Patrick M. Schwaferts

Improving Practical Relevance of Bayes Factors

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 21.10.2021

Patrick M. Schwaferts

Improving Practical Relevance of Bayes Factors

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

1. Berichterstatter: Prof. Dr. Thomas Augustin
2. Berichterstatter: Prof. Dr. Volker Schmid
3. Berichterstatter: Prof. Dr. Martin Spieß

Eingereicht am 21.10.2021
Disputation am 21.02.2022

Zusammenfassung

Bayes-Faktoren werden neuerdings als Ersatz für die stark kritisierten frequentistischen Hypothesentests propagiert. Allerdings beruhen Bayes-Faktoren häufig auf denselben statistischen Hypothesen, die in den frequentistischen Verfahren verwendet und wegen mangelnder praktischer Relevanz kritisiert wurden. Um Bayes-Faktoren vor ähnlichen Unzulänglichkeiten zu bewahren, wird in der vorliegenden Dissertation versucht, herauszuarbeiten, wie die praktische Relevanz von Bayes-Faktoren verbessert werden kann. Es zeigt sich, dass eine formale Definition des Begriffs der praktischen Relevanz innerhalb der statistischen Entscheidungstheorie zu finden ist. Die Relevanz eines Ergebnisses hängt natürlich davon ab, wofür es verwendet wird, und eine solche Verwendung ist - formal gesehen - eine Entscheidung. Dementsprechend wurden die Bayes-Faktoren im Rahmen der Bayes'schen Entscheidungstheorie dargestellt und bewertet, wobei die Spezifikation der Verlustfunktion das größte Hindernis für ihre Anwendung zu sein scheint. Typischerweise sind die Informationen über Konsequenzen einer Entscheidung knapp, vage und mehrdeutig, was eine eindeutige und präzise Spezifikation der Verlustfunktion nahezu unmöglich macht. Um dieses Spezifikationsproblem zu lösen, werden zwei Möglichkeiten diskutiert: Erstens kann die Verlustfunktion durch die Verwendung eines hypothesenbasierten Ansatzes vereinfacht werden, und zweitens können die geforderten Spezifikationen mengenwertig, d.h. verallgemeinert, anstelle von präzisen Werten aufgefasst werden. In diesem Sinne wurde eine zweifache Verallgemeinerung der Bayes-Faktoren in die Entscheidungstheorie und in das Feld der verallgemeinerten Wahrscheinlichkeiten entwickelt und anschließend in einen anwenderfreundlichen statistischen Leitfaden verdichtet. Außerdem wurde das Wesen von statistischen Hypothesen kritisch bewertet, wobei gezeigt wurde, dass sie - im Gegensatz zur gängigen Auffassung in der Literatur über Bayes-Faktoren - lediglich Teilmengen des Parameterraums sind.

Die vorliegende kumulative Dissertation besteht aus **acht veröffentlichten Beiträgen**, die jeweils unterschiedliche Aspekte dieses Themas abdecken. Zusammen behandeln diese Beiträge, wie die praktische Relevanz von Bayes-Faktoren verbessert werden kann, indem sie die dazu notwendigen konzeptionelle Grundlagen (Definition der praktischen Relevanz, Natur von statistischen Hypothesen), methodologische Grundlagen (zweifache Verallgemeinerung von Bayes-Faktoren) und Methoden für Anwendungen (benutzerfreundliche Schritt-für-Schritt-Anleitungen, Vergleich mit anderen Methoden) ausarbeiten.

Abstract

Bayes factors are recently promoted as replacement for the heavily criticized frequentist hypothesis tests. Yet, Bayes factors are oftentimes based on the same statistical hypotheses that were employed in frequentist procedures and criticized for lacking practical relevance. To guard Bayes factors of similar shortcomings, the present dissertation attempts to elaborate on how to improve the practical relevance of Bayes factors. It appears that a formal definition of the notion of practical relevance is located within the framework of statistical decision theory. The relevance of a result naturally depends on what it is used for, and – formally speaking – such a use is a decision. Accordingly, Bayes factors were depicted and evaluated within the framework of Bayesian decision theory, in which the specification of the loss function seems to be the major obstacle to its application. Typically, information about the consequences of a decision are scarce, vague, partial, and ambiguous, prohibiting an unambiguous specification of the loss function. To deal with these specification issues, two options are discussed: First, the loss function can be simplified by employing a hypothesis-based account and, second, the required specifications can be allowed to be set-valued, i.e. imprecise, instead of precise values. In this regard, a twofold generalization of Bayes factors into the framework of decision theory and into the framework of imprecise probabilities was developed and condensed into a straightforward framework for applications. Besides, the nature of statistical hypotheses was critically evaluated, showing that – in contrast to the current conception within the literature of Bayes factors – they are merely subsets of the parameter space.

