Bachelor Thesis

Department of Statistics

Ludwig-Maximilians-Universität Munich

The Bayes Factor for independent two-sample comparisons in psychological research:

A discussion and generalization in the context of Imprecise Probabilities

Luisa Ebner

supervised by Patrick Schwaferts

Munich, August 2018

Contents

Abstract

The Bayes Factor is the Bayesian tool for hypothesis comparison. As the ratio of two marginal likelihoods, it quantifies statistical evidence in favor of one hypothesis over another. Thereby, it is sensitive to the prior distributions of unknown parameters, contained in the respective likelihood functions.

In recent years, scientists in psychological research promoted the Bayes Factor as an alternative to the frequentist two sample t-test, in turn being a major case of application in that domain.

The major target of this Bachelor thesis is to generalise the Bayes Factor for independent two-sample comparisons (BF) in the context of Imprecise Probabilities. The ensuing *Imprecise Bayes Factor* (IBF) is then promoted as an enhancement of BF in situations, where subjective prior knowledge is insufficient to meet the demand for a precise specification of the normally distributed effect size prior representing the alternative hypothesis.

The thesis dialectically discusses the BF in the context of the effect size prior, argumented to be BF's only test relevant prior. The demand for its precise specification is herein revealed as BF's predominant deficiency.

The IBF approach counteracts this shortcoming through an explicit consideration and modeling of partial prior knowledge. Drawing on the theory of Imprecise Probabilities, the prior's hyperparameters are subjectively specified as intervals. Based thereon, a credal set is established to substitute one precise prior distribtion by a set of infinitely many, potential prior distributions. Finally, the IBF is defined as an interval, bounded by the minimal and the maximal resultant BF value. The latter are accomplished over an optimisation of the conventional BF calculation in conformity with the predefined credal set.

The IBF approach increases the feasibility of BF calculations in scientific practice. It reduces error-proneness, enables for an inclusion of multiple perspectives and encourages cautious, more realistic conclusions. Furthermore, it is likelier to contain the prior distribution, that matches the real word situation.

To sum it: The IBF states a beneficial alternative to the BF in any situation, where prior knowledge does not allow for a precise specification of the effect size prior and an interval is considered a satifactory result.

1 Introduction

The evaluation of formal, specialized hypotheses is among the principal targets of applied sciences, particularly including psychological research. [Liu and Aitkin, 2008, p. 363]
Scientific hypotheses may thereby be conceived as tentative predictions of cause and effect in the context of a certain research question. As such, hypotheses signify a researcher's set of considerations, expectations and beliefs prior to a scientific analysis. [Rouder et al., 2009, p. 229] In order to gain subject-related relevance, hypotheses need to pass a formal, scientifically approved evaluation within the context of statistical inference. [see Augustin et al., 2014, p. 136]

This Bachelor thesis adresses the Bayes Factor as the tool for hypothesis comparison within the framework of Bayesian inference. The thesis focuses on independent two-sample comparisons as a standard, statistical problem in psychological research. In this context, the Bayes Factor serves to quantify statistical evidence in the comparison of a point null hypothesis - representing the assumption of equal group means - and a respective, composite alternative.
Despite increasing popularity, the Bayes Factor resulting from the so-called *Bayesian two-sample t test* [Gönen et al., 2005, p. 252] comprises considerable shortcomings regarding prior sensitivity and the handling of partial prior knowledge. [Morey et al., 2016, p. 15] Finally, the Imprecise Bayes Factor is proposed as generalisation of the BF in the context of Imprecise Probabilities and promoted as an enhancement to the conventional approach.

The thesis starts off, embedding the Bayes Factor into the major principles of Bayesian inference. After that, Harold Jeffrey's Bayes Factor is presented on a general account. Thereby, a step-by-step description of the approach, that yields the Bayes Factor as its final outcome, is given. The description is subdivided into the hypothesis setup, hypothesis specifications and finally, hypothesis comparison. Based on the latter, the Bayes Factor is presented as the ratio of the marginal likelihoods under both hypotheses under consideration. Finally, it's ambition of use as well as reasonable interpretations of, and conclusions from Bayes Factor results are submitted.
In chapter 3, the *Bayesian two-sample t-test* is presented, leading to the BF as a relevant special case application in psychological research. Lastly, a closed-form for BF calculation, developed by Gönen et al. [2005, p. 253] is introduced and applied thenceforward.
In chapter 4, BF is dialectically discussed in the context of its most divisive component, the effect size prior. The discussion is threefold, adressing prior necessity, prior sensitivity and the restrictions on prior implementation as matters of debate.

Finally, the demand for a precise, subjective choice for the test- relevant prior is concluded to be BF's predominant deficiency.

Chapter 5 eventually proposes the *Imprecise Bayes Factor* (IBF) as a generalisation and enhancement of the BF. Initially, the enhancement target is explained herein. Then, the application of a subjective credal set is presented as main idea for BF enhancement in the context of Imprecise Probabilities. Based thereon, the IBF is defined and its calculation is explained. Differences in interpreting and conclusion making are highlighted. Subsequently, a simulated application sample serves to illustrate the IBF approach in scientific practice and in the end, IBF's advantages over its conventional counterpart are clarified.

2 The Bayes Factor - key figure for Bayesian hypothesis comparison

In line with the principles of Bayesian inference, Harold Jeffreys originated a methodology to quantify comparative evidence in favor a scientific hypothesis. [Kass and Raftery, 1995, p. 773; Ly et al., 2016, p. 19] The degree, to which observed data prefer one hypothesis over another, is accounted as a real value, that was later designated as the *Bayes Factor*. [Kass and Raftery, 1995, p. 773]
Different from frequentist hypothesis testing, Jeffreys' approach targets hypothesis comparison and goes by interpreting sample data as statistical evidence in favor of one hypothesis over another. [Morey et al., 2016, p. 16; Aitkin, 1991, p. 113] As such, a Bayes Factor is solely not apt for decision making among two alternative hypotheses. [Morey et al., 2016, p. 17; Augustin et al., 2014, p. 140]

2.1 The major principles of Bayesian inference

The comparative analysis of scientific hypotheses depicts an important method of Bayesian inference and as such follows both of its major principles: [Rouder et al., 2018, p. 105; Morey et al., 2016, p. 10]

First, Bayesian inference embraces the ”epistemic interpretation” [Etz and Vandekerckhove, 2018, p. 6] of probability. That is to say, a person's incomplete state of prior knowledge is expressible as a probability distribution or conversely, probability is employed to quantify a person's uncertainty or stength of belief based on subjective knowledge. [Morey et al., 2016, p. 10]
Still, Etz and Vandekerckhove [2018, p. 6] clarify:

> ”The fact that epistemic probabilities [...] are subjective does not mean that they are *arbitrary*. Probabilities are not acts of will; they are subjective merely in the sense that they may differ from one individual to the next. That is just to say that different people bring different information to a given problem.”

Second, sample data serve the purpose of updating prior degrees of belief. This updating is implemented according to *Bayes Theorem* or *Bayes' Rule* and results in posterior probabilities. The latter finally represent the degree of belief in a quantity of interest, when composing prior knowledge and sample information. [Augustin et al., 2014, p. 140; Etz and Vandekerckhove, 2018, p. 9; Raftery, 1995, p. 126; Morey et al., 2016, p. 10; Royall, 2010, p. 128, Goldstein, 2006, p. 403]
To cite Royall [2010, p. 128] in that regard,

> ”the main subject matter of statistics is the study of how data sets

> change degrees of belief; from prior, by observation of A, to posterior. They change by Bayes' theorem."

Due to the fact that prior beliefs are transformed into posterior beliefs solemnly through the inclusion of observed sample data, the transformation itself is considered as statistical evidence provided by the data. [Kass and Raftery, 1995, p. 776]

2.2 The Bayesian approach to hypothesis comparison

In the following, a detailed step-by-step description of the Bayesian approach to hypothesis comparison shall be given. Guided by the exposition of Liu and Aitkin [2008, p. 363], the approach shall be subdevided into three major stages. These are the hypothesis setup, hypothesis specification and hypothesis comparison. Finally, the Bayes Factor shall be presented as the key outcome of hypothesis comparison.

2.2.1 Hypothesis setup

Initially, a concrete research question may have arisen in the context of scientific research. In order to become accessible to a statistical analysis, the same need to be transformed into a set of competing hypotheses, whereas the latter depict a simplified, theoretical representation of the real-world matter of interest, it concerns. [Morey et al., 2016, p. 16; Liseo, 2012, p. 198]

In a forward look on this thesis, the considered hypothesis set may have the structure

$$H_0 : \delta = \delta_0 \qquad \text{vs.} \qquad H_1 : \delta \neq \delta_0 \,. \tag{2.1}$$

[Berger and Delampady, 1987, p. 317]
Herein, the subject of interested is represented by δ and δ_0 is a precise value, that a respective researcher assumes, could be approximately true. [Marden, 2000, p. 1316]

A suchlike point-valued conception of the null hypothesis is indeed typical. After all, δ_0 is deemed particularly plausible or of special interest for the scientific inquiry. [Etz and Vandekerckhove, 2018, p. 21; Liu and Aitkin, 2008, p. 363; Ly et al., 2016, p. 19]

2.2.2 Hypothesis specification

In order to submit the considered hypotheses to a statistical analysis, the Bayesian framework initially claims a subjective assignment of prior probabilities $\pi(H_0)$ and $\pi(H_1)$ on the considered hypotheses themselves. [Aitkin, 1991, p. 112; Rouder et al., 2018, p. 105] In the Bayesian understanding, a researcher is expected to have a certain degree of prior belief in his specified hypotheses. [Etz and Vandekerckhove, 2018, p. 10] According to the above stated, Bayesian notion of probabiliy (see section 2.1), he may set his prior beliefs as $\pi(H_0) = \mathbb{P}(H_0)$. Finally, $\pi(H_0)$ and $\pi(H_1)$ are required to add up to 1. [Gönen et al., 2005, p. 252]

Next up, Jeffreys' Bayesian analysis compares considered hypotheses according to their ability to predict an observed sample data set. [Vanpaemel, 2010, p. 492;

Morey et al., 2016, p. 8; Rouder et al., 2018, p. 105]
For this reason, well defined, Bayesian hypotheses need to be represented by a *statistical model* [Liu and Aitkin, 2008, p. 362, 363], that makes precise predictions about the probability of each possible outcome x *under* H_j. The statistical models are yielded over a marginalisation of unambigiously specified, posterior likelihood functions $f(x|H_j)$, $j = 0, 1$, defined below. [Aitkin, 1991, p. 111]
In this regard,

> "[l]ikelihoods can be thought of as how strongly the data are implied by a hypothesis. *Conditional* on the truth of an hypothesis, likelihood functions specify the probability of a given outcome [...]."

[Etz and Vandekerckhove, 2018, p. 9]

Following Liu and Aitkin [2008, p. 363], one may postulate or at least imagine a true probability density function $f^t(x|\theta^t)$, from which an observed sample set $\boldsymbol{x}$ was drawn. In it, θ^t denotes the true parameter (set). Apparently, no prior parameter distributions are occur therein. After all, there is no need to specify uncertainty about the prarmeter's true values.
With the objective to optimally approximate the true density $f^t(x|\theta^t)$, one specifies two candidate likelihood functions $f(x|H_0)$ and $f(x|H_1)$ to represent the probability of each possible outcome x under H_0 and H_1, respectively. [Liseo, 2012, p. 198]

Under H_0, δ is stated to have the value δ_0, precisely. However, the related likelihood function generally depends on a number of other, unknown parameters. These may be denoted jointly as the parameter set θ, hereafter. Thus, the likelihood under H_0 may be defined as $f(x|\theta, \delta = \delta_0)$, where θ is an unknown element of the parameter space Θ. In a Bayesian analysis, every unknown parameter is given its own probability density distribution. [Raftery, 1995, p. 126]
As to that, $\pi_\theta(\theta)$ may denote the parameter prior distribution, that represents prior uncertainty about the true value of θ. This $\pi_\theta(\theta)$ needs to be specified subjectively by the respective analyst. [Liu and Aitkin, 2008, p. 363; Etz and Vandekerckhove, 2018, p. 21]
According to Bayes' Rule, the likelihood function is then multiplied point by point with the respective prior distribution to receive the "posterior likelihood function" [Gallistel, 2009, p. 441]

$$f(x|H_0) = f(x|\theta, \delta = \delta_0)\pi_\theta(\theta). \tag{2.2}$$

Under H_1, also the value of δ - apart from not being δ_0 - is unknown. [Morey et al., 2016, p. 16] Consequently, the likelihood function contains both θ and δ as unknown parameters. In this regard, H_1 can be understood as an "extension of $[H_0]$ by inclusion of a new parameter" [Ly et al., 2016, p. 22]. In order to yield the posterior likelihood function under H_1, an analyst needs to redefine H_1 over an inclusion of subjective prior knowledge about δ. [Royall, 2010, p. 128; Rouder et al., 2009, p. 228]

To both include available information and reveal an appropriate degree of uncertainty about the value of δ under H_1, a continuous prior probability distribution $\pi_\delta(\delta)$ is specified across a range of potential parameter values, indicating that δ is not δ_0. [Rouder et al., 2009, p. 229]
In order to specify a suitable prior parameter distribution $\pi_\delta(\delta)$, data-external information is once more put to use. Based on personal beliefs, professional expertise and other relevant resources, the plausibility of different δ-values is assessed. The ensuing (personal) beliefs are then transformed into a probability density distribution $\pi_\delta(\delta)$ over the parameter space Δ, in turn containing all possible δ-values. [Rouder et al., 2009, p.229-233]
Consequently, $\pi_\delta(\delta)$ represents a researcher's personal state of uncertainty about δ before data is at hand. [Kass, 1992, p. 553; Ly et al., 2016, p. 21; Gallistel, 2009, p. 440]

The above stated hypothesis set (2.1) may hence be restated as

$$H_0 : \delta = \delta_0 \qquad \text{vs.} \qquad H_1 : \delta \sim \pi_\delta(\delta). \tag{2.3}$$

As this notation reveals, the alternative hypothesis is explicitly defined over the parameter prior $\pi_\delta(\delta)$. [Liseo, 2012, p. 199; Liu and Aitkin, 2008, p. 363; Lavine and Schervish, 1999, p. 119]

Finally, the posterior likelihood function under H_1 ensues as

$$f(x|H_1) = f(x|\theta, \delta)\pi_\theta(\theta)\pi_\delta(\delta) \tag{2.4}$$

As the parameter (set) θ occurs in the posterior likelihood of both H_0 and H_1, the same may be referred to as the *common* parameter (set). Now, as δ exclusively enters the marginal likelihood under H_1, it may be called "test-relevant parameter". [Ly et al., 2016, p. 22, 23].

