigma_\delta^2)$ conditional on σ^2 (H_1), the fictitious research group is unable to precisely specify the test-relevant prior due to a lack of overly excessive information and

1. Technically, one could also argue that the convex hull of $\mathscr{M}$ can be considered.

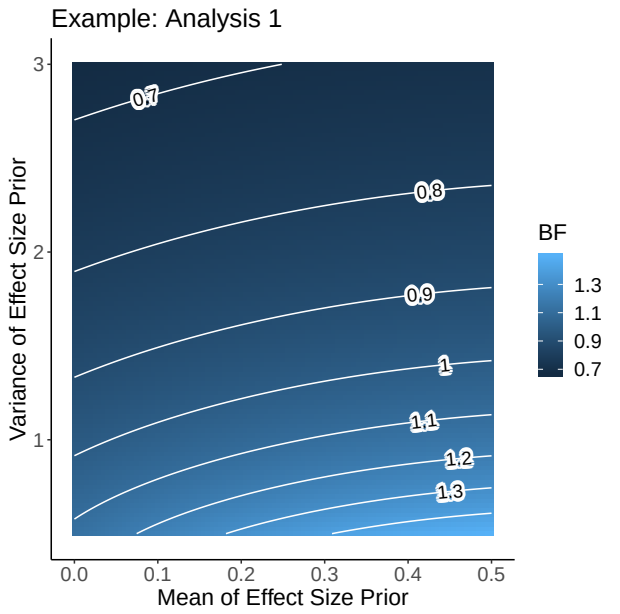

Figure 1: Dependence of the Bayes Factor value (color) on the mean μ_δ (x-axis) and variance σ_δ^2 (y-axis) of the normal effect size prior within the first exemplary analysis.

therefore employs the hypotheses as in equation (17). In accordance with the previous studies, if there is a gender effect (H_1), δ might be positive but rather small. In that, the research team figures out that normal prior distributions for δ might be plausible with a mean parameter μ_δ ranging from 0 to 0.5 and with a variance parameter σ_δ^2 being within the interval $[0.5, 3]$, leading to

$$\mathscr{M} = \{N(\mu_\delta, \sigma_\delta^2) | \mu_\delta \in [0, 0.5], \sigma_\delta^2 \in [0.5, 3]\}. \quad (22)$$

Note, that these considerations need to be based on previous knowledge, which might be available more profoundly in a real-world investigation (as it is the scientist performing the investigation, who knows most about the effect of interest) than in this simple example.

The research group now assess the recurrence rate scores x and y of $n = 10$ men and $m = 10$ women, respectively, which yield $t = 1.46$, $n_\delta = 5$ and accordingly

$$rBF = [0.67, 1.50]. \quad (23)$$

Figure 1 illustrates the dependence of the Bayes Factor value on the hyperparameters μ_δ and σ_δ^2.

Due to the disagreement within the previous studies, the research team did not prefer any hypothesis over the other, prior to the investigation, so they set $P(H_1) = P(H_0) = 0.5$ as prior probabilities of the hypotheses, leading to posterior odds with the same range (equation (23)).

Therefore, the data z favor H_1 0.67 to 1.5 times as much as H_0 and there is no unambiguous evidence for either hypothesis, because rBF contains both values larger and smaller than 1. Analogously, expressed by the posterior odds, the research team cannot believe in one hypothesis more strongly than in the other. However, if the test-relevant prior would have been specified precisely, there might have been a single Bayes Factor value that might have favored one of the hypotheses, but this conclusion would have been arbitrary and therefore potentially misleading. In that, given that the available prior information is only imprecisely available within this example, the data is inconclusive about the hypotheses, so the research team can neither state that recurrence rates are similar for both women and men nor that they are larger for women than for men.

In order to obtain more evidence, the research team assess another 20 women and 20 men, so that $n = m = 30$. The new results are

$$rBF = [0.18, 0.42] \quad (24)$$

with $t = 0.65$ and $n_\delta = 15$. Now, the data might be interpreted as favoring the null hypothesis H_0 $1/0.42 = 2.4$ to $1/0.18 = 5.5$ as much as the alternative hypothesis H_1, being not inconclusive anymore. Analogue, Figure 2 illustrates the dependence of the Bayes Factor value on the hyperparameters μ_δ and σ_δ^2. The data might be treated as (slightly) favoring the hypothesis of similar recurrence rates between women and men and, based on the prior probabilities of the hypotheses, the research team believes into H_0 2.4 to 5.5 times as much as into H_1.

As illustrated by this example, the imprecision of prior information leads to an inconclusive, but robust and less arbitrary result that indicates a lack of information even after collecting the first data set, which might have been masked by pretending an arbitrary precision and is tackled appropriately by collecting more data.