The present cumulative dissertation thesis consists of **eight published contributions**, each delineating different topics within this framework. Together, these contributions elaborate on how to improve the practical relevance of Bayes factors by providing conceptual foundations (definition of practical relevance, nature of statistical hypotheses), methodological foundations (twofold generalization of Bayes factors), and methodologies for applications (user-friendly step-by-step guides, comparison to other methods).

Acknowledgments

I want to thank my Ph.D. supervisor **Prof. Dr. Thomas Augustin**, the examination committee Prof. Dr. Volker Schmid, Prof. Dr. Martin Spieß, Prof. Dr. Christian Heumann, and Prof. Dr. Annika Hoyer, my colleagues Hannah Blocher, Dr. Eva Endres, Dr. Paul Fink, Cornelia Fütterer, Dr. Christoph Jansen, Dominik Kreiss, Malte Nalenz, Aziz Omar, Dr. Julia Plass, and Dr. Georg Schollmeyer, the backbone of our institute Elke Höfner and Brigitte Maxa, as well as Luisa Ebner, Prof. Dr. Scott Ferson, Prof. Dr. Timo von Oertzen, Aaron Schaal, Christian Schwaferts, and Irma Talić!

I want to dedicate this thesis to my brother Andreas.

Preface

When I started this thesis in September 2017, I secretly hoped that I could finish it without diving into statistical decision theory. Apparently, it did not work out. This is science. It is not about what I want or what I do not want. It is about what is correct. If I need statistical decision theory to understand relevant concepts, I must not object to it.

Tutzing, October 2021

Patrick

Statement of Author's Contributions

I prepared the initial drafts of the decision theoretic contributions submitted to ISIPTA (Schwaferts and Augustin, 2019, 2021c). TA and I critically discussed, edited, and finalized the initial submission and the revised version after peer review.

LE elaborated on the robust Bayes factor in her Bachelor's thesis under the supervision of TA and me. LE and I prepared the initial draft of the contribution about the robust Bayes factor submitted to ISIPTA (Ebner et al., 2019). LE, TA, and I critically discussed, edited, and finalized the initial submission. TA and I critically discussed, edited, and finalized the revised version after peer review.

I wrote the technical reports (Schwaferts and Augustin, 2020, 2021e) and the preprints (Schwaferts and Augustin, 2021d,a,b). TA provided minor revisions.

TA: Prof. Dr. Thomas Augustin
LE: Luisa Ebner

ISIPTA: International Symposium on Imprecise Probability: Theories and Applications

Contents

Part A

Overall Discussion

Chapter 1

Introduction

1.1 Goal of the Dissertation

Statisticians and methodologists develop statistical methods and empirical scientists employ them. To use them, certain essential quantities need to be specified properly by the empirical scientist, such that they correctly reflect their counterpart within the applied field of science. How to specify these quantities properly is one of the fundamental issues at the interface of statistics and applied science: On the one hand, the applied scientist who seeks methodological guidance wants to know how to specify these quantities properly. On the other hand, the statistician who develops statistical methods assumes that these quantities are specified properly, as their specification is an applied, not a statistical problem, and depends on the actual field of applied research. As a consequence, the body of scientific elaborations about this issue is quite small compared to the amount of methodological developments or the amount of applied studies. Nevertheless, a correct specification of the essential quantities required within statistical methodologies is of paramount importance. If these quantities are misspecified, results will inform past the research question of interest. In severe cases, incorrect results might render the scientific endeavor as a waste of time and money.

This general lack of methological guidance on how to specify the essential quantities properly might lead to an increasing use of default specifications, a trend in applied sciences with a very long and striking history. In the case of frequentist hypothesis tests, default specifications of the employed hypotheses have become that omnipresent that critics refer to their use as a mindless ritual (Gigerenzer, 2004) called Null-Hypothesis-Significance-Testing (NHST). Naturally, mindlessly analyzing empirical data does impair the value scientific results and the integrity of science. While characteristics of NHST have been criticized for decades (with critique dating back to Berkson, 1938), critique intensified (e.g.

Dienes and Mclatchie, 2018; Kruschke and Liddell, 2018) in the context of the replicability crisis (Ioannidis, 2005).

In line with the critique against frequentists methods, especially against NHST, and the increasing popularity of Bayesian statistics (see e.g. van de Schoot et al., 2017), a massive amount of effort (see e.g. Kass and Raftery, 1995; Gönen et al., 2005; Rouder et al., 2009) was put into the development of Bayes factors – a Bayesian quantity in the context of hypothesis comparisons (dating back to Jeffreys, 1961). Now, Bayes factors are argued to replace frequentist hypothesis tests.