2.2.3 Hypothesis comparison

Up to that point, posterior likelihood functions are specified for both hypotheses under consideration. However, the latter have not been compared as yet.

In order to do so, both hypotheses have to undergo a confrontation with actually *observed* sample data $\boldsymbol{x}$. However, the support for their statistical models depends on the extent, to which their predictions of $\boldsymbol{x}$ match the observed data points in $\boldsymbol{x}$. [Morey et al., 2016, p. 8]

For the further analysis, such an i.i.d. sample $\boldsymbol{x} = (x_1, ..., x_n)$ may be assumed given. Finally, the comparison of the considered hypotheses can be undertaken through a comparison of their respective marginal likelihoods.
Denoted as $m_j(\boldsymbol{x})$, $j = 0, 1$, the marginal likelihood or rather statistical model of H_j is defined as the integral of its posterior likelihood function $f(x|H_j)$ with respect to its respective parameter vector. [Gallistel, 2009, p. 441; Raftery, 1995, p. 128; Liu and Aitkin, 2008, p. 363; Etz and Vandekerckhove, 2018, p. 22] Finally, the statistical models read as:

$$m_0(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x}|H_0)\, d\theta = \int_\Theta f(\boldsymbol{x}|\theta, \delta = \delta_0)\pi_\theta(\theta)\, d\theta \tag{2.5}$$

$$m_1(\boldsymbol{x}) = \int_\Theta \int_\Delta f(\boldsymbol{x}|H_1)\, d\theta\, d\delta = \int_\Theta \int_\Delta f(\boldsymbol{x}|\theta, \delta)\pi_\theta(\theta)\pi_\delta(\delta)\, d\delta\, d\theta. \tag{2.6}$$

[Wang and Liu, 2016, p.196]
In these formulas, the parameter priors $\pi_\theta(\theta)$ and $\pi_\delta(\delta)$ are part of the hypothesis specifications and the values of the likelihood functions are determined by the observed data $\boldsymbol{x}$. [Gallistel, 2009, p. 441; Wang and Liu, 2016, p. 196; Wasserman, 2000, p. 95]

Contentwise, the marginal likelihood can be understood as the weighted average of the likelihood computed over all parameter values θ (and δ) according to the respective, specified parameter prior distributions. [Etz and Vandekerckhove, 2018, p.16; Matthews, 2011, p. 844]
As such, it depicts a possible, joint consideration of a class of precise probability density functions $f(x|\theta)$ under H_0 and $f(x|\theta, \delta)$ under H_1.

Now, that $m_j(\boldsymbol{x})$ and $\pi(H_j)$, $j = 0, 1$, are given, this

> "turns the problem of statistical inference into a problem of probabilistic deduction, where the posterior distribution [...] can be calculated by Bayes rule."

[Augustin et al., 2014, p. 140]

The same applies as

$$\pi(H_j|\boldsymbol{x}) = \frac{\pi(H_j)m_j(\boldsymbol{x})}{\pi(H_0)m_0(\boldsymbol{x}) + \pi(H_1)m_1(\boldsymbol{x})}, \tag{2.7}$$

which provides the researcher with the posterior probability of either hypothesis. [Liseo, 2012, p. 199; Ly et al., 2016, p. 20; Gönen et al., 2005, p. 252; Wang and Liu, 2016, p. 196]

Yet, the primary concern of hypothesis comparison is to gain insights about which of the competing hypotheses the data support *more* strongly. One wishes to quantify the degree, to which the data are indicative of one hypothesis *over* another. [Rouder et al., 2018, p. 105]

Hence, the comparison of H_0 and H_1 is implemented as a "posterior odds ratio" [Liu and Aitkin, 2008, p. 363]:

$$\underbrace{\frac{\pi(H_0|\boldsymbol{x})}{\pi(H_1|\boldsymbol{x})}}_{\text{Posterior Odds}} = \underbrace{\frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})}}_{\text{Bayes Factor}} \times \underbrace{\frac{\pi(H_0)}{\pi(H_1)}}_{\text{Prior Odds}} \tag{2.8}$$

[Ly et al., 2016, p. 20]
Herein, the Prior Odds states the degree, to which a person's prior beliefs favor H_0 over H_1 beforehand a data analysis. [Matthews, 2011, p. 848; Morey et al., 2016, p. 12] As such, the prior probabilities $\pi(H_0)$ and $\pi(H_1)$

> "[...] reflect our prior beliefs/knowledge, and have no effect on the balance of evidence from data - rather, they shape how this evidence is used to arrive at a new belief state."

[Matthews, 2011, p. 848]

By contrast, the Posterior Odds denote the relative plausibility of H_0 over H_1, taking account of the data. [Rouder et al., 2009, p. 228; Morey et al., 2016, p. 12] Apparently, they are dependent on the Prior Odds. As the latter is a purely subjective quantity, the Posterior Odds is usually considered unfit for an exclusive quantification of evidence in the data. [Kass and Raftery, 1995, p. 773]
Instead, Harold Jeffreys promoted the Bayes Factor as the central measure for hypothesis comparison. [Gönen et al., 2005, p. 252]

2.3 The Bayes Factor

The Bayes Factor is sometimes referred to as the "centerpiece" [Kass and Raftery, 1995, p. 773] or "corner-stone" [Johnson, 2005, p. 689] of the Bayesian approach to hypothesis comparison, as it pools most of the analysis within one value.
It is the final result of hypothesis comparison, based on previous hypothesis speci-

fications and may among others affect a subsequent hypothesis selection. [Sinharay and Stern, 2002, p.196]

2.3.1 Definition

While retaining to the above stated hypothesis set (2.3), the Bayes Factor $BF_{01}(\boldsymbol{x})$ takes values in $(0, \infty)$ and is defined as

$$BF_{01}(\boldsymbol{x}) = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{\int_\Theta f(\boldsymbol{x}|\theta, \delta = \delta_0)\pi_\theta(\theta)\, d\theta}{\int_\Theta \int_\Delta f(\boldsymbol{x}|\theta, \delta)\pi_\theta(\theta)\pi_\delta(\delta)\, d\delta\, d\theta}\,. \tag{2.9}$$

[Marden, 2000, p. 1318; Wang and Liu, 2016, p. 196]

Verbalized, the Bayes Factor - comparing the null hypothesis H_0 to the the alternative hypothesis H_1 - is the ratio of two marginal likelihoods $m_j(\boldsymbol{x})$, $j = 0, 1$.

In the words of Liu and Aitkin [2008, p. 363],

> ”[t]he Bayes Factor is an extension of the standard likelihood ratio, where the likelihood is defined as the probability of the observed data, given a model with specified parameter values. Whereas the likelihood ratio compares two models by their respective likelihoods, the Bayes factor compares two model classes by their respective marginal likelihoods. [...] In brief, a Bayes factor is a likelihood ratio for two model classes.”

The subscripts within BF_{01} state the order, in which the competing hypotheses are compared to each other. In definition 2.9, $m_0(\boldsymbol{x})$ is in the numerator, compared to $m_1(\boldsymbol{x})$ in the denominator. [Etz and Vandekerckhove, 2018, p.22]
As such, high values above 1 suggest a high plausibility of H_0 compared to H_1. The closer the Bayes Factor is to 0, the more strongly do the data favour H_1 over H_0. Finally, a Bayes Factor of value 1 states that the data are just as much evidence for H_0 as for H_1. [Augustin et al., 2014, p. 152; Raftery, 1995, p. 129; Ly et al., 2016, p. 22]

The order, according to which the hypotheses are compared, is freely exchangable under the relation

$$BF_{01}(\boldsymbol{x}) \;=\; \frac{1}{BF_{10}(\boldsymbol{x})} \;. \tag{2.10}$$

[Wang and Liu, 2016, p. 196]

2.3.2 Ambition of use

The use of Bayes Factors rests on the general ambition to compare scientific hypotheses - an endeavour, that is scientifically met by a confrontation of the latter with an observed sample $\boldsymbol{x}$ under the primary question:
Which of the two competing hypotheses do the data favor and how strongly do they favor it?, or similarly: *What evidence do the data hold about the two hypotheses comparatively?* The answer, however, shall be provided by the Bayes Factor.
Finally, one calculates a Bayes Factor under the ambition to quantify the extent, to which observed data endorse or negate one considered hypothesis over another. [Lavine and Schervish, 1999, p. 119; Gallistel, 2009, p. 441; Kass, 1992, p. 551]

2.3.3 Interpretation

Now, that the ambition behind the use of Bayes Factors is posed, a precise interpretion of the Bayes Factor shall be given. This shall serve to compare the Bayes Factor's actual meaningfulness with the above stated, desired manner of use.
As a direct component of the odds notation for model comparison (2.8), the Bayes Factor is the multiplier, transforming Prior Odds into Posterior Odds. [Johnson, 2005, p. 689; Liu and Aitkin, 2008, p. 363]

$$\frac{\pi(H_0)}{\pi(H_1)} \times BF_{01}(\boldsymbol{x}) = \frac{\pi(H_0|\boldsymbol{x})}{\pi(H_1|\boldsymbol{x})} \tag{2.11}$$

In other words, it expresses the data induced change of belief when going from the former to the latter. [Rouder et al., 2018, p. 105; Rouder et al., 2009, p. 228; Ly et al., 2016, p. 19]

Transposing the equation for the Bayes Factor, the same may as well be defined as the ratio of the Posterior Odds to the Prior Odds:

$$BF_{01}(\boldsymbol{x}) = \frac{\pi(H_0|\boldsymbol{x})}{\pi(H_1|\boldsymbol{x})} \times \left[\frac{\pi(H_0)}{\pi(H_1)}\right]^{-1} \tag{2.12}$$

[Morey et al., 2016, p. 12]
In this sense, it measures the degree, to which the observed data $\boldsymbol{x}$ prompt a revision of the odds in favor of a hypothesis when going from the prior to the posterior. [Rouder et al., 2009, p. 105; Matthews, 2011, p. 851]

From a mathematical point of view, the Bayes Factor simply denotes the ratio of two marginal likelihoods $m_j(\boldsymbol{x})$, j = 0,1. Denoted as statistical models, these respectively measure the degree, to which model-based predictions of x match the observed sample $\boldsymbol{x}$. As to that, the Bayes Factor quantifies the relative predictive accuracy of one model over the other, whereby it applies that

> "[i]f the probability of observed data is high, then the model predicted the observed data to be where they were observed. If the probability of data is low, then the model did not predict the observations well."

[Rouder et al., 2018, p. 105]

Nevertheless, Bayes Factor's mathematical structure is compatible with its common interpretation as statistical evidence on account of the *likelihood principle* and the *law of likelihood.* [Royall, 2010, p. 122, 123]
The former states that the entire evidence from a dataset $\boldsymbol{x}$ - so far as relevant for the evaluation of the considered hypotheses - is comprised in the likelihood.
The latter says, if an observation $\boldsymbol{x}$ is more plausible under one hypothesis than under another - in a sense of providing higher predictive accuracy - then $\boldsymbol{x}$ is evidence in support of that hyphothesis. The degree, to which $\boldsymbol{x}$ supports H_0 over H_1 or vice versa, is then defined as the ratio of their respective model's likelihoods.
When extending the scope of these two precepts from likelihoods to marginal likelihoods, they are argued to be applicable to the Bayes Factor and thus faciliate its interpretation as strength of statistical ecidence in favor of one out of two competing hypotheses. [Etz and Vandekerckhove, 2018, p. 24; Rouder et al., 2018, p. 105; Rouder et al., 2009, p. 228]

2.3.4 Arguable conclusions

While clearing away typical misconceptions about the Bayes Factor's explanatory power, the following section shall reveal, what may arguably be concluded from a

Bayes Factor result.

A Bayes Factor may be conceived as a statement about the evidence, an observation $\boldsymbol{x}$ delivered about the odds on H_0 relative to H_1. Within the odds notation of Bayes' Rule, it causes a redistribution of probability between two competing hypotheses.

However, a Bayes Factor makes no actual belief statement about the probability of either hypothesis. In that regard, it differs from the Posterior Odds. Other than the latter, a Bayes Factor does not indicate one's final belief about H_0 relative to H_1. Instead, it should be conceived as a learning factor, that different researchers may adapt to their initially held Prior Odds. In other words, a Bayes Factor, that favors H_0 over H_1, does not imply that a researcher's final belief in the null hypothesis will top that in the alternative. It can merely raise his degree of belief compared to that, he held a priori. [Etz and Vandekerckhove, 2018, p. 10; Lavine and Schervish, 1999, p. 121]

Furthermore, nothing about the adequacy of a hypothesis may be concluded from a Bayes Factor alone. That is because of two major properies, the latter holds as a measure of statistical evidence:
First, evidence provided by a Bayes Factor is "relational" [Morey et al., 2016, p. 8]. That is to say, a sample $\boldsymbol{x}$ alone - in a sense of being isolated from the considered hypotheses - does not hold any evidence. $\boldsymbol{x}$ function as evidence only through their impact on the probabilities of the stated hypotheses. In fine, a Bayes Factor's evidential content lies solely in the relation between data and hypotheses.
By implication, a Bayes Factor, stating that the model under H_0 matches the data $\boldsymbol{x}$ better than that under H_1, does not rule out, that both models might fit the data poorly. [Morey et al., 2016, p. 17] Hence, the substantial worth of a Bayes Factor presumes a meaningful choice of the candidate hypotheses. Derived conclusions must be assessed relational to the models, it concerns.
Second, evidence provided by a Bayes Factor is relative. All support from data can do is to make one hypothesis more plausible than one other. However, it cannot expose a hypothesis well-suited all by itself. To quote Lavine and Schervish [1999, p. 121], a Bayes Factor may be the answer to:

> "How well, relative to each other, do the hypotheses explain the data?"