4. Discussion

This paper depicts the robust Bayes Factor both as a generalization of the conventional Bayes Factor and also as a possibility to tackle one of the main criticisms against the Bayes Factor, namely the arbitrariness of specifying a precise prior distribution. Clearly, this asks for a discussion of rBF's effective advantages in scientific practice.

Put simply, the robust Bayes Factor generalizes the classical Bayes Factor in a way to render it more compatible with scientific reality. It faces up to the fact, that numerically precise credences are hardly ever attainable in practice and precise prior choices can thus be alleged arbitrariness or unjustified make-belief of precision (see e.g. Goldstein, 2006; Kass and Raftery, 1995). Following a truly intuitive generalization principle, the robust Bayes Factor is constructed to

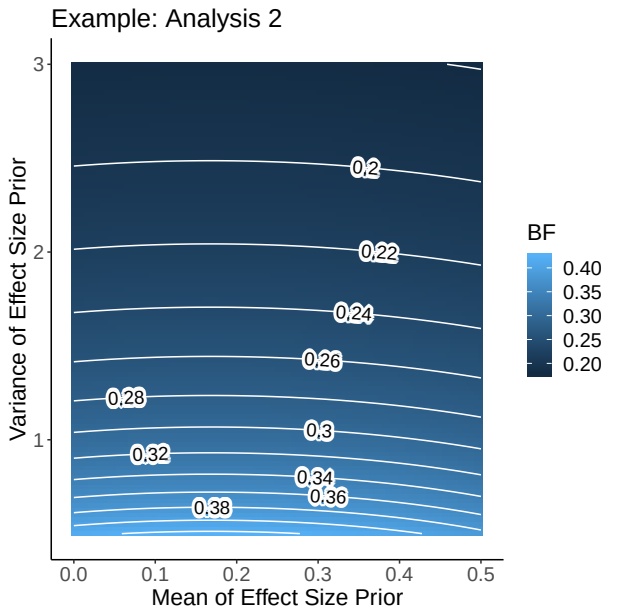

Figure 2: Dependence of the Bayes Factor value (color) on the mean μ_δ (x-axis) and variance σ^2_δ (y-axis) of the normal effect size prior within the second exemplary analysis.

provide reliable results also in situations where prior knowledge is partial: If one is unable to specify precise parameter values in accordance with their prior knowledge, one might still be able to locate parameters in value *ranges* and thus specify intervals, which allow to represent the available uncertainty in a more comprehensive way. At the same time, the robust Bayes Factor approach upholds the notion that subjective prior knowledge is a gain to statistical analyses (compare e.g. Gelman et al., 2017; Matthews, 2011; Vanpaemel, 2010; Kass and Raftery, 1995). In that, it prompts the respective researcher to reason about suitable parameter values and claim choices on parameter bounds, such that the interval length reflects, but not exceeds, the actual amount of uncertainty. In addition, prior assumptions are laid out transparently through the set of prior distributions. Furthermore, the robust Bayes Factor approach may be approved for encouraging scientific consensus by enabling multiple prior perspectives on the parameter of interest to be merged into the set of prior distributions. The resulting robust Bayes Factor might then yield greater acceptance in the face of prior disagreement on a single precise prior distribution (see e.g. Berger, 1990). One may even state that the *rBF* result provides an analyst with an extended overall impression of comparative evidence. Based on the resulting interval length, (s)he may reflect about the Bayes Factors overall robustness against differing hyperparameter assumptions or individual uncertainty. As the resulting *rBF* interval is considered and interpreted as an entity of its own, cautious and solid conclusions are encouraged. The demand for any evidence statement to be expressed with reference to inherent prior imprecision, makes conclusions less over-precise and withal more honest (see e.g. Augustin et al., 2014).

Of course, the robust Bayes Factor approach has its limitations. For the certain context employed within this paper, the resulting robust Bayes Factor is a convex interval of values. This, however, is not given in general and in certain situations the robust Bayes Factor might only be a non-convex set of values rather than an interval, which bears difficulties for its interpretation. Assume a robust Bayes Factor set contains two values, e.g. 3.0 and 3.2, but not those values in between. The correct interpretation would be that the data are evidence favoring H_1 3.0 or 3.2 times as much, but not e.g. 3.1 times as much, as H_0. More research is necessary on how to deal with this issue.