However, also the calculation of the Bayes factor requires certain essential quantities to be specified properly in accordance with the research problem of interest. While discussions and developments about these issues are still ongoing, one might find all sorts of different opinions about it, ranging from default specifications (see e.g. Rouder et al., 2009; Ly et al., 2016) to an emphasis on individual specifications that capture the research context (see e.g. Rouder et al., 2018a; Dienes, 2019; Gronau et al., 2019). In the course of the last decades, there are also some researchers, who changed their view on this topic (cp. e.g. Rouder et al., 2009, 2018b), underlining that the current development of Bayes factors is far from being completed.

There are, however, two aspects that are predominantly missing in this development of Bayesian methods for hypothesis comparisons:

- While it is easy to see that misspecifying essential quantities might lead to irrelevant results, an exact formal definition of a relevant result is missing. Yet, by developing the notion of *practical relevance* on a formal level, mathematical concepts and frameworks are opened up, which are required to ensure the relevance of a result. These concepts and frameworks might then be incorporated into the statistical methods in the context of Bayes factors.
- Essential quantities are typically required to be precisely specified. To do so, the applied scientist needs to have quite a high level of certainty about the relevant information for this specification. However, information is rarely that abundant to allow unambiguous precise specifications. The applied topic under investigation is still not fully known, else it might not be subject to a scientific endeavor. In that, it suggests itself to refrain from requiring the applied scientists to provide precise specifications and allowing rather imprecise[1] specifications that include the available (partial) information and uncertainty as is.

[1] A conventional conception is that the mathematical formalization, which is used to elaborate on a certain phenomenon of interest in a scientific way, is merely an idealization of the actual real world characteristics. In that sense, one might argue that it is only natural to have a certain (ideally small) degree of misspecification within the employed mathematical formalization. Yet, a distinction has to be made:

– A misspecification might arise because the phenomenon of interest in the real world is quite compli-

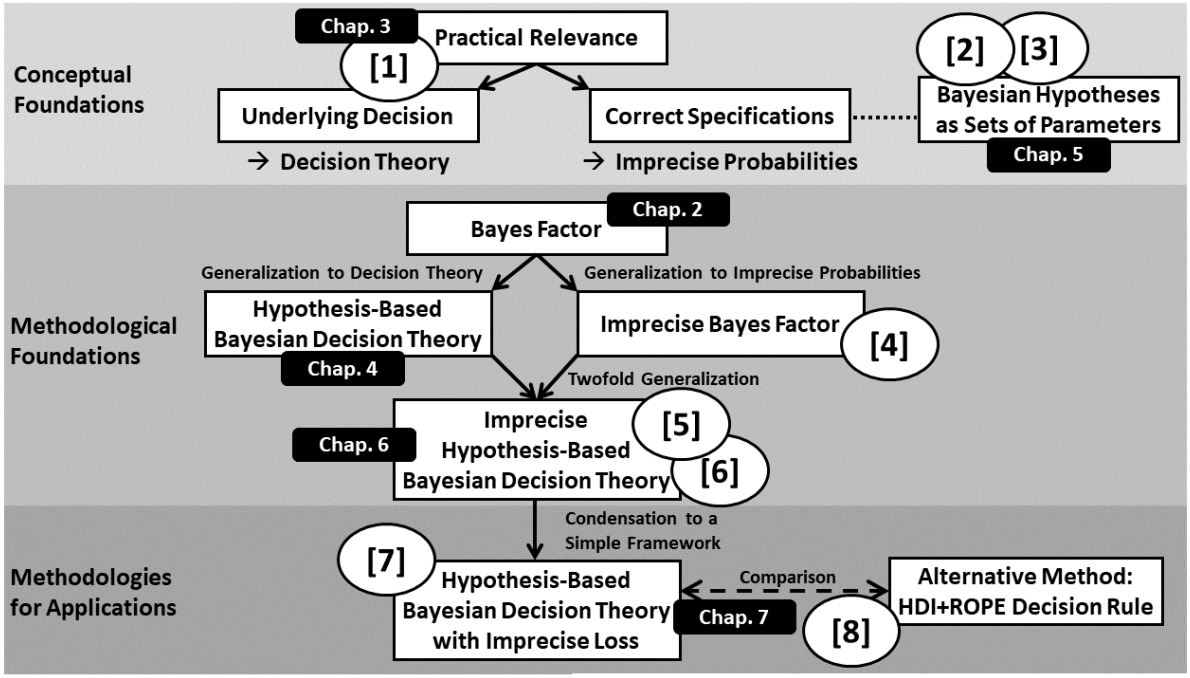

Figure 1.1: Thesis Overview. The elaborations within this dissertation thesis can be divided coarsely into conceptual foundations (light gray), methodological foundations (gray), and methodologies for applications (dark gray). The main concepts (squared boxes) and their relations (arrows) are shown, to which all contributions of this thesis (circles) are related to. The numbers of the contributions relate to the listing in Section 1.2. Black rounded boxes denote the chapters within the following elaboration.