It cannot give information about how well the hypotheses explain the data *overall*. Finally, valid conclusions from a Bayes Factor remain comparative. Under no circumstances are they absolute. [Morey et al., 2016, p. 8]

Summing up, the Bayes Factor is suited for hypothesis comparison under the assumption of an adequate hypothesis set and the consideration of the internal prior distributions. It makes a comparative evidence statement, applicable for data-induced belief updating.

Yet, a Bayes Factor alone is inapt for hypothesis selection. Deciding for one hypothesis implies, that its underlying model is considered ”good enough in some way” [Morey et al., 2016, p. 8]. However, the comparative character of a Bayes Factor bans suchlike, absolute statements.

Rather, it is up to the researcher concerned to reflect about the meaning of a Bayes Factor within the context of the underlying research question. [Etz and Vandekerckhove, 2018, p. 5]

3 The Bayes Factor for independent two-sample comparisons in psychological research

Up to that point, Bayesian hypothesis comparison - as being centered around the Bayes Factor - was given a general account. Going forward, it shall be applied to a special case, commonly referred to as the Bayesian two-sample t test. [Gönen et al., 2005, p. 252; Rouder et al., 2009, p. 225, Fox and Dimmic, 2006, n.p.]
Finally, this application will yield an easily calculable, closed-form Bayes Factor (BF herafter). [Gönen et al., 2005, p. 253]

3.1 The scientific initial situation

Initially, the circumstances of the considered application case shall be described, whereby the considered research discipline will be psychological research.
The research question may concern the presence or absence of an effect, indicated by the difference between two independent groups. [Rouder et al., 2018, p. 102] Consider for instance a gender or skin colour effect being at question.
The corresponding research setting consists of the two independent groups. Their experimental conditions differ according to the question of interest. The consequent research question reads as: *Do the groups differ?*, together with: *Is the resulting in-sample difference big enough infer an effect?* [Fox and Dimmic, 2006, n.p.]

3.2 The experimental setup

The proposed experimental setup corresponds to that of a classical two-sample t test. In order to examine a potential group difference, independent samples need to be drawn from the two considered groups. Consequently, there may be a sample data set $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}\}$ of size $n = n_1 + n_2$, composed of two independent random samples

$$\boldsymbol{x_1} = \{x_{11}, ..., x_{1n_1}\} \quad \text{and} \quad \boldsymbol{x_2} = \{x_{21}, ..., x_{2n_2}\}. \tag{3.1}$$

Both are assumed to be drawn from normally distributed populations. The corresponding population means may be μ and $\mu + \Delta\mu$, respectively. Thereby, $\Delta\mu$ may be referred to as the total effect. [Rouder et al., 2009, p. 234] The normal variance, denoted as σ^2, is assumed to be identical within both groups. [Wang and Liu, 2016, p. 195] Under the assumption of conditional independence, the sample $\boldsymbol{x}$ is modeled as

$$x_{11}, ..., x_{1n_1} \overset{iid.}{\sim} N(\mu, \sigma^2) \quad \text{and} \quad x_{21}, ..., x_{2n_2} \overset{iid.}{\sim} N(\mu + \Delta\mu, \sigma^2), \tag{3.2}$$

respectively. [Gönen et al., 2005, p. 253; Wang and Liu, 2016, p. 195; Ly et al., 2016, p. 23]

Within the frequentist framework, one would hereafter proceed with a calculation of the t-statistic

$$t = \frac{\bar{\boldsymbol{x_1}} - \bar{\boldsymbol{x_2}}}{s_p \, / \sqrt{n_\delta}} \, . \tag{3.3}$$

Thereby, $\bar{\boldsymbol{x_i}} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ is the sample mean of group i, $i = 1, 2$.
Furthermore,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{3.4}$$

depicts the pooled variance estimate with s_i^2 being the sample variance of group i. Finally, $n_\delta = (\frac{1}{n_1} + \frac{1}{n_2})^{-1}$ is commonly called the effective sample size. [Gönen et al., 2005, p. 252]

The according p-value is then defined as $p = 2\,\mathbb{P}\,(T_\nu > |t|)$, whereby T_ν follows the T- distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.
At it, H_0 is rejected in favor of H_1, if p is less than a prespecified significance level α. [Wang and Liu, 2016, p. 195]

However, the follwing chapter will be dedicated to the presentation of the *Bayesian* approach, yielding a special Bayes Factor.

3.3 The "Bayesian two-sample t test"

In the following, the Bayesian approach to hypothesis comparison shall be applied to the above stated case of an independent two-sample comparison. Thereby, conformity to the general depiction in chapter 2.2 shall be preserved.

3.3.1 Hypotheses setup

Still in conformity with the classical two-sample t test, the question of interest (see section 3.1) may be transferred into the hypothesis set

$$H_0 : \Delta\mu = 0 \quad vs. \quad H_1 : \Delta\mu \neq 0. \tag{3.5}$$

[Wang and Liu, 2016, p. 195; Rouder et al., 2018, p. 105]

Generally, it is helpful to reparameterise the total effect $\Delta\mu$ into the so-called standardized effect size $\delta = \frac{\Delta\mu}{\sigma}$ and concurrently revise the hypothesis set to

$$H_0 : \delta = 0 \quad vs. \quad H_1 : \delta \neq 0. \tag{3.6}$$

[Killeen, 2005, p. 346; Rouder et al., 2009, p. 230; Ly et al., 2016, p. 22]
As a "dimensionless quantity" [Gönen et al., 2005, p. 253], δ follows a scale, based on which researchers may set generally applicable benchmarks for small, medium or large effects within psychological research. [Cohen, 1988, n.p.]
The reparameterisation eases both the assessment and the comparison of effects, without changing the basic nature of the hypothesis set.

Verbalised, one compares a precise null hypothesis H_0 against a composite alternative H_1 with respect to δ. As δ is representative of the considered group difference, H_0 implies equal group means and H_1 assumes a group difference of a yet unspecific extent.

Finally, this is where the Bayesian and the frequentist approach part company.

3.3.2 Hypothesis specification

The natural, Bayesian approach to compare the considered hypotheses it to calculate a Bayes Factor. This initially presumes a specification of respective marginal likelihoods.

H_0 assignes δ the precise point value 0. By implication, the entire sample $\boldsymbol{x} =$

$\{x_{11}, ..., x_{1n_1}, x_{21}, ..., x_{2n_2}\}$ is assumed to be drawn from a population $X \sim N(\mu, \sigma^2)$. Consequently, likelihood function is defined as $f(x|\mu, \sigma^2, \delta = 0)$. At it, μ and σ^2 depict unknown parameters for whom prior densities $\pi_\mu(\mu)$ and $\pi_{\sigma^2}(\sigma^2)$ need to be specified. Finally, the posterior likelihood function under H_0 ensues as

$$f(x|H_0) = f(x|\mu, \sigma^2)\pi_\mu(\mu)\pi_{\sigma^2}(\sigma^2)\,. \tag{3.7}$$

H_1 implies that the population means are not equal. As the disparity's degree is not regarded in $\delta \neq 0$, Ly et al. [2016, p. 23] claim that H_1 primarily loosens H_0's restriction on δ. In fact, H_1 an unspecific amount of uncertainty about the extent, to which the group means are assumed to differ.
At this point, the Bayesian framework demands a specification of this uncertainty in form of a parameter prior distribution $\pi_\delta(\delta)$. [De Santis and Spezzaferri, 1997, p. 503] The same shall spread probability mass over a range of potential δ-values according to their (subjectively ascribed) plausibility under H_1. [Morey et al., 2016, p. 11,14]
Accordingly, the Bayesian hypothesis set ensues as

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \sim \pi_\delta(\delta)\,. \tag{3.8}$$

[Rouder et al., 2009, p. 230]
Now, indicating that δ is distributed precisely according to $\pi_\delta(\delta)$, H_1 turns from a general hypothesis into a specific one. This allows for the attribution of a posterior likelihood function as

$$f(x|H_1) = f(x|\mu, \sigma^2, \delta)\pi_\mu(\mu)\pi_{\sigma^2}(\sigma^2)\pi_\delta(\delta)\,. \tag{3.9}$$

In order to become definite, π_μ, π_{σ^2} and π_δ need to be specified by the analyst according to his prior information and respective beliefs.

The specification of the prior on δ is given an emphasized position within this evaluation process. This is because π_δ - other than π_μ and π_{σ^2} - will later on enter the BF only through the marginal likelihood under H_1. As such, it considerably affects on it's outcome. π_δ may thus be stated the (only) test-relevant prior. [Ly et al., 2016, p. 23]

In this context, a normal distribution shall be chosen to represent prior knowledge about the value of δ. The choice of a normal distribution for the effect size prior is chiefly promoted in psychological research. [Berger and Sellke, 1987, p. 112; Gönen

et al., 2005, p. 253]
On the one hand, 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 evenly declines as the distance to the mean increases. [Rouder et al., 2009, p. 232; Matthews, 2011, p. 844]
On the other hand, mean and variance are quite intuitive measures to be specified subjectively. Their respective effect on the shape of the prior are easy to imagine, also for non-statisticians. This faciliates reasonable hyperparameter choices and in turn an alternative hypothesis that has a reasonable counterpart in the real-world. [Rouder et al., 2009, p. 229-233]
Consequently, $\pi_\delta(\delta)$ is specified as $N(\mu_\delta, \sigma_\delta^2)$ and with it

$$H_1 : \delta \sim N(\mu_\delta, \sigma_\delta^2). \tag{3.10}$$

Therein, μ_δ and σ_δ^2 are commonly referred to as hyperparameters. [Gönen et al., 2005, p. 254] Under the categorical assumption of a normal distribution, these are the only variables to be chosen (subectively) by the respective analyst. [Berger and Sellke, 1987, p. 112]

After that, prior distributions need to be posed on the remaining model parameters μ and σ^2.
Denoted as $\pi_\mu(\mu)$ and $\pi_{\sigma^2}(\sigma^2)$, they enter the posterior likelihood functions under *both* hypotheses. It is commonly argued, that this largely depletes their effect on the BF outcome and in turn makes their specificaion less critical. [Rouder et al., 2009, p. 231] Hence, Jeffreys proposed non-informative priors, that have quite evolved into standard and shall herein be adopted as

$$\mu \propto const. \quad \text{and} \quad \sigma^2 \propto \frac{1}{\sigma^2}. \tag{3.11}$$

[Wang and Liu, 2016, p. 196; Gönen et al., 2005, p. 253]

3.3.3 Hypothesis comparison

In the following, steps are taken towards the actual hypothesis comparison.
Within the Bayesian framework, hypotheses are compared by how well they relatively predict observed sample data. In other words, support for a scientific hypothesis depends on how its marginal likelihood is geared to an observed sample in comparison to that of the other hypothesis under consideration. [Morey et al., 2016,

S.8]
This initially demands the incorporation of $\boldsymbol{x} = \{x_{11}, ..., x_{1n_1}, x_{21}, ...x_{2n_2}\}$ into the ongoing analysis. Given $\boldsymbol{x}$, the calculation of the marginal likelihoods under either hypothesis may follow. By analogy with the equations 2.6, the marginal densities $m_0(\boldsymbol{x})$ and $m_0(\boldsymbol{x})$ are composed as

$$m_0(\boldsymbol{x}) = \iint f(\boldsymbol{x}|\mu, \sigma^2, \delta = 0)\pi_{\sigma^2}(\sigma^2)\pi_\mu(\mu)\, d\mu\, d\sigma^2 \tag{3.12}$$

and

$$m_1(\boldsymbol{x}) = \iiint f(\boldsymbol{x}|\mu, \sigma^2, \delta)\, \pi_{\sigma^2}(\sigma^2)\, \pi_\mu(\mu)\, \pi_\delta(\delta)\, d\delta\, d\mu\, d\sigma^2 \tag{3.13}$$

in this very case.

Within the Bayesian concept, $m_0(\boldsymbol{x})$ represents the model under the assumption of equal group means. In reverse, $m_1(\boldsymbol{x})$ signifies that under the assumption of a $N(\mu_\delta, \sigma^2_\delta)$ - distributed, non-zero effect size δ. Within both marginal likelihoods, the so-called "nuisance" parameters [Gönen et al., 2005, p. 254] μ and σ^2 are assigned the improper priors $\pi_\mu(\mu) \propto const.$ and $\pi_{\sigma^2}(\sigma^2) \propto \frac{1}{\sigma^2}$, respectively.

Finally, the actual hypothesis comparison is accomplished by the ratio of $m_0(\boldsymbol{x})$ and $m_1(\boldsymbol{x})$, which is in turn amounts to the subsequent definition of the BF.

3.4 The Bayes Factor for independent two-sample comparisons

3.4.1 Definition

Finally, $m_0(\boldsymbol{x})$ and $m_1(\boldsymbol{x})$ are compared to yield the relative strength of evidence, the data $\boldsymbol{x}$ hold for one of the considered hypothesis over the other.
According to definition 2.9, this is done by a calculation of the respective BF.

$$BF_{01}(\boldsymbol{x}) = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{\iint f(\boldsymbol{x}|\mu, \sigma^2, \delta = 0)\, \pi_{\sigma^2}(\sigma^2)\, \pi_\mu(\mu)\, d\mu\, d\sigma^2}{\iiint f(\boldsymbol{x}|\mu, \sigma^2, \delta)\, \pi_{\sigma^2}(\sigma^2)\, \pi_\mu(\mu)\, \pi_\delta(\delta)\, d\delta\, d\mu\, d\sigma^2} \tag{3.14}$$

The numerator measures the marginal likelihood of $\boldsymbol{x}$ under the scientific assumption of equal group means. The denominator depicts the equivalent under the assumption that $\delta \sim N(\mu_\delta, \sigma^2_\delta)$. As such, the above stated BF is treated as the statistical evidence, the data $\boldsymbol{x}$ hold for the absence of an effect in comparison to a $\pi_\delta(\theta_\delta)$ - distributed effect size.