It may also be countered that the strengths of the robust Bayes Factor approach are at cost of more vague statements of comparative evidence. The expressiveness and clarity of conclusions implies reasonably narrow *rBF* intervals and if the *rFB* bounds are not either both above or below 1, comparative evidence remains somewhat ambiguous, as in the first part of the example (Section 3.2). If the specified prior intervals of the hyperparameters are too broad to yield conclusive results, one could either try to narrow them by collecting additional information prior to the experiment or collect additional data, as illustrated within the second part of the example (Section 3.2). Finally, if neither is possible, Berger (1990, p. 307) reasons that

> "[...] then there are legitimate differences or uncertainties in opinion which lead to different conclusions, and it seems wisest just to conclude that there is no answer; more evidence is needed to solve the ambiguity. Any 'alternative' [approach] which claims to do more, would simply be masking legitimate uncertainty by 'sweeping it under the carpet'. "

5. Outlook

The robust Bayes Factor was described for a first context of two independent normally distributed samples with an imprecise normal effect size prior within this paper. Besides employing it within an applied scientific investigation, its further development might comprise two different steps. First, the robust Bayes Factor might be extended to different experimental setups, such as those that assess correlations or dependent variables within more than two groups. Second, the restriction of the prior distributions being normal within the prior set of distributions might be removed to allow all desired shapes of prior distributions. Latter, however, might require a solution to interpreting non-convex

sets of Bayes Factor values and advanced computational methods to calculate respective Bayes Factor values, which could be avoided within this paper due to the availability of close form formulas.

Appendix A. R Code

R code to replicate the example and generate Figures 1 and 2 is provided electronically.

Acknowledgments

We want thank all reviewers for their valuable comments and their open sharing of thoughts. In addition, PS wants to thank the LMU Mentoring Program, which supports young researchers.

References

Thomas Augustin, Gero Walter, and Frank P.A. Coolen. Statistical inference. In Thomas Augustin, Frank P.A. Coolen, Gert de Cooman, and Matthias C.M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. John Wiley & Sons, 2014.

Daryl J. Bem, Jessica Utts, and Wesley O. Johnson. Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101(4): 716–719, 2011.

James O. Berger. Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328, 1990.

James O. Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82:112–122, 1987.

James O. Berger and Robert L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA, second edition, 1988.

Jacob Cohen. The earth is round (p<. 05). *American Psychologist*, 49:997–1003, 12 1994.

Andrew Gelman, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10), 2017.

Michael Goldstein. Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1:403–420, 2006.

Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall. The Bayesian two-sample t test. *The American Statistician*, 59:252–257, 2005.

Ian Hacking. *Logic of statistical inference*. Cambridge University Press, 1965.

Harold Jeffreys. *Theory of Probability*. Oxford, Oxford, England, third edition, 1961.

James M. Joyce. A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, 24:281–323, 2010.

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

John K. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2015.

John K. Kruschke. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

Charles C. Liu and Murray Aitkin. Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52:362–375, 2008.

Alexander Ly, Josine Verhagen, and Eric-Jan Wagenmakers. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32, 2016.

William J. Matthews. What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? *Judgment & Decision Making*, 6: 843–856, 2011.

Richard D. Morey and Jeffrey N. Rouder. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16:406–19, 07 2011.

Richard D. Morey, Jan-Willem Romeijn, and Jeffrey N. Rouder. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18, 2016.

David Ríos Insua and Fabrizio Ruggeri, editors. *Robust Bayesian Analysis*. Springer Science & Business Media, 2012.

Jeffrey N. Rouder. Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2):301–308, 2014.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16:225–237, 2009.

Jeffrey N. Rouder, Julia M. Haaf, and Joachim Vandekerckhove. Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25:102–113, 2018.

Felix D. Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2):322, 2017.

Patrick Schwaferts and Thomas Augustin. Imprecise hypothesis-based Bayesian decision making with simple hypotheses. Conditionally accepted subject to minor revision for: Jasper de Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler, editors, *Proceedings of the 11th International Symposium on Imprecise Probability: Theory and Applications (ISIPTA '19, Ghent), Proceedings in Machine Learning Research*, 2019.

Rens Van De Schoot, Sonja D. Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, and Sarah Depaoli. A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22:217–239, 2017.

Hanna M. van Loo, Steven H. Aggen, Charles O. Gardner, and Kenneth S. Kendler. Sex similarities and differences in risk factors for recurrence of major depression. *Psychological Medicine*, 48:1685–1693, 2017.

Wolf Vanpaemel. Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54:491–498, 2010.

Eric-Jan Wagenmakers, Tom Lodewyckx, Himanshu Kuriyal, and Raoul Grasman. Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, 60(3):158–189, 2010.

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L.J. Van Der Maas. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3):426–432, 2011.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

Min Wang and Guangying Liu. A simple two-sample Bayesian t-test for hypothesis testing. *The American Statistician*, 70:195–201, 2016.