Following the first aspect, it appears that decision theoretic concepts are required to formally define the notion of practical relevance (Schwaferts and Augustin, 2021d). To follow the second aspect, imprecisely specified quantities and their integration into statistical methods is the primary matter of the field of imprecise probabilities. Therefore, in order to further develop the existing Bayesian methods for hypothesis comparisons (especially Bayes factors) w.r.t. these two neglected aspects, their generalization into both the framework of decision theory (see e.g. Berger, 1985; Robert, 2007) and the framework of imprecise

cated and the researcher decides to employ a simpler (and therefore idealized) formalization, such that a scientific treatment of the phenomenon of interest becomes feasible.

– A misspecification might rise because knowledge about the phenomenon of interest in the real world is scarce, vague, partial, and ambiguous, rendering different possible formalizations plausible, such that choosing any of these formalizations might be considered arbitrary.

These two types of misspecifications are completely different! While the context of idealization requires sufficient information about the phenomenon of interest to argue that such an idealization is appropriate, the context elaborated on in this dissertation thesis assumes that information is – in general – insufficiently available for precise specifications. Accordingly, precise specifications cannot be treated as idealization if the available information is insufficient. They would rather constitute arbitrary choices, emphasizing the need to use imprecisely specified quantities to formalize partial information.

probabilities (see e.g. Walley, 1991; Augustin et al., 2014) is crucial.

Working towards this two-fold generalization is the goal of this dissertation thesis, thereby allowing Bayes factors to be employed in a more practically relevant manner (Figure 1.1).

1.2 Overview of Contributions

This dissertation thesis consists of eight contributions[2]. These are listed with regard to a logical structure of their content (Figure 1.1):

[1] Schwaferts P. and Augustin T. (2021d). Practical relevance: A formal definition. URL `http://arxiv.org/abs/2110.09837`

[2] Schwaferts P. and Augustin T. (2021e). Updating consistency in Bayes factors. Technical Report 236, Ludwig-Maximilians-University Munich, Department of Statistics. URL `http://dx.doi.org/10.5282/ubm/epub.75073`

[3] Schwaferts P. and Augustin T. (2021a). Bayes factors can only quantify evidence w.r.t. sets of parameters, not w.r.t. (prior) distributions on the parameter. URL `http://arxiv.org/abs/2110.09871`

[4] Ebner L., Schwaferts P., and Augustin T. (2019). Robust Bayes factor for independent two-sample comparisons under imprecise prior information. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 167–174. PMLR. URL `http://proceedings.mlr.press/v103/ebner19a.html`

[5] Schwaferts P. and Augustin T. (2019). Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh*

[2] Besides these contributions, two additional scientific works were published that do not constitute this dissertation thesis:

- Hilbert S., McAssey M., Bühner M., Schwaferts P., Gruber M., Goerigk S., and Taylor P.C.J. (2019). Right hemisphere occipital rTMS impairs working memory in visualizers but not in verbalizers. *Scientific Reports*, 9(1):1–8. URL `http://dx.doi.org/10.1038/s41598-019-42733-6`
- Schwaferts C., Schwaferts P., von der Esch E., Elsner M., and Ivleva N.P. (2021). Which particles to select, and if yes, how many? subsampling methods for Raman microspectroscopic analysis of very small microplastic. *Analytical and Bioanalytical Chemistry*, 413(14):3625–3641. URL `http://dx.doi.org/10.14459/2021mp1596628`

International Symposium on Imprecise Probability: Theories and Applications, volume 103 of *Proceedings of Machine Learning Research*, pages 338–345. PMLR. URL http://proceedings.mlr.press/v103/schwaferts19a.html

[6] Schwaferts P. and Augustin T. (2021c). Imprecise hypothesis-based Bayesian decision making with composite hypotheses. In A. Cano, J. De Bock, E. Miranda, and S. Moral, editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, page 280–288. PMLR. URL https://proceedings.mlr.press/v147/schwaferts21a.html

[7] Schwaferts P. and Augustin T. (2021b). How to guide decisions with bayes factors. URL http://arxiv.org/abs/2110.09981

[8] Schwaferts P. and Augustin T. (2020). Bayesian decisions using regions of practical equivalence (ROPE): Foundations. Technical Report 235, Ludwig-Maximilians-University Munich, Department of Statistics. URL http://dx.doi.org/10.5282/ubm/epub.74222