3.4.2 Special implementation

For precisely the above stated case, Gönen et al. [2005, p. 253] devised a closed-form implementation, which allows for a BF formula solely dependent on the pooled-variance two-sample t-statistic under H_0 and H_1, each.
The concrete implementation applies as

$$BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma^2_\delta) = \frac{T_\nu(t\,|\,0, 1)}{T_\nu(t\,|\,n_\delta^{1/2}\mu_\delta, 1 + n_\delta\sigma^2_\delta)}. \tag{3.15}$$

Herein, t stands for the pooled variance two-sample t-statistic (see equation 3.3). μ_δ and σ^2_δ are the hyperparameters of the normally distributed effect size prior π_δ under H_1 (see definition 3.10).
Finally, $T_\nu(.|a, b)$ depicts the probability density function of a random variable $Y/\sqrt{U/\nu}$, where $Y \sim N(a, b)$ and $U \sim \chi^2(\nu)$ independent of Y. [Gönen et al., 2005, p. 253]

It is withal not unusual to set $\mu_\delta = 0$ in practical applications to reflect uncertainty about the direction of the effect being at question. [Wang and Liu, 2016, p. 196; Rouder et al., 2009, p. 229; Rouder et al., 2018, p. 104]
In this case, the special implementation stated above, can further be simplified to

$$BF_{01}(\boldsymbol{x}, \sigma^2_\delta) = \left[\frac{1 + t^2/\nu}{1 + t^2/\{\nu(1 + n_\delta\sigma^2_\delta)\}}\right]^{(\nu+1)/2} \times (1 + n_\delta\sigma^2_\delta)^{-1/2} \tag{3.16}$$

Herein, $\pi_\delta(\theta_\delta)$ is specified as $N(0, \sigma_\delta^2)$, which makes σ_δ^2 the only parameter to be chosen by the analyst. [Gönen et al., 2005, p. 253; Wang and Liu, 2016, p. 195, 196]

Irrespective of the choice of μ_δ, the following shall be pointed out on the whole:
For the above stated special case of an independent two-sample comparison, the Bayes Factor depends on observed data only through their corresponding t-statistic. This enables for a facile calculation and standardized software implementations - pleasant features, that are otherwise unusual in the context of Bayesian analysis. To obtain a BF, all the user has to do, is insert the respective sample data and specify the test relevant prior π_δ by means of the hyperparameters σ_δ^2 and μ_δ according to his prior knowledge and skilled beliefs. Finally, this affords him an easy opportunity for an eventual sensitivity analysis.

4 A dialectical discussion of the Bayes Factor in the context of the test relevant prior

The BF approach for independent two-sample comparisons has become quite popular in psychological research and a number of other research domains. [Van De Schoot et al., 2017, p. 218-221; Rouder et al., 2018, p. 102
According to Wang and Liu [2016, p. 195]

> ”[t]he Bayesian approach to statistical design and analysis is emerging as an increasingly effective and practical alternative to the frequentist one.”

More than few scientists have even promoted the BF as a superior measure for respective theory evaluations. [Rouder et al., 2009; Vanpaemel, 2010; Gallistel, 2009; Rouder et al., 2018]

As this braces the question on BF's effective qualities, the following chapter shall give a detailed overview of the advantages of and the current controversy around the BF.
First, some chief arguments for BF-approval shall be submitted in this regard.
After that, the BF shall dialectically be discussed over its most critical component, the test-relevant prior π_δ. The discussion shall be threefold, respectively approaching one concrete matter of debate around π_δ. Whereas the first matter refers to the general necessity of π_δ, the second attaches the sensitivity of BF to varying π_δ-choices and the last is about the standards of π_δ- implementation. Every matter is first given a neutral description and then critical arguments are opposed to approving counterarguments.
At last, the contrasting notions shall be balanced for the purpose of drawing a conclusion about BF's particular strengths and weaknesses in the light of π_δ.

4.1 Arguments for Bayes Factor - approval

First of all, scientists approve the BF for its logically sound, philosophical underpinning through Bayes Rule. They support the Bayesian notion of probability as a statement of personal knowledge and point out, how the focus of Bayesian analyses lies on the rational updating of the Prior Odds in the light of empirical data. A scientist may take BF as guidance from data, telling him how to rationally revise his pre-existing information. [Rouder et al., 2018, p. 102] Thus, BF is said to match the basic idea of scientific research work. [Kass and Raftery, 1995, p. 792; Rouder et al., 2009, p. 228]

Furthermore, advocats point out the Bayes Factor's clear interpretative framework. [Matthews, 2011, p. 846]
A BF_{01} of value 2 for example, reveals that the data $\boldsymbol{x}$ provide twice as much statistical evidence for H_0 than for H_1. This provides the analyst with a clear evidence statement.

Even more, the calculation of a BF allows the analyst to define the alternative hypothesis specifically in accord with his research interest. [Berger and Delampady, 1987, p. 319]
This is valuable, because it can often be irrelevant that δ is not exactly 0, as long as the difference remains too small to be meaningful in the research context.

Moreover, the BF is liable to the likelihood principle. [Rouder et al., 2018, p. 105]
It is commonly argued that its valid conclusions thus depend only on observed sample data and not on theoretically assumed, but practically unobserved data. [Matthews, 2011, p. 846]

Finally, BF allows a researcher to gain relative evidence in favor of H_0. Other than in frequentist hypothesis tests, the analyst may thereby also gain relative support the invariance of a certain variable. [Rouder et al., 2018, p. 105]
Advocats of the Bayes Factor commonly point out, how the demonstration of sameness or invariance constitutes a major part of scientific findings. [Rouder et al., 2009, p. 225, 228, 233; Kass and Raftery, 1995, p. 791; Gallistel, 2009, p. 439]

Respecting the above stated, favourable votes, BF nevertheless remains controversial. It's validity underlies a critical debate, largely confined to the specification of the test-relevant prior π_δ. [Kass and Raftery, 1995, p. 792; Sinharay and Stern, 2002, p. 196; Morey et al., 2016, p. 16] Within the whole BF-approach, the same is being criticised most often and for a variety of reasons. [Vanpaemel, 2010, p. 491; Liseo, 2012, p. 197]

4.2 Discussion of prior necessity

The first matter of debate affects the categorical necessity to specify proper prior distributions on all unknown parameters implicit to H_0 and H_1. [Johnson, 2005, p. 689; Morey et al., 2016, p. 15, 16; Kass and Raftery, 1995, p. 781; Vanpaemel, 2010, p. 492; Rouder et al., 2009, p. 229] As Morey et al. [2016, p. 16] formulates it aptly,

> " the prior distribution [...] ensures that the model has a definite marginal likelihood, and thus establishes a bridge between the hypothesis and the data."

This particularly requires the analyst to set the test-relevant prior π_δ within H_1. As the same is assumed to be a normal distribution, it is up to the respective analyst to specify precise values for the mean μ_δ and the variance σ_δ^2 in order to yield a BF. Respective of his paricular choice, probability mass is spread on a series of δ-values according to $N(\mu_\delta, \sigma_\delta)$. [Liseo, 2012, p. 197; Rouder et al., 2009, p. 230]
In short, BF (see equation 3.15) cannot possibly be calculated without a precise specification of the hyperparameters μ_δ and σ_δ^2.

4.2.1 Reproving argumets on prior necessity

(a) **Excessiveness of practical efforts**

Critics commonly oppose the necessity for π_δ by itself. They deem the implied specification of μ_δ and σ_δ^2 an additional burden on the analyst, making theory evaluation needlessly complex. [Goldstein, 2006, p. 417]
Especially, when prior knowledge is vague, the selection of precise values for μ_δ and σ_δ^2 strikes them as a disproportionate practical effort. [Goldstein, 2006, p. 413; Kass and Raftery, 1995, p. 776, 781]
Finally, the specification of π_δ deters the user and thus prevents him from routined BF applications. [Liseo, 2012, p. 205]

(b) **Proneness to misspecifications**

Moreover, π_δ is prone to missspecification in several points. The case of sufficient prior knowledge or profound prior information from alike data sets is rare. In fact, only fewest researchers would be safe to say, their prior choice for μ_δ and σ_δ was beyond dispute. Not only are subjective beliefs typically too rough to define abstract parameter values, it is also a simplistic approximation to assume that π_δ is normally distributed.
Summing it, the specification of π_δ comes along with a considerable amount of uncertainty and that implies error-proneness. Based on the typical extent of

initial information on δ, it is likely to set unfit hyperparameters.

The bottom line is: The BF approach imposes great importance on a distribution that is highly error-prone, likely to be incorrect and can at best be approximate.

(c) **Dependence on model accuracy**

In order to be useful to a certain research question, BF is predicated on a meaningful set of hypotheses. [Liu and Aitkin, 2008, p. 367; Morey et al., 2016, p. 16, 17]

After all, π_δ co-specifies the marginal likelihood that enters the BF under H_1. [Liu and Aitkin, 2008, p. 363; Liseo, 2012, p. 199]

As so, explicitly H_1 and implicitly BF imply a fine choice of π_δ. However, an implausible choice of π_δ locate BF somewhere between meaningless and wrong. [Morey et al., 2016, p. 16; Liseo, 2012, p. 213]

(d) **Implausibility of prior precision**

The next point of criticism hits the necessity of a *precise* prior π_δ and can aptly be summed, quoting Joyce [2010, p. 281]:

> "Belief is not all-or-nothing. Opinions come in varying gradations of strength which can range from full certainty of truth, through equal confidence in truth and falsehood, to complete certainty of falsehood."

For the considered special case, this criticises to the demand imposed on analysts to have *precise* credences on μ_δ and σ_δ in order to assign them precise, numerical values. In fact, a whole series of scientific work has been devoted to the psychological absurdity of numerically sharp degrees of belief. It is hereby argued that human knowledge can at best be represented through a value range. [Joyce, 2010, p. 282, 283]

Clearly, this viewpoint particularly applies to prior knowledge of the unknown effect size parameter δ. To put it in the accusing words of Goldstein [2006, p. 414],

> "[A] true subjective formulation should start by recognising the limited abilities of the individual to make large collections."

Consequently, critics allege the choice of π_δ arbitrariness and unjust make-belief of precision. [Goldstein, 2006, p. 411; Kass and Raftery, 1995, p. 781]

4.2.2 Approving counterarguments

(a) **Incorporation of data-external information**

Proponents challenge the first two points of criticism. They regard the need to specify π_δ less as an effort or burden, but more as a chance to incorporate valuable, data-external information into the analysis. [Gelman and Hennig, 2017, p. 207; Matthews, 2011, p. 848, 852; Vanpaemel, 2010, p. 491; Kass and Raftery, 1995, p. 776]

With the specification of π_δ, analysts are given the opportunity to make use of their relevant prior knowledge. Most often, they are professionals with long-term experience in their respective research field. More than likely can they draw on related studies, similar test results or simply worth-while working experience. The case of sheer ignorance, however, may be dismissed as very unlikely in the considered context. This makes an absurd choice π_δ at least unlikely. From this perspective, π_δ enhances the evaluation process, rather than impeding it.

(b) **Endorsement of a reasonable alternative hypothesis**

The importance of a well-specified π_δ in view of reasonable BF outcomes can not be opposed. Still, advocats claim the contribution of π_δ to yield a reasonable alternative hypothesis. [Vanpaemel, 2010, p. 491; Matthews, 2011, p. 848]
Within the Bayesian framework, Vanpaemel [2010, p. 491] declares the considered hypotheses as ”quantitatively instantiated theories”, that stem from rational beliefs. On that note, an accurate hypothesis specification requires an accurate exposition of beliefs and for this purpose, the analyst may at will draw on π_δ. Finally, one may also call π_δ co-reponsible for a meaningful alternative hypothesis.

4.3 Discussion of prior sensitivity

The second matter of debate concerns the sensitivity of BF to varying choices of its hyperparameters μ_δ and σ_δ^2. [Johnson, 2005, p. 689; Ríos Insua and Ruggeri, 2012, p. 197]

In order to comprehend the effect of π_δ on the BF, one should recall the same as a ratio of marginal likelihoods (see equation 3.14).
To quote Rouder et al. [2009, p. 229], the marginal likelihood is

> "[...] the weighted avergage of the likelihood over all constituent point hypotheses, where priors serve as the weight."

BF's particular prior densities - in question to have the effect of suchlike *weights* - are π_μ, π_{σ^2} and π_δ. However, π_μ and π_{σ^2} occur both in both marginal likelihoods under consideration. As such, their weighing effect is argued to largely diminish. [Rouder et al., 2009, p. 231] Finally, BF's dependence on prior densities is predominantly down to the - as the name implies - test-relevant prior π_δ.
One may conclude: The value of BF depends on the prior density π_δ inherent to H_1. As such, BF is sensitive to changes of the hyperparameters μ_δ and σ_δ^2 building up π_δ. [Rouder et al., 2009, p. 229]

BF's prior sensitivity may best be explained by a concrete, numerical example:
Let $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ be simulated sample sets drawn from $N(0, 1)$ and $N(0.5, 1)$, respectively. Let the sample sizes be $n_1 = n_2 = 10$. The hyperparameter μ_δ may be set to 0. Finally, σ_δ^2 may vary between 0 and 3 to demonstrate, how the value of BF_{10} changes due to different (subjective) σ_δ^2 - choices.

Figure 1 clearly illustrates, what is meant by *prior sensitivity* . For $\sigma_\delta^2 = 0$, H_0 is equal to H_1. Clearly, BF attains the value 1 in this case. As σ_δ^2 is increased, greater relative weight is placed on certain effect sizes under H_1. As $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ in fact stem from normal distributions with different population means, BF desirably prefers H_1 over H_0 from then on. In this graphic, the maximal BF value is attained under $\sigma_\delta^2 = 1.1$. In other words, $H_1 : \delta \sim N(0, 1.1)$ is the most preferable alternative hypothesis compared to $H_0 : \delta = 0$. Further increases of σ_δ^2 presume ever-growing, observed effect sizes. In this example, σ_δ^2 - increases beyond 1.1 constantly lower the comparative statistical evidence for H_1. Finally, one can see that BF is particularly sensitive to σ_δ^2 - changes between 0 and 0.5. Within this range, the value of BF increases fourfold.