The fist contribution (Schwaferts and Augustin, 2021d) develops a formal definition of the notion of practical relevance. To do so, previous elaborations on the concept of practical significance (see e.g. Kirk, 1996, 2001) were assessed and it appeared that all elaborations had two common characteristics: An applied decision that should be guided and – implicitly employed – a way to determine the optimal action for each possible effect. Formally, these characteristics relate to a decision problem with all potential actions and a loss function in the context of statistical decision theory (see e.g. Berger, 1985; Robert, 2007). Naturally, the practical relevance of some results depends on what the results are used for, and this use is formalized as a decision problem. In that, it appears that the notion of practical relevance is a decision theoretic concept. The contribution distinguishes between two different, but connected concepts of practical relevance: The practical relevance of an *effect* and *hypotheses* which incorporate the notion of practical relevance. With these definitions, one might employ the context of decision theory to understand the requirements of specifying hypotheses reasonably in an applied study.

Unfortunately, this understanding of how to specify hypotheses in a practically relevant manner does not match with the current elaborations about hypothesis specifications in the Bayes factor literature. While the previously developed decision theoretic elaboration states that a reasonable hypothesis specification depends on the potential actions and the loss function of the underlying decision problem only, the current conception of hypotheses in the context of Bayes factors relates them to the Bayesian prior distribution (cp. e.g. Vanpaemel, 2010; Vanpaemel and Lee, 2012; Morey et al., 2016; Rouder et al., 2018a; Dienes, 2019; Tendeiro and Kiers, 2019). However, this prior distribution formalizes only prior knowledge (or uncertainty or degrees of belief or information), but not a hypothesis,

about the effect of interest. How can this disagreement be explained? Where are potential shortcomings?

The second contribution (Schwaferts and Augustin, 2021e) is a groundwork to investigate these questions. In the context of Bayesian statistics, a prior distribution gets updated by the data to a posterior distribution. It is typically assumed that – if specified correctly – the prior distribution reflects all available knowledge before the data were obtained, such that the posterior distribution reflects all available knowledge after the data were obtained. In that, the posterior distribution might be employed as a prior distribution for the assessment of a second data set (of the same structure). Naturally, the final posterior distribution should be identical whether both data sets are considered subsequently or merged first and considered at once. If so, updating is called consistent, else it is called inconsistent. This contribution provides a proof that updating with Bayes factors is consistent. This detailed depiction of updating with Bayes factors emphasizes that it is important to consider that the complete prior distribution gets updated by observing the data, not only parts of it. If only some parts of the prior distribution get updated, updating is inconsistent.

The third contribution (Schwaferts and Augustin, 2021a) uses these insights to discuss the notion of a statistical hypothesis. In the context of Bayes factors, there are two hypotheses being contrasted against each other. Each of those hypotheses formalizes a theory, which are contrasted against each other in the research question. The goal of the scientific investigation is to assess the quality of these theories, i.e. to alter their plausibility or even to dismiss one of the theories in favor of the other. In such a contrasting setting, it is incorrect to change the theories themselves by seeing the data, only their plausibility should be adapted. If the contrast of the initially stated theories is of interest, changing them does not inform the research question. Based on these considerations about the nature of statistical hypotheses and the technical elaborations of the previous contribution, the question about the relation between the prior distribution and statistical hypotheses can be answered: Prior distributions change by seeing the data, statistical hypotheses must not change by seeing their data (only their plausibility should), so prior distributions cannot be used to formalize the theory that should be reflected by the statistical hypothesis. This conclusion argues against recommendations in the Bayes factor literature (cp. e.g. Vanpaemel, 2010; Vanpaemel and Lee, 2012; Morey et al., 2016; Rouder et al., 2018a; Dienes, 2019; Tendeiro and Kiers, 2019), but shows that the previously elaborated decision theoretic concept of statistical hypothesis is not inherently contradicting – despite disagreements with the current literature. A clear definition of and distinction between the employed mathematical concepts is paramount.

The fourth contribution (Ebner et al., 2019) starts tackling generalization of Bayes factors into the framework of imprecise probabilities by allowing imprecisely specified prior distributions. The generalized Bayes factor is then called robust Bayes factors. The elaboration was based on a common, widely used case within the framework of Bayes factors: Normally distributed data, where the mean parameter has a normal prior distribution. The hyper-

parameters of this normal prior were now allowed to be interval-valued, such that they can better incorporate the (potentially available) uncertainty inherent to the Bayesian prior specification. The resulting interval-valued robust Bayes factor does no longer sweep this uncertainty under the carpet and results are closer to the applied context.

The fifth contribution (Schwaferts and Augustin, 2019) elaborates on the generalization of hypothesis-based Bayesian decision theory into the framework of imprecise probabilities (see e.g. Walley, 1991; Augustin et al., 2014). In order to employ the framework of hypothesis-based Bayesian decision theory, the applied scientist has to specify hypotheses, a prior distribution, and a loss function. Typically, information about these three quantities is partial, such that their unambiguous precise specification is not possible. Forcing the applied scientist to make a commitment to precise values might result in arbitrary specifications. In that, it is better to allow for imprecise specifications of hypotheses, prior distributions, as well as loss functions. To keep the notation within appropriate limits, this contribution was restricted with its generalization into the framework of imprecise probabilities to simple hypotheses, i.e. hypotheses which hypothesize only a single parameter value.

The sixth contribution (Schwaferts and Augustin, 2021c) extends this generalization of hypothesis-based Bayesian decision theory into the framework of imprecise probabilities to composite hypotheses. This is crucial for its employment in applied studies, as they typically employ composite hypotheses. Again, an imprecise specification of (now composite) hypotheses, prior distributions, and loss functions is allowed.

The seventh contribution (Schwaferts and Augustin, 2021b) delineates such a hypothesis-based Bayesian decision theoretic account in a user-friendly manner and illustrates the involvement of Bayes factors therein. Also, an extra focus was put on how to use (pre-existing) Bayes factors in a subsequent, more comprehensive decision theoretic account. While the loss function is a component of a decision theoretic account, the hypotheses and the prior distribution are components of every hypothesis-based Bayesian analysis, especially Bayes factors. In that, it seems that a precise loss specification might be the biggest obstacle in applying a decision theoretic framework. Therefore, this contribution restricts itself to employ only an imprecise loss function (with precise hypotheses and a precise prior distribution). Nevertheless, imprecisely specified hypotheses or prior distributions might still be included into the hypothesis-based Bayesian decision theoretic account, as they might be included into analyses with Bayes factors. A special feature of the framework with imprecise loss is that it also allows for results stating that information is insufficient to yield reliable conclusions. In that, no precision will be pretended which is not available.

The eighth contribution (Schwaferts and Augustin, 2020) deals with another method in the context of Bayesian hypothesis-based analyses, that allows for results to be indecisive: A decision rule that compares Bayesian highest density intervals (HDI) with regions of practical equivalence (ROPE), and is therefore referred to as HDI+ROPE decision rule

(Kruschke, 2015, 2018). This contribution establishes the decision theoretic foundation of the HDI-ROPE decision rule and shows that it is not grounded on imprecise specifications of relevant quantities. Instead the employed decision theoretic quantities seem to be rather artificially specified than to be connected with an underlying applied decision problem. Consequently, it seems that by using a hypothesis-based Bayesian decision theoretic account with imprecise specifications one might obtain results that are closer to the applied context, i.e. more practically relevant, than by using the HDI+ROPE decision rule.

In summary, it seems that with the first three contributions a lot of foundational work (Figure 1.1, light gray) was done within the field of methodologies, especially Bayes factors, before the two-fold generalization of Bayes factors towards the decision theoretic framework and towards the framework of imprecise probabilities was tackled (Figure 1.1, gray). This, however, is crucial to connect the status quo in the context of Bayes factors with the provided generalizations. Without this connection to the available literature about Bayes factors, the two-fold generalization might be of little use to the applied scientist. The resulting framework and its condensed version for applications (Figure 1.1, dark gray) might then seem as something completely different, although, in fact, it is merely an advancement of methods already employed.

The following elaborations are structured as follows (Figure 1.1, black boxes). First, the mathematics of Bayes factors are outlined (Chapter 2) and special emphasis was put on the decomposition of the prior distribution in the context of hypotheses. By defining the notion of practical relevance on a formal level (Chapter 3), a connection was established between hypotheses, the practical relevance of research results, and statistical decision theory. This connection was elaborated on mathematically in the context of hypothesis-based Bayesian decision theory (Chapter 4), pointing out that the employment of hypotheses corresponds to an assumption on the loss function (Section 4.2). The nature of hypotheses was then compared to current conceptions about hypotheses in the literature about Bayes factors. Their disagreement was assessed and solved (Chapter 5). With a unified understanding about hypotheses, the hypothesis-based Bayesian decision theoretic framework was generalized into the context of imprecise probabilities (Chapter 6). An easy and straightforward framework was delineated in detail that might be seen as the next development from the current status quo about Bayes factor in the literature (Chapter 7). With this new understanding about Bayes factors in the context of decision theory, the interpretation of Bayes factors are finally evaluated w.r.t. their practical usefulness (Chapter 8).