Generally speaking, one may add the following: Once the chosen hyperparameter pair implies unreasonably large effects, this penalizes the marginal likelihood under H_1 and conversely lifts the relative support for H_0 within BF. [Rouder et al., 2009,

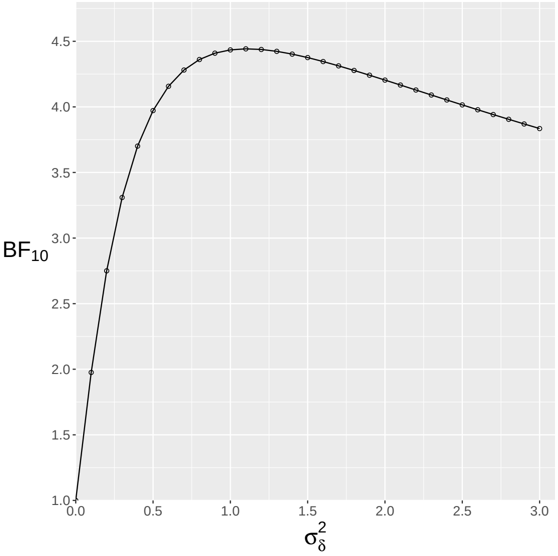

Figure 1: Plot of BF_{10} against σ_δ^2

p. 229] Finally, unrealistic choices for H_1 yield a BF, that provides unbounded support for H_0 over H_1. [Rouder et al., 2018, p. 105]
To sum it: The less realistic one chooses μ_δ and σ_δ^2, the farther will model-generated data under H_1 be apart from the empirical data and the more will BF comparatively favor H_0.

4.3.1 Reproving arguments on prior sensitivity

Oversensitivity to varying hyperparameter choices

In that regard, the major point of criticism is apparent: BF is declared critically sensitive to varying hyperparameter choices. [see Morey et al., 2016, p. 15; see Vanpaemel, 2010, p. 491; Kass and Raftery, 1995, p. 792] In reference to the above stated example, one might criticise that a change of σ_δ^2 by only 0.5 quadruples the value of BF. Especially in combination with prior uncertainty regarding the specification of π_δ, insufficient robustness is commonly deemed problematic. [Berger et al., 2012, p. 2]

4.3.2 Approving counterarguments

Adaptability to the reseach question

The counterargument thereon may be put straight, quoting Etz and Vandekerckhove [2018, p. 24]:

> ”[T]he answer we get naturally depends on the question we ask.”

That is to say: Yes, different priors entail different BF, but this is just in line with its Bayesian conception. [Morey et al., 2016, p. 15]

BF employs data to compare two contrasting beliefs, formalised as hypotheses, whereby π_δ co-specifies the belief under H_1. Contentwise, this makes π_δ part of establishing an alternative hypothesis with a precise marginal likelihood. Finally, BF responds to exactly that research question, stating that:

The sample $\boldsymbol{x}$ is BF times more likely to have been generated under H_0 than under H_1, *taking into account the prior* π_δ. [Liu and Aitkin, 2008, p. 363]

In this sense, different priors respond to (at least slightly) different *research questions.*[Morey et al., 2016, p. 15; Matthews, 2011, p. 848]

Accordingly, BF's sensitivity to π_δ equals sensitivity to the question asked - in turn being a favourable feature, not a weakness. Vanpaemel [2010, p. 491] even goes so far as to call the marginal likelihood an apt measure to compare scientific hypotheses precisely *because* of its sensitivity to the prior.

4.4 Discussion of prior implementation

The third matter of debate treats the actual implementation of π_δ and the constraints an analyst underlies at it.

As explained above, the specification of π_δ is rather crucial within the BF approach. Yet, a proper choice of π_δ demands for profound prior information and even then, the resulting BF would remain at risk of being neglected by others on the ground of subjectivity. To get around this, Bayesian analysts typically put noninformative prior densities to use. In several applications, this depicts a seemingly attractive way to state indifference and ignorance about unknown parameter values whilst lowering the influence of the prior on the sought outcome. However, this does not apply for the choice of π_δ. [Liseo, 2012, p. 202]
First of all, noninformative priors are usually improper, meaning they are defined merely up to an arbitrary constant. Employing an improper prior for π_δ would cause BF itself to be indeterminate. [Johnson and Rossell, 2010, p. 144]
However, as the multiple of an arbitrary constant, BF is no longer of any worth for hypothesis comparison. [De Santis and Spezzaferri, 1997, p. 504]
This categorically rules out improper priors for the specification of π_δ.
Above that mathematical concern, (very) diffuse prior distribtions are most often unfit to represent π_δ. [Aitkin, 1991, p. 113]
Per definition, the same spread probability mass over a large range of values, resulting in flat probability distributions. Recalling that π_δ represents an analyst's initial beliefs and uncertainty about the values of δ, a flat prior is unintuitive. Yet, one would thereby place almost equal weights on wholly unrealistic effect sizes as on small, plausible ones. [De Santis and Spezzaferri, 1997, p. 504; Morey et al., 2016, p. 16]
Finally, the heavy weighing of devious δ-values completely lowers the marginal likelihood under H_1 and in turn leads BF to an overstated support of H_0. [De Santis and Spezzaferri, 1997, p. 510]
Putting it drastically: Even if the data indicated a just meaningful effect, this might not become apparent from a BF comprising a noninformative π_δ. As the latter would pose H_1 that unrealistic, BF tends to prefer H_0, *comparatively.*

In consequence, BF analyses are mostly bound to proper, informative prior choices, where the respective information may stem from relevant test results, related literature and - for most parts - subjective knowledge. [Aitkin, 1991, p. 113]
In fact, not only the applied knowledge itself, but also the transformation of knowledge into a probability distribution is grounded on individual assessments of the analyst. [Kass and Raftery, 1995, p. 781; Matthews, 2011, p. 844]

4.4.1 Reproving arguments on prior implementation

(a) **The constraint to knowledgeability**

In this matter, critique first of all adresses the demand for knowledgeability, imposed on anyone, who wants to make use of a BF analysis. It condemns the penalty, the BF forces on vague choices for π_δ. In that regard, critics claim that knowledgeability by implication is questionable. [Liu and Aitkin, 2008, p. 367]
In most practical cases, analysts simply *are* unsure about the value of δ, often to a remarkable extent. The demand for an informative π_δ leaves them with pretty much two options. Either they back away from the BF approach in general or they pretend knowledgeability, they do not actually have. In the latter case, π_δ is at risk of being arbitrary, rather than well-founded.

(b) **Decline of scientific validity**

Still others feel uncomfortable with the subjectivity of π_δ as they worry how it diminishes the scientific validitiy of their research results. [Goldstein, 2006, p. 411; Matthews, 2011, p. 851; Rouder et al., 2018, p. 106]
Depending on individual knowlegde and differing information sources, researchers may legitimately hold different subjective prior assumptions about the value of δ. This in turn makes it fairly easy to refute another person's BF result just by advocating a different π_δ. One might therefore put BF's overall evidential worth into question.
Lastly, due to a subjectively chosen π_δ, scientists may constantly be accused of "engineer[ing] any result they wish". [Rouder et al., 2009, p. 233] The sheer possibility of adjusting the prior in favor of a desired finding, undermines the integrity of BF outcomes. [Rouder et al., 2018, p. 106]

(c) **Endorsement of scientific dicord**

Finally, prior subjectivity is being criticised to cause a number of researcher to draw different conclusions based on the same data.[Matthews, 2011, p. 851]
As a group of different researchers is very unlikely to advocate the exact same prior π_δ, there may co-exist a series of seemingly equally valid BF-values within the same research context. Such unsteadiness is claimed to be adverse for scientific consensus and generally approved findings. [see Goldstein, 2006, p. 407]

4.4.2 Approving counterarguments

(a) **The deficiency of noninformative priors**

By contrast, advocats of the subjective BF approach criticise the overall endeavor to use flat priors. [Matthews, 2011, p. 846; Morey et al., 2016, p. 16; Vanpaemel, 2010, p. 494; Gelman and Hennig, 2017, p. 969, 970]

On behalf of objectivity, many scientists try to avoid personal decisions and hide away from statistical procedures, that cannot be received solely from the data at hand. Flat, noninformative priors are said to tempt primarily by simplicity of use. After all, tuning decisions can hereby be handed over to some algorithm. [Gelman and Hennig, 2017, p. 969, 971]
However, noninformative priors pass over the theoretical meaning of π_δ and its hyperparameters. Concretely, μ_δ represents the mean of the standardized effect size δ and σ_δ^2 indicates, which effect size values are likely, which unlikely and which incredible under H_1 before data are seen. [Vanpaemel, 2010, p. 491]
Taking this into account, vague priors are nothing but unwise. On the one hand, they spread weight over improbable up to even impossible effect sizes. On the other, they are not apt to represent indifference or objectivity. [Morey et al., 2016, p. 16] With a flat prior, a scientist intends to act as if he was fully unsuspecting of possible effect sizes and as if no study had ever been made on a related issue. This is rather absurd. A scientist may be expected to know, which hypotheses he wants to compare. [Matthews, 2011, p. 849] Finally, this should at least enable him to rule out impossible δ-values. [Vanpaemel, 2010, p. 494]
Affirming this, Lindley [see critique of O'Hagan, 1995] once urged:

> "[I]t is better to think about [the parameter] and what it means to the scientist. It is his prior that is needed and not the statistician's. No one who does this has an improper distribution."

(b) **Reasonableness of prior knowlegdge**

Especially in a context, where the hyperparameters have such a clear meaning, scientists may be assumed to have some intuition and knowledge about them. Mean and variance are particularly intuitive variables in psychological research. Although prior knowledge may sometimes be not quite advanced, assuming complete ignorance would anyway underrate a researcher's ability. Finally, as prior knowledge is reasonable, informative priors are feasible and subjectivity is down to rational beliefs. It can primarily enrich and adjust the evalutation. [Vanpaemel, 2010, p. 493]

(c) **Validity through flexibility and context awareness**

Next up, proponents reverse the accusation that a subjective prior diminishes the validity of BF results. [Matthews, 2011, p. 849; Gelman and Hennig, 2017, p. 970, 975]
According to them, an informative prior distribution gives rise to flexibility and context dependence and thus rather *enhances* the validity of scientific results. After all, it allows a researcher to adapt the alternative hypothesis perfectly to his research question. He may tune μ_δ and σ_δ^2 with direct regard to the research context. [Gelman and Hennig, 2017, p. 969; Morey et al., 2016, p. 16] With σ_δ^2,

he may for instance contour the size of δ, that he consideres (just) meaningful. Obviously, this varies according to the respective research context. As to that, Vanpaemel [2010, p. 494] calls on researchers to embrace the prior, rather than treating it as a "nuisance necassary to get the Bayesian modelling machinery going".

(d) **Admission of multiple perspectives**

It is further approved, how subjective priors π_δ admit of multiple perspectives within a scientific community. [Goldstein, 2006, p. 410; Gelman and Hennig, 2017, p. 975]

As Gelman and Hennig [2017, p.975] expounds, "[...] reality and facts are accessible only through individual perspectives" Now, different researchers possess different knowledge, draw on various sources of information, hold different personal beliefs and carry out research in different contexts. With a subjective prior, different perspectives can be expressed and made transparent. Finally, (slightly) different BF results may even endorse scientific progress, when being composed as different perspectives on a common subject of study. [Morey et al., 2016, p. 15]

(e) **Endorsement of scientific communication**

Finally, advocats of the subjective prior defend the same against the accusation of causing discord and dismissal among researchers. In fact, they claim the opposite to be the case. Accordingly, researchers must be aware that hypothesis comparison is from ground up subjective. [Gelman and Hennig, 2017, p. 971] Like Rouder et al. [2009, p. 235] put it straight: "For any data set, the null will be superior to some alternatives and inferior to others." Nevertheless, specific choices have to be made in order to conduct scientific research. Having that in mind, researchers need to open for communication in any case. After all, the latter is every day practice in scientific work. Thereby, it makes no difference, whether experimental methods, valid interpretations or the appropriateness of π_δ are at issue. [Morey et al., 2016, p. 15]

Moreover, the choice of π_δ is usually revealed transparently. This benefits direct criticism.

In many cases, when individual notions do not contradict, boudaries my be set, on which a number of researchers agrees. If the respective interval is sufficiently narrow, this can be seen as a meaningful, scientific finding. If no consensus can be found, however, this is an equally reasonable outcome. It shows up scientific differences, once more endorsing scientific communication. Finally,

> "the view of negotiated alternatives is vastly preferable to the current practice, in which significance tests are mistakenly regarded as

> objective. [...] The sooner we adopt inference based on specifying alternatives, the better.”

[Rouder et al., 2009, p. 235]

4.5 Conclusion

Within this Bayesian framework, BF provides a logically sound evidence statement. Both the mathematical and the conceptual setup of BF are rather consistent with the general idea of scientific progress. Among others, the approach satisfies with a transparent handling of subjectivity and the potential to support the null hypothesis.

The controversy around the qualification of BF is largely confined to the test-relevant prior π_δ. In fact, π_δ is an integral component of BF. It imposes the precise specification of the hyperparameters μ_δ and σ_δ^2 on any analyst, willing to make use of BF outcomes. As default procedures are mostly ineligible, this makes the evaluation process a bit more complex. On the other hand, π_δ enables for an inclusion of relevant, data-external information and warrants great flexibility for context awareness.

BF is sensitive to changes of π_δ through varying hyperparameter choices. This may be seen both good and bad. Certainly, one may approve it as adaptability to fine changes in the research context. In order to cope with prior sensitivity, hyperparameter choices should nevertheless be reasoned, honestly displayed and kept open for discussion. Additionally, many scientists advise a routine sensitivity analysis. That is to say, BF should additionally be calculated over a range of other reasonable hyperparameter choices. [Liu and Aitkin, 2008, p. 364-366; Sinharay and Stern, 2002, p. 196]

Furthermore, the BF approach commonly demands an informative π_δ. This assumes a considerable amount of prior knowledge and relevant information on part of the analyst. One may appreciate the implied invitation to reflect on reasonable effect sizes, interesting results and the given context. Moreover, one may expect a scientist in psychological research to have certain, personal resources on the mean and the variance of an effect under study.