Chapter 2

Bayes Factors

2.1 Prior Situation

In the context of Bayesian statistics, the data $x = (x_1, ..., x_n)$ is typically assumed to be parametrically distributed according to the density $f(x|\theta)$, where $\theta \in \Theta$ is the parameter of interest and Θ is the parameter space. There is a Bayesian prior distribution with density $\pi(\theta)$ on the parameter that formalizes the available knowledge[1] about the parameter. Bayes rule

$$\pi(\theta|x) = \frac{f(x|\theta) \cdot \pi(\theta)}{f(x)}\,, \tag{2.1}$$

with

$$f(x) = \int_\Theta f(x|\theta) \cdot \pi(\theta)\, d\theta \tag{2.2}$$

being the marginal density of the data x, dictates how to update the prior density $\pi(\theta)$ into the posterior density $\pi(\theta|x)$ once the data x were observed. As it is assumed that the prior distribution formalizes all the relevant knowledge about the parameter before the data were observed, the posterior distribution formalizes all relevant knowledge about the parameter after the data were observed. In that sense, results are derived solely from the posterior distribution in a Bayesian analysis.

In the context of Bayes factors, there need to be two statistical hypotheses h_0 and h_1, that are contrasted against each other. These hypotheses

$$h_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad h_1 : \theta \in \Theta_1 \tag{2.3}$$

[1] There are also other interpretations of Bayesian parameter distributions than knowledge (e.g. in Jaynes, 2003), e.g. uncertainty (e.g. in Kruschke, 2015) or degrees of belief (e.g. in Jeffreys, 1961) or information (e.g. in Berger, 1985). Within this dissertation thesis, the exact interpretation of the parameter distribution is irrelevant, so the term *knowledge* was arbitrarily chosen to denote the interpretation of parameter distributions. Nevertheless, the other possible interpretations might instead be employed as well.

are subsets $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ of the parameter space that are typically assumed to be a non-overlapping partition: $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. Unless stated otherwise, this assumption will be maintained in the following. Based on these two hypotheses, the prior density $\pi(\theta)$ might be used to calculate both the prior probabilities

$$p(h_0) = \int_{\Theta_0} \pi(\theta)\, d\theta \tag{2.4}$$

$$p(h_1) = \int_{\Theta_1} \pi(\theta)\, d\theta \tag{2.5}$$

of the hypotheses and the within-hypothesis prior parameter densities

$$\pi_0(\theta) := \pi(\theta|h_0) = \frac{1}{p(h_0)} \cdot \pi(\theta) \cdot \mathbb{1}(\theta \in \Theta_0) \tag{2.6}$$

$$\pi_1(\theta) := \pi(\theta|h_1) = \frac{1}{p(h_1)} \cdot \pi(\theta) \cdot \mathbb{1}(\theta \in \Theta_1)\,, \tag{2.7}$$

where $\mathbb{1}(s)$ is the indicator function that equals $\mathbb{1}(s) = 1$ if the statement s is true and $\mathbb{1}(s) = 0$ if s is false. With these quantities the initial overall prior density $\pi(\theta)$ can be decomposed as (cp. also Rouder et al., 2018b)

$$\pi(\theta) = p(h_0) \cdot \pi_0(\theta) + p(h_1) \cdot \pi_1(\theta)\,. \tag{2.8}$$

An illustration of such a prior decomposition is provided in Figure 2.1.

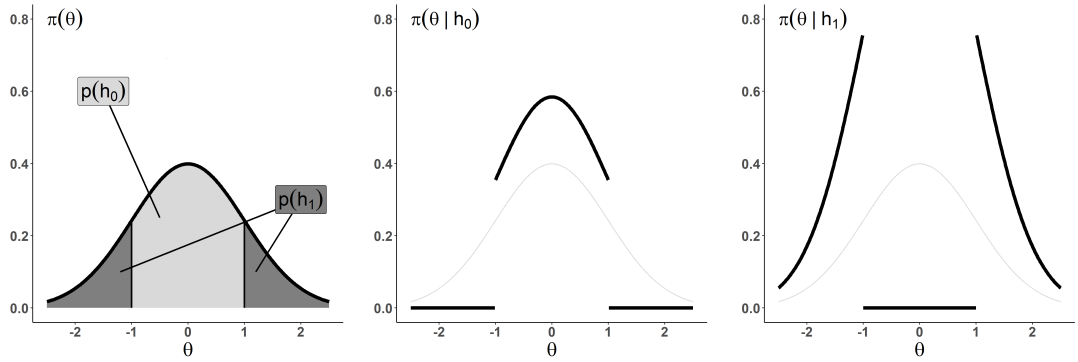

Figure 2.1: Prior Decomposition. Assume the parameter θ is distributed according to a standard normal distribution $N(0, 1)$ and the hypotheses are defined by $\Theta_0 = [-1, 1]$ and $\Theta_1 = (-\infty, -1) \cup (1, \infty)$. The left plot depicts the overall prior density $\pi(\theta)$ and the prior probabilities $p(h_0)$, $p(h_1)$ of the hypotheses. The middle and right plot depict the within-hypothesis prior densities $\pi_0(\theta)$ and $\pi_1(\theta)$, respectively. This figure was taken from (Schwaferts and Augustin, 2021b) with an adapted notation.