Doubts regarding BF are primarily concerned with the practical feasibility or at least the considerable efforts to achieve a reasonable π_δ. However, this prior co-determines BF's meaningfulness. [Pericchi and Walley, 1991, p. 3; Liu and Aitkin, 2008, p. 367; Goldstein, 2006, p. 407]
BF's benefits become apparent only on condition of profound prior information. However, they are yet less instrumental in the general case of moderate prior knowledge and even problematic, when relevant information is meager.

One may hereof claim: The distinctive deficiencies of BF first and foremost trace back to the demand for *precise* value assignments on π_δ's hyperparameters.

It is not the basic concept of BF and not the general inclusion of a subjective element like π_δ; it is the *demand for prior precision* that causes most of BF's deficiencies.
First, personal knowledge is in rarest cases directly transferable into a precise, numerical value or a unique distribution.
Second, prior knowledge is pretty much never determinate enough for justified unambigous value assignments of μ_δ and σ^2_δ. This promts arbitrary choices, misspecifications and amplifies discord among the scientific community. Researchers might agree on certain boundaries, but unlikely on an exact pair of values $(\mu_\delta, \sigma^2_\delta)$.
Third, BF is rather prone to disavowal and as such impedes generally accepted findings. In combination with prior sensitivity, accusations of adjusted results are often hard to turn away. [Rouder et al., 2009, p. 233]
Finally, the demand for prior precision may be stated as the major reason for researchers to account the "pain gain ratio" [Goldstein, 2006, p. 407] of BF to be too high.

To sum up, BF is a valid tool for two-sample comparisons to anyone, who approves the Bayesian conception of inference. It has groudedly become popular for a variety of preferable properties in view of hypothesis comparison.
The points, for which BF is groundedly criticised or backed away from, may for a large part be ascribed to the demand for a precise test-relevant prior π_δ.
Finally, this is where one could effectively draw on, when willing to enhance BF.

5 The Imprecise Bayes Factor - An enhancement proposal in the context of Imprecise Probabilities

The following chapter is designated to make an enhancement proposal for the conventional Bayes Factor for independent two-sample comparisons (BF) descibed as yet. The approach specifically draws on the problems BF entails concerning its strict demand for a single, precise prior distribution on the effect size parameter δ. Imbedded in the theory of Imprecise Probabilites, this restriction shall be eased. The generalised BF outcome will finally be titled the *Imprecise Bayes Factor* (IBF).

5.1 The enhancement target

The principal target in enhancing the conventional BF grounds on a problematic situation, in which many researchers find themselves, when willing to calculate a BF. Under the endeavour to ascertain a difference in group means, they are demanded to define prior distributions for all unknown parameters occuring in the hypothesis' model classes. The prior distributions on μ and σ^2 are argued not to be of noteworthy consequence for the BF result. Thus, they are confidently set according to Jeffreys' noninformative proposal (see 3.11). [Rouder et al., 2009, p. 231] However, BF is indeed sensitive to the choice of the test-relevant prior $\pi_\delta(\theta_\delta) = N(\mu_\delta, \sigma_\delta^2)$ and thus to the choice of its hyperparameters μ_δ and σ_δ^2. [Sinharay and Stern, 2002, p. 196]

Now, even though a reseacher most often has some applicable prior knowledge about the sought effect size δ, the same is (most) often not sufficient to decide on one, unique prior distribution $N(\mu_\delta, \sigma_\delta^2)$. [Goldstein, 2006, p. 418; Wolfenson and Fine, 1982, p. 80, Etz and Vandekerckhove, 2018, p. 7]

However, a suchlike specification presupposes "very detailed prior knowledge" [Augustin et al., 2014, p. 205] or rather "extremely definite beliefs" [Joyce, 2010, p. 285].

It is debatable, whether precise credences are ever attainable in practice and thus, whether a single prior distribution is ever justified to model uncertainty about δ under H_1. [see Berger, 1990, p. 305, 306, Joyce, 2010, p. 283; Pericchi and Walley, 1991, p. 3]

Notwithstanding this, prior knowledge is incomplete and contentious in many practical situations and if so, any precise choice can be accused of arbitrariness and are indeed quite likely to be misstated. [Goldstein, 2006, p. 418; Walley et al., 1996, p. 458] Accrding to Kass and Raftery [1995, p. 784],

> "[a]ny approach that selects a single model [...] leads to underestimation of the uncertainty about quantities of interest, sometimes to a dramatic extent."

Especially under BF's high level of prior sensitivity,

> "[...] 'overprecision' by too rigorous assumptions may destroy the practical relevance of the results obtained."

[Augustin et al., 2014, p. 146]

Beyond the problem of individual prior uncertainty, groups of researchers often need to achieve joint conclusions despite of indiviually diverse prior notions. [Berger, 1990, p. 304, 305] In such cases, single prior distributions are not only hard to agree upon, but also negligent of multiple perspectives.

Suchlike precarious combinations of prior uncertainty and prior sensitivity - which unfortunately apply to many cases in scientific practice - are regarded as a major drawback of Bayesian analyses long-since. [Berger et al., 2012, p. 2]
In order to antagonize the problematic nature of Bayesian analyses in that regard, procedures were developed to deal with its prior sensitivity in a state of partial prior knowledge, where it is usually difficult to assign prior distributions on unknown parameters. [see Pericchi and Walley, 1991; see Ríos Insua and Ruggeri, 2012]

The traditional approach to deal with the prior sensitivity of Bayes Factors in scientific practice is to submit an obtained Bayes Factor result to a so-called *Bayesian sensitivity analysis* [Walter, 2013, p. 40] or *Robust Bayesian analysis.* [Berger et al., 2012, p. 1]
Applied to the concrete case of BF, its basic concept goes as follows: In principle, one assumes the existence of a 'correct' prior distribution π_δ, which could ideally model prior uncertainty regarding the effect size δ and thus lead to an 'ideal' BF analysis. Due to the unfortunate case of incomplete prior knowledge, this distribution is not attainable. [Walley et al., 1996, p. 462]
Thus, BF is calculated over a class of individually plausible candidate prior distributions. Often, this class is conceived as a sort of "neighbourhood" [Pericchi and Walley, 1991, p. 1, 2] to one "central element" [Walter, 2013, p. 40] that is intended to be chosen for π_δ and whose single-valued BF outcome shall be primarily reported afterwards.
Finally, the major goal of a robust Bayesian analysis is to assess the sensitivity of a BF result to reasonable variations of the hyperparameters μ_δ and σ_δ in order to confirm the BF as a stable result or to transparently unfold how other prior choices would have lead to different BF results. [Berger et al., 2012, p. 1, 7]
However, robust analyses do not ease the analyst past the general dilemma to assign precise values to μ_δ and σ_δ^2 when lacking in sufficient prior knowledge. Still, any choice's adequacy remains unclear and unanimity among different researchers cannot be ensured. Finally, the one BF result selected for reports or further analyses still comprises the analyst's prior uncertainty in its entirety.

5.2 The enhancement proposal

The major target of an IBF analysis is to relax BF's deficient demand for a precise hyperparameter choice. It's major idea for enhancement draws on the theory of *Imprecise Probabilities* and may be broken down as follows:
The hyperparameters are defined as closed intervals instead of single values. This leads to a *subjective credal set* comprising infinitely many precise prior distribtutions instead of only one.

5.2.1 Imprecise Probabilities

Imprecise Probabilities aim at an adequate modelling of situations, where prior information is partial, prior beliefs are imprecise and thus, a precise specification of π_δ is problematic and ambiguous. [Augustin et al., 2014, p. 145, 146, 148]
Over time, Imprecise Probabilities have evolved into a full-fledged theretical framework to

> "[...] encompass and extend the traditional concepts and methods of probability and statistics by allowing for incompleteness, imprecision and indecision, and provide new modelling opportunities where reliability of conclusions from incomplete information is important."

[Augustin et al., 2014, p. xiii]

However, the overall, basic concept of Imprecise Probailities can be broken down to the rather simple and indeed natural idea of replacing a (hardly attainable or unjustified) precise probability measure $P(A)$ by an interval of probability measures $[\underline{P}(A); \bar{P}(A)]$. The latter spawns an inprecise set of probability measures, defined over $\underline{P}(A)$ and $\bar{P}(A)$ as the interval's lower and upper bound, respectively. [Walter, 2013, p. 33; Joyce, 2010, p. 281]
The systematic use of intervals rests upon the notion, that prior uncertainty can *never* be modelled adequately under the use of precise probabilities, parameter values or models. It is argued that the latter would certainly underrate the actual uncertainty associated with their results. [Kass and Raftery, 1995, p. 784; Walley et al., 1996, p. 462]
Consequently, statistical conclusions are submitted solely through lower and upper bounds. [Walley et al., 1996, p. 463]
As such, the theory of Imprecise Probabilities objects the classical, Bayesian interpretation of probability as stated in chapter 2.1.

In the following, one special approach within the framework of Imprecise Probabilities shall be employed to model prior uncertainty about the hyperparameters μ_δ and σ^2_δ.

5.2.2 The subjective credal set

As stated above, the IBF approach refers to situations, in which a researcher intends to calculate BF whilst being unsure about the choice of μ_δ and σ^2_δ required to specify the effect size prior. In accordance with his prior knowledge, he can at best locate the hyperparameters within certain value *ranges*. [cf. Joyce, 2010, p. 283]

Whereas a conventional BF analysis cannot incorporate incomplete prior knowledge, an IBF analysis aims at the flexible consideration of the latter. In order to explicitly model prior uncertainty, the IBF approach resorts to a common, imprecise-probabilistic tool, termed *credal set*. [Walter, 2013, p. 37; Augustin et al., 2014, p. 19]
Credal sets generally denote non-empty sets or classes of (precise) probability distributions, consistent with prespecified lower and upper bounds on the parameters contained in these distributions. [Walter, 2013, p. 37, 38; Walley et al., 1996, p. 462]

Transferred to the specific case of imprecise prior knowledge on μ_δ and σ^2_δ, a suchlike credal set may be constructed as follows:
Instead of deciding on one value pair, the researcher expresses his hyperparameter choices in terms of closed *intervals* $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma^2_\delta}$. These shall depict two ranges of μ_δ- and σ^2_δ- values, seperately accounted reasonable in the light of available prior knowledge.
Concretely, this requires the researcher to set 4 precise interval boundaries, which he shall choose according to the motto:
As narrow as possible and as broad as necessary, so that the width of the inteveral accurately reflects the amount of uncertainty in the prior choices.
Finally, his specifications are of the form

$$\mathcal{I}_{\mu_\delta} = [\underline{\mu}_\delta; \bar{\mu}_\delta] \quad \text{and} \quad \mathcal{I}_{\sigma^2_\delta} = [\underline{\sigma}^2_\delta; \bar{\sigma}^2_\delta]. \tag{5.1}$$

Thereby, $\underline{\mu}_\delta$ and $\bar{\mu}_\delta$ one by one depict the lower and the upper bound of the interval $\mathcal{I}_{\mu_\delta}$. Taken together, they are sufficient to repesent incomplete prior knowledge about μ_δ. The analog applies to the definition of $\underline{\sigma}^2_\delta$ and $\bar{\sigma}^2_\delta$.

Mathematically speaking, the IBF approach expands the hyperparameter specification from a single point $(\mu_\delta, \sigma^2_\delta)$ in the two-dimensional space to a rectangular hyperparameter area $[\underline{\mu}_\delta, \bar{\mu}_\delta] \times [\underline{\sigma}^2_\delta, \bar{\sigma}^2_\delta]$, comprising infinitely many, potential value pairs.

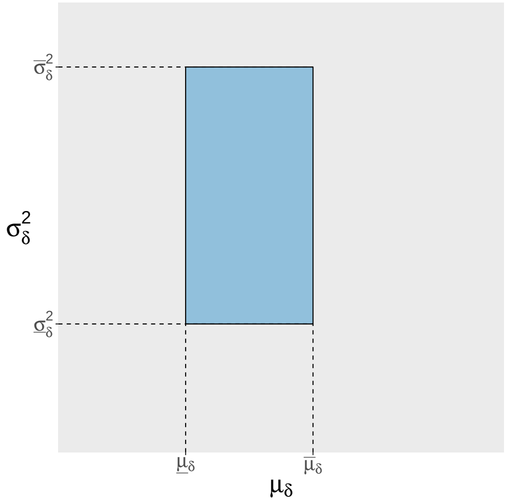

Figure 2: Rectangular area of hyperparameter choices

Based thereon, the credal set $\mathcal{M}$ may be defined as

$$\mathcal{M} = \{N(\mu_\delta, \sigma_\delta^2) : \ \mu_\delta \in \mathcal{I}_{\mu_\delta}, \ \sigma_\delta^2 \in \mathcal{I}_{\sigma_\delta^2}\} \tag{5.2}$$

or in the style of Berger and Sellke [1987, p. 115],

$$\mathcal{M} = \{\text{all normal distributions } N(\mu_\delta, \sigma_\delta^2) : \ \underline{\mu}_\delta \leq \mu_\delta \leq \bar{\mu}_\delta, \ \underline{\sigma}_\delta^2 \leq \sigma_\delta^2 \leq \bar{\sigma}_\delta^2\}. \tag{5.3}$$

Verbalized, $\mathcal{M}$ depicts a closed set of all normal prior distributions, that match the analyst's interval-valued prior belief in the values of μ_δ and σ_δ^2. [cf. Augustin et al., 2014, p. 147]
As the applied lower and upper bounds were specified merely upon subjective prior knowledge and personal beliefs, $\mathcal{M}$ shall explicitly be termed a *subjective* credal set.

In sum, the IBF approach asks the analyst to specify two hyperparameter intervals $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma_\delta^2}$ instead of two precise hyperparameter values μ_δ and σ_δ^2. Thereby, it counteracts his dilemma to make precise, subjective decisions in a state of prior uncertainty and instead enables for a flexible, more realistic expression of partial knowledge. [cf. Wolfenson and Fine, 1982, p. 80]

By now, IBF's final outcome may be determined.
Consequently, the following chapter is dedicated to the accomplishment, definition and interpretation of an IBF result.

5.3 The Imprecise Bayes Factor

By now, imprecise prior knowledge is modelled as a set of infinitely many, precise prior distributions.

However, the principal target of an IBF analysis remains to make conclusions about statistical evidence in favor of H_0 against H_1 or vice versa in the light of a sample $\boldsymbol{x}$.

As a imprecise-probabilistic method, $\mathcal{M}$ is considered as an "entity of its own" [Walter, 2013, p. 40] As no single prior distribution is believed to yield a justified BF result, no particular meaning is given to individual prior distributions and nothing is concluded from any single BF result. Instead, $\mathcal{M}$ yields an interval of BF results that is expressed solely over lower and upper bounds. [Walley et al., 1996, p. 462]

Finally, this intervall may be introduced as the *Imprecise Bayes Factor*.

5.3.1 Derivation & Definition

In concrete terms, the same depicts an interval of multiple Bayes Factors, bounded by the maximal and the minimal BF value derivable from the elements of the credal set $\mathcal{M}$.

$$IBF_{01}(\boldsymbol{x}, \mathcal{I}_{\mu_\delta}, \mathcal{I}_{\sigma_\delta^2}) = \left[\min_{\substack{\mu_\delta \in \mathcal{I}_{\mu_\delta} \\ \sigma_\delta^2 \in \mathcal{I}_{\sigma_\delta^2}}} BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma_\delta^2);\ \max_{\substack{\mu_\delta \in \mathcal{I}_{\mu_\delta} \\ \sigma_\delta^2 \in \mathcal{I}_{\sigma_\delta^2}}} BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma_\delta^2) \right] \tag{5.4}$$

As the IBF is defined over two BF values, which seperately take values $(0, \infty)$, the same value range applies to IBF's upper and lower bound. For reasons of clarity, IBF's bounds may denoted as $\underline{BF}$ and $\overline{BF}$ hereafter.

Finally, the following relation may be inferred:

$$IBF_{01} = \left[\underline{BF}_{01};\ \overline{BF}_{01} \right] \quad \Leftrightarrow \quad IBF_{10} = \left[(\overline{BF}_{01})^{-1};\ (\underline{BF}_{01})^{-1} \right] \tag{5.5}$$

5.3.2 Calculation

In order to calculate an IBF, the sample data set $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}\}$ (see definition 3.1) and the subjectively specified hyperparameter intervals $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma_\delta^2}$ are demanded as input. In order to process an IBF result, the two BF extrema over $\mathcal{M}$ need to be identified. Indeed, $\mathcal{M}$'s closure property warrants their finite existence. However, their detection requires the calculation of (theoretically infinitely many) BF values, resultant from each and every prior distribution inherent to $\mathcal{M}$.

In the hypothetical case of a finite credal set, a suchlike endeavour could be met through an application of the *Generalized Bayes' Rule.*
According to Walter [2013, p. 39], the same enables to coherently "transfer the basic aspects of traditional Bayesian inference to the generalised [imprecise] setting". In the case of a given credal set $\mathcal{M}$, it purports:

> "The prior credal set $\mathcal{M}$ is updated element by element to obtain the posterior credal set [...] cosisting of all posterior distributions [...] obtained by traditional Bayesian updating of elements of the prior credal set"

[Walter, 2013, p. 39]
Now, instead of gradually updating an imprecise prior distribution to an imprecise posterior distribution via Bayes' Rule, one might also deploy each hyperparameter combination $(\mu_\delta, \sigma_\delta)$ in $\mathcal{M}$ - once at a time - to calculate a set of corresponding BFs and finally set the maximum and the minimum value as seeked interval boundaries.

However, $\mathcal{M}$ - as defined in 5.2 - in fact contains infinitely many prior distributions. On that account, IBF's interval boudaries cannot be calcuated in practice.
Instead, the latter turns into a numerical optimization problem. One might herefore apply the `R`-function `optim()` and optimize the function $BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma^2_\delta)$ (see 3.15) over μ_δ and σ^2_δ, simultaneously. One optimization process may thusly lead to the seeked minimum and a second may serve to yield the maximum. (see Appendix A)

5.3.3 Interpretation

Just like its credal set $\mathcal{M}$, an IBF result is to be interpreted as one cohesive entity. That is to say, an IBF result is interpreted in terms of its lower and upper bounds. One does explicitly not regard to single BF values contained in the interval so to declare the imprecise-theoretic notion that none of them is reasonable for themselves.

A generally valid interpretation of IBF_{01} might then read as: Given the subjective credal set $\mathcal{M}$ as an expression of imprecise belief regarding π_δ's hyperparameters μ_δ and σ_δ, the sample $\boldsymbol{x}$ is at least $\underline{\mathrm{BF}}_{01}$ and at most $\overline{BF}_{01}$ times as much statistical evidence for H_0 as for H_1.
Rather colloquially speaking, one might say: If we locate μ_δ somewhere between $\underline{\mu}_\delta$ and $\bar{\mu}_\delta$ and assume σ^2_δ to lie within $\underline{\sigma}^2_\delta$ and $\bar{\sigma}^2_\delta$, then we would expect to obtain a BF_{01} within a range of $\underline{\mathrm{BF}}_{01}$ and $\overline{BF}_{01}$.

To sum it: One does no longer state comparative statistical evidence in precise terms, but merely delimits the same to an interval of a certain size. As such, one makes a less distinct, but on the other hand more cautious and transparent, comparative evidence statement. The same is apt to give a rough, inclusive impression about BF results in the light of prior uncertainty.

5.3.4 Arguable conclusions

Arguable conclusions from an IBF result indeed demand for some deliberation. However, the same are not always quite straighforward.

In fact, if $\underline{\text{BF}}_{01}$ is greater than 1, one might conclude that $\boldsymbol{x}$ consistently favors H_0 over H_1, albeit to varying degrees. If $\overline{BF}_{01}$ remains below 1, just the opposite may be inferred. Otherwise, an IBF explains how the comparative evidence remains somewhat ambigious, given the extent of prior imprecision. In such cases, any single BF result could be disputed on grounds of unjust precision.

As stated above, an IBF might also result from a collection of differing prior beliefs held by a research group. In such cases, one may conclude to what extent these differences are relevant in the sense of being reflected in the overall BF variation.

In any case, one may incorporate the size of the IBF interval relative to that of $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma^2_\delta}$ into his conclusions. However, the former stongly depends on the size of the hyperparameter intervals used to represent the uncertain inputs. [cf. Berger et al., 2012, p. 9, cf. Etz and Vandekerckhove, 2018, p. 27] Of course, the expressiveness and clarity of conclusions implies reasonably narrow IBF intervals. If the latter are too broad to be conclusive, there are two options that can be taken so as to yield a satisfactory result. One is to narrow the hyperparameter intervals further and the other is to collect additional data. Of course, a more detailed specification of prior knowledge needs to be justified by an attainment of certain, additional prior information. 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'. "

In accordance with Goldstein [2006, p. 418], one might claim that even then "the value of an incomplete analysis overweighs possible missspecifications and wrong results". Of course, *overly* vague, ambigious intervals are no longer informative. [cf. Walter, 2013, p. 50]

5.4 A simulated application example

Hereafter, an exemplary IBF analysis shall be performed on the basis of a simulated data set. The hypothetical research question shall hereby adapt to a relevant topic in the field of psychological research. Primary reference applies to the article *Sex similarities and differences in risk factors for recurrence of major depression*, published in the journal *Psycholgical Medicine.* [van Loo et al., 2017]

According to a number of studies, women are approximately twice as likely to experience major depression (MD) than men. [see van Loo et al., 2017, p. 1695; Noel-Hoeksema, 2001, p. 173] However, it is rather uncertain, whether this gender difference persists after disease onset. Previous studies turned out controversial results. [see van Loo et al., 2017, p. 1695]

In a loose reference thereto, the following, hypothetical research situation may be imagined:
A scientist in psychological research wants to pursue the question, whether women are more likely to experience a *recurrence* of MD than men. He formulates his research question as: *Is the overall risk of a MD recurrence different in both sexes?* Or rather: *Is there a gender effect within the recurrence risk of MD?*
The overall risk of recurrence may captured by a score, calculated over a number of different risk predictors. [cf. van Loo et al., 2017, p. 1687-1689] In this illustrative example, it may be assumed that the score results are normally distributed. As the results may stem from a preliminary inquiry, they are standardised with the men's group mean and the pooled standard deviation.
For the purpose of a statistical analysis, the researcher may draw on a (simulated) sample data set $\boldsymbol{x}$ composed of two independent, normally distributed group samples $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$.
Thereby, $\boldsymbol{x_1}$ includes the scores of 10 men and $\boldsymbol{x_2}$ comprises the score values of 10 women. In both cases, the data are true, fictitious values, drawn from two different normal distributions. (see Appendix A)
In order to examine a potential gender difference in the average recurrence risk, the researcher intends to conduct a Bayesian analysis.
Consequently, he constructs two contrasting hypotheses of the form

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \sim N(\mu_\delta, \sigma_\delta^2). \tag{5.6}$$

Whereas the null hypothesis states equal, average recurrence risk for women and men, the alternative implies a normally distributed effect size δ around the mean μ_δ and with a variance of σ_δ^2.

Up to this point, all requirements are met for the researcher to carry out a conventional BF analysis. (see chapter 3.4)
In order calculate a BF, the latter finally expects him to specify precise hyperparameter values μ_δ and σ_δ^2 based on his personal prior knowledge. (see definition 3.15)

After all, the researcher lacks a sufficient knowledge base to confidently make precice choices in that regard. On that account, he decides to make use of BF's enhancement proposal, the IBF approach, for which he proceeds as follows:

In order to yield a credal set $\mathcal{M}$ of prior distributions for δ under H_1, he specifies two hyperparameter intervals. According to his prior knowledge, he chooses lower and upper bounds for μ_δ and σ_δ^2, each. Given H_1, he expresses his assumptions regarding μ_δ in form of the interval

$$\mathcal{I}_{\mu_\delta} = [0; 0.5]$$

and localizes σ_δ^2 - indicative for the expected effect range - within the interval

$$\mathcal{I}_{\sigma_\delta^2} = [0.5; 3].$$

Thereby, he constructed an area of potential hyperparameter pairs, visualisable as:

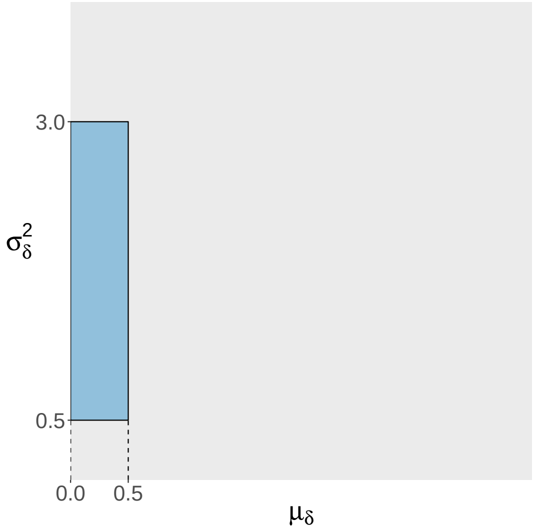

Figure 3: Subjectively specified hyperparameter area

In accordance with his previous interval specifications, the subective credal set $\mathcal{M}$ holds as

$$\mathcal{M} = \{\text{all } N(\mu_\delta, \sigma_\delta^2) \text{ distributions} : \ 0 \leq \mu_\delta \leq 0.5, \, 0.5 \leq \sigma_\delta^2 \leq 3\}.$$

Finally, the corresponding IBF_{10} is defined as

$$IBF_{10}(\boldsymbol{x}, [0; 0.5], [0.5; 3]) = \left[\min_{\substack{\mu_\delta \in \, [0;0.5] \\ \sigma_\delta^2 \in [0.5;3]}} BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma_\delta^2); \max_{\substack{\mu_\delta \in \, [0;0.5] \\ \sigma_\delta^2 \in [0.5;3]}} BF_{01}(\boldsymbol{x}, \mu_\delta, \sigma_\delta^2) \right] \tag{5.7}$$

By means of an appropriate, numerical optimization algoritm, the minimal and the maximal BF_{10}-value may be computed successively. (see section 5.3.2)
Using the R-function `optim()`, the IBF interval yields as

$$IBF_{10}(\boldsymbol{x}, [0; 0.5], [0.5; 3]) = [1.836817; \ 5.994383].$$

Finally, the researcher is in a position to say that $\boldsymbol{x}$ is about 1.8 to 6 - times as much statistical evidence for H_1 than for H_0. Taking into account his prior uncertainty about δ under H_1, the BF consistently favors H_1 over H_0. Merely the preference degree varies, so that the value of upper bound is about 3-times that of the lower. He may conclude or report: The data $\boldsymbol{x}$ imply 1.8 to 6-times more statistical evidence for a certain gender effect than for equal recurrence risk of MD in both sexes.

To allow for a visualisation of the interval boundaries' accomplishment, the hyperparameter intervals $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma^2_\delta}$ may be discretised into respective segments (at intervals of 0.05 and 0.1 hereafter).
Consequently, $\mathcal{M}$ comprises only a finite number of normal distributions, based on which just as many BF values may be calculated and plotted in form of a heatmap.

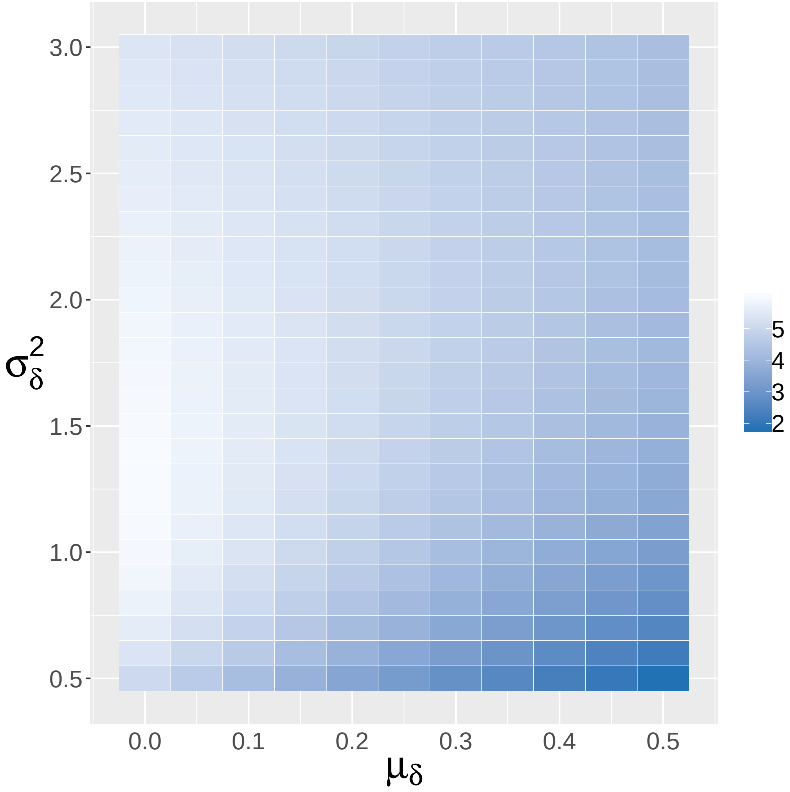

Figure 4: Heatmap of discretised BF results

Finally, the IBF interval boundaries are the minimum and the maximum BF value occuring in this heatmap. Their values are 1.836817 and 5.994381, respectively. As such, the discretised bounds correspond to the optimized up to the sixth decimal point.
As highlighted through the white grid, these BF values have been calculated discretely. Yet, the graphical representation suggests the assumption of BF's continuity regarding the hyperparameters μ_δ and σ^2_δ. A mathematical proof exceeds the scope of this thesis, but as continuity nevertheless stands to reason, a smoothed heatmap is displayed .

5.5 An enhancement record

Eventually, IBF's overall capacity to enhance BF shall be evaluated. As to that, the following key points provide a general view over IBF's major advantages over its conventional counterpart.

(a) **Awareness of context dependence**

First of all, the IBF approach - as a direct generalisation of the BF approach - cleaves to the notion, that subjective prior knowledge is a gain to statistical analyses. Just like BF, it applies personal beliefs to yield the *subjective* credal set $\mathcal{M}$. As such, it equally prompts the researcher to think about reasonable hyperparameter values. Even more, it promts him to specifiy $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma^2_\delta}$ *as narrow as possible*. Finally, IBF similarly encourages the analyst to use his avaiable background information in order to raise context dependence and adjust the analysis to his intended research question. [cf. Gelman and Hennig, 2017, p. 973- 975]

(b) **Increase of practicability**

The IBF offers a pragmatic alternative to the infeasible, time-consuming specification of one precise hyperparameter pair $(\mu_\delta, \sigma_\delta)$ in situations, where prior knowledge is partial. This eases the burden, usually imposed on the analyst. In order to compute an IBF, the same is no longer ought to the unpleasant situation of decision making, when precise judgements lack a sufficient knowledge base. Instead, IBF exclusively demands for practical assessments. Finally, this makes IBF-analyses more attractive and qualified for common use.

(c) **Reduction of error proneness & Increase of real-world correspondence**

Compared to BF results, IBF intervals are less error prone. Obviously, the risk of misstating hyperparameter intervals of an arbitrary, finite size is lower than that of being wrong about two precise values. Even more, $\mathcal{M}$ is much likelier to contain a prior that matches with the real world situation. The case that this holds true for a single distribution $N(\mu_\delta, \sigma^2_\delta)$ can practically be ruled out. As such, IBF is a better approximation of the real world situation it refers to. Finally, the reduced danger of misspecifiactions makes IBF results more meaningful and harder to confute.

(d) **Extension of interpretability**

An IBF result affords the analyst an extended, overall impression of comparative, statistical evidence. After all, it provides insight about the overall range of BF values deemed reasonable according to respective prior beliefs. This range may likewise result from different, personal notions or from prior uncertainty. Based on IBF's interval size, one may reflect about BF's overall robustness against differing hyperparameter assumptions or individual uncertainty. However, the IBF implies a sensitivity analysis, just under a different rationale.

(e) **Flexible consideration of prior uncertainty**

The IBF approach can flexibly adapt to the available amount of prior information. It can cope with situations, where the latter is sparse and at the same time it does not impede thorough prior knowledge. [cf. Augustin et al., 2014, p. 145] To sum it: Applying the IBF approach, one can explicitly model any (partial) degree of prior knowledge. [cf. Augustin et al., 2014, p. 158] As such, IBF simply expands the BF approach by loosening its restrictions on prior specification.

(f) **Awareness of multiple perspectives & Encouragement of consensus**

The conventional BF is often criticised to counter general agreement. However, conclusions based on a certain prior $N(\mu_\delta, \sigma_\delta^2)$ can be discounted as irrelevant for anybody who would have chosen another. [cf. Gelman and Hennig, 2017, p. 989]

The fact that IBF incorporates prior knowledge in form of intervals, enables for a union of multiple, individual prior notions about μ_δ and σ_δ^2. As such, a number of different researchers may arrive at and agree on one joint IBF result. [cf. Berger, 1990, p. 304] Indeed, this is no longer a definite conclusion, but considering that "[m]ultiple perspectives are a reality to be reckoned with and should not be hidden" [Gelman and Hennig, 2017, p. 975], a range of BF values states an eligible compromise between the ideal of overall agreement and the reality of debatable, different opinions.

The composition of common hyperparameter intervals in turn provides a good opportunity for scientific reasoning and communication about reasonable hyperparameter values. [cf. Gelman and Hennig, 2017, p. 975]

(g) **Encourament of cautiousness & Transparency**

An IBF result as such itself requests the analyst to draw cautious conclusions. It demands that any evidence statement is expressed with reference to the resective prior imprecision. This makes IBF conclusions less over-precise and withal more honest. [cf. Augustin et al., 2014, p. 145]

Through $\mathcal{I}_{\mu_\delta}$ and $\mathcal{I}_{\sigma_\delta^2}$, prior assumptions are laid out transparently. [cf. Gelman and Hennig, 2017, p. 989] This makes it easy for different researchers to equalize the latter with their own beliefs and thus decide whether they want to support the obtained result. [Goldstein, 2006, p. 409]

(h) **Automized applicability**

Finally, the IBF can easily be implemented in an application software. The computation of respective interval boundaries can be achieved through variant optimization algorithms, applied to the closed-form BF function introduced by Gönen et al. [2005, p. 253]. The programming language R, for instance, offers the optimization algorithms `optim()` or `optimize()` for this purpose. For the practical use in psychological reseach, one may thus imagine an interactive software package, which requests the 4 interval boundaries for μ_δ and σ_δ from the analyst and processes them into an interval-valued IBF output.

6 Conclusion

On the whole, this thesis was dedicated to the Bayes Factor as the result of Bayesian hypothesis comparison. In particular, it intended to present, discuss and enhance the Bayes Factor approach, referred to as the "Bayesian two-sample-t-test".

This section summarizes the most important results and clarifies the strengths and limitations of the proposed Imprecise Bayes Factor.

Chapter 1 introduced the general Bayes Factor as a component of the odds notation of Bayes' Rule. It presented the same as the ratio of two marginal likelihoods. The Bayes Factor was interpreted as the amount of evidence, a sample $\boldsymbol{x}$ holds in favor of one scientific hypothesis against another. Furthermore, the Bayes Factor was exposed as a relational and relative measure of statistical evidence. As such, its explanatory power depends on the meaningfulness of the considered hypothesis set and any valid conclusions remain comparative. By itself, the Bayes Factor is inapt for hypothesis selection or absolute hypothesis evaluations.

In chapter 2 presented the "Bayesian two sample t-test" as an exceedingly common, statistical problem in psychological research. This special case holds a facile, closed-form for BF calculations. In fact, Gönen et al. developed a formula, in which BF is dependent only on $\boldsymbol{x}$ and the two hyperparameters of the normally distributed effect size prior, representing H_1. This enables for a simple calculation and a high ease of use for non-statisticians. However, facile applicability is not the only reason for the increased popuarity of the BF approach.

As stated in chapter 3, proponents endorse BF's clear interpretative framework and the great flexibility to adapt H_1 to the research question and the scientific context. Moreover, they approve BF's comparative nature, which allows to gain comparative evidence in favor of H_0.

The controversy around the BF is largely confined to the test-relevant prior π_δ. Misgivings primarily hit the practical feasibility and the additional efforts to specify a reasonable prior distribution $\pi_\delta = N(\mu_\delta, \sigma_\delta^2)$ over a reasonable pair of hyperparameters. Doubts adress the necessity for a subjective specification of π_δ as well as BF's (high) sensitivity to varying hyperparameter choices. The discussion in chapter 3 turned out that BF's demand for prior precision may largely be deplored. It was reasoned that precise prior knowlegde is rarely attainable in scientific practice. However, this prompts arbitrary hyperparameter choices, causes misspecifications and thus boosts the risk of meaningless BF results and wrong conclusions. Finally, a relaxation of the demand for prior precision was declared as the chief working point for BF enhancement.

The Imprecise Bayes Factor proposed in chapter 5 is the direct response thereto. It enables the analyst to specify the hyperparameters in form of intervals, whose lengths represent subjective prior uncertainty. Through a corresponding credal set, the IBF approach explicitly models partial prior knowledge. This generalisation increases the feasibility of π_δ-specifications in scientific practice, reduces error-proneness and enables for an inclusion of multiple perspectives. As the resulting IBF interval is considered and interpreted as an entity of its own, cautious, more realistic conclusions are encouraged. Finally, an IBF result is likelier to contain the prior distribution, that matches the real word situation.

A final, concluding remark may read as follows: The IBF enhances the conventional BF in situations, where prior knowledge does not allow for precise prior specifications. The generalisation focused on a realistic dealing with uncertainty to reach more honest and stable results. It raises awareness of subjectivity, prior uncertainty and the reality of different prior notions. Yet, these strengths are at cost of a more vague, possibly ambigious statements of comparative evidence. Too broad hyperparameter intervals might lead to cumbersome, uninformative results. As to that, the respective analyst needs to deem an interval a satisfactory outcome.

As yet, the IBF applies to one special case of application. Extensions to other, more general research questions are certainly conceivable. However, these would commonly involve a much more complex computation of lower and upper bounds.

A Digital Supplement

This Bachelor thesis involves a digital supplement, which consists of the folder `Bachelor_Thesis_IBF`. It contains the `R`-code to reproduce all figures and calculations reported in the thesis.

References

M. Aitkin. Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:111–142, 1991.

T. Augustin, F. Coolen, G. de Cooman, and M. C. Troffaes, editors. *Introduction to Imprecise Probabilities.* John Wiley & Sons, 2014.

J. Berger, D. R. Insua, and F. Ruggeri. Bayesian robustness. In D. Ríos Insua and F. Ruggeri, editors, *Robust Bayesian Analysis.* Springer Science & Business Media, 2012.

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

J. O. Berger and M. Delampady. Testing precise hypotheses. *Statistical Science*, pages 317–335, 1987.

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

J. Cohen. *Statistical power analysis for the behavioral sciences. 2nd.* Hillsdale, N.J.: L.Erlbaum Associates, 1988.

F. De Santis and F. Spezzaferri. Alternative Bayes factors for model selection. *Canadian Journal of Statistics*, 25:503–515, 1997.

A. Etz and J. Vandekerckhove. Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, 25:5–34, 2018.

R. J. Fox and M. W. Dimmic. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, 7:n.pag., 2006.

C. Gallistel. The importance of proving the null. *Psychological Review*, 116:439–453, 2009.

A. Gelman and C. Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180:967–1033, 2017.

M. Gönen, W. O. Johnson, Y. Lu, and P. H. Westfall. The Bayesian two-sample t test. *The American Statistician*, 59:252–257, 2005.

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

V. E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67:689–701, 2005.

V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:143–170, 2010.

J. M. Joyce. A defense of imprecise credences in inference and decision making 1. *Philosophical Perspectives*, 24:281–323, 2010.

R. E. Kass. Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42:551–560, 1992.

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

P. R. Killeen. An alternative to null-hypothesis significance tests. *Psychological Science*, 16:345–353, 2005.

M. Lavine and M. J. Schervish. Bayes factors: What they are and what they are not. *The American Statistician*, 53:119–122, 1999.

B. Liseo. Robustness issues in Bayesian model selection. In D. Ríos Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*. Springer Science & Business Media, 2012.

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

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

J. I. Marden. Hypothesis testing: From p values to Bayes factors. *Journal of the American Statistical Association*, 95:1316–1320, 2000.

W. 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.

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

S. Noel-Hoeksema. Gender differences in depression. *Current Directions in Psychological Science*, 10:173–176, 2001.

A. O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:99–138, 1995.

L. R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, 59:1–23, 1991.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

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

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

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

R. Royall. The likelihood paradigm for statistical evidence. In M. l. Taper and S. R. Lele, editors, *The nature of scientific evidence: Statistical, philosophical, and empirical considerations.* University Chicago Press, 2010.

S. Sinharay and H. S. Stern. On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56:196–201, 2002.

R. Van De Schoot, S. D. Winter, O. Ryan, M. Zondervan-Zwijnenburg, and S. Depaoli. A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22:217–239, 2017.

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

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

P. Walley, L. Gurrin, and P. Burton. Analysis of clinical data using imprecise prior probabilities. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45:457–485, 1996.

G. Walter. *Generalized Bayesian inference under prior-data conflict.* PhD thesis, Ludwig Maximilians Universität München, 2013.

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

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

M. Wolfenson and T. L. Fine. Bayes-like decision making with upper and lower probabilities. *Journal of the American Statistical Association*, 77:80–88, 1982.

Statutory declaration

I declare that I have developed and written the enclosed Bacelor Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. This thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Signature Date