In that, it is possible to specify the prior situation for an analysis with Bayes factors in two different ways:

- Specify the overall prior density $\pi(\theta)$ and the hypotheses as subsets Θ_0, Θ_1 of the parameter space.
- Specify the hypotheses as subsets Θ_0, Θ_1 of the parameter space and supplement them with specifications of their prior probabilities $p(h_0)$, $p(h_1)$ and the within-hypothesis prior parameter densities $\pi_0(\theta)$, $\pi_1(\theta)$.

Both ways of specification are equivalent as they provide exactly the same amount of information. The respective unspecified information might by obtained by equations (2.4), (2.5), (2.6), (2.7) or by equation (2.8), respectively. Naturally, this equivalence depends on the previously made assumption[2] that the hypotheses form a non-overlapping partition of the parameter space.

2.2 Posterior Situation

Applying Bayes rule (2.1), the prior densities and probabilities get updated by observing the data x to the within-hypothesis posterior parameter densities

$$\pi_0(\theta|x) = \frac{f(x|\theta) \cdot \pi_0(\theta)}{f(x|h_0)} \tag{2.9}$$

$$\pi_1(\theta|x) = \frac{f(x|\theta) \cdot \pi_1(\theta)}{f(x|h_1)}\,, \tag{2.10}$$

where

$$f(x|h_0) = \int_{\Theta_0} f(x|\theta) \cdot \pi_0(\theta)\, d\theta \tag{2.11}$$

$$f(x|h_1) = \int_{\Theta_1} f(x|\theta) \cdot \pi_1(\theta)\, d\theta \tag{2.12}$$

[2]Without this assumption the second type of specification contains more information than the first type of specification, as there might be multiple different specifications of the second type that lead to the same specification of the first type. This, however, requires the hypotheses to be overlapping, a case that is typically considered problematic in the Bayes factor literature (cp. Morey and Rouder, 2011). If the true parameter values lies within the overlapping part, the Bayes factor value will never approach ∞ or 0 (i.e. lead to unambiguous evidence (cp. also Rouder et al., 2018b, p. 105)) even if the sample size will be increased infinitely (in this context compare also Jeffreys (1961, e.g. p. 269) arguments for his decisions on default prior distributions, especially the Cauchy distribution on the normal mean: It seems that he tried to avoid the possibility of unambiguous evidence at all costs.). Interpreting this characteristic, one might say that the scientific setting allows for result that are barely capable of distinguishing between the two theories of interest that are contrasted against each other in the research question. It might be argued that such a setting is suboptimal for optimizing the information gain of the scientific investigation and it is recommended to rethink the investigational design, such that non-overlapping hypotheses might be employed.

are the marginal densities of the data x under the respective hypotheses h_0, h_1, and to the posterior probabilities

$$p(h_0|x) = \frac{f(x|h_0) \cdot p(h_0)}{f(x)} \tag{2.13}$$

$$p(h_1|x) = \frac{f(x|h_1) \cdot p(h_1)}{f(x)} \tag{2.14}$$

of the hypotheses, respectively.

The overall posterior density $\pi(\theta|x)$ can, again, be depicted as a decomposition

$$\pi(\theta|x) = p(h_0|x) \cdot \pi_0(\theta|x) + p(h_1|x) \cdot \pi_1(\theta|x) \tag{2.15}$$

of the within-hypothesis posterior parameter densities (equations (2.9) and (2.10)) and the posterior probabilities (equations (2.13) and (2.14)). Naturally, the calculation of the overall posterior density via this hypothesis-based decomposition (equation (2.15)) yields the same density as simply updating the overall prior distribution $\pi(\theta)$ with Bayes rule (equation (2.1)).

Inspecting this hypothesis-based Bayesian update in greater detail, one might relate the probabilities of the hypotheses with each other. The ratio $p(h_0)/p(h_1)$ of the prior probabilities forms the prior odds and the ratio of the posterior probabilities forms the posterior odds

$$\frac{p(h_0|x)}{p(h_1|x)} = \frac{\frac{f(x|h_0)\cdot p(h_0)}{f(x)}}{\frac{f(x|h_1)\cdot p(h_1)}{f(x)}} = \frac{f(x|h_0)}{f(x|h_1)} \cdot \frac{p(h_0)}{p(h_1)}\,, \tag{2.16}$$

where

$$BF_{01} := \frac{f(x|h_0)}{f(x|h_1)} \tag{2.17}$$

is defined as the Bayes factor[3]. Therefore, the posterior odds

$$\frac{p(h_0|x)}{p(h_1|x)} = BF_{01} \cdot \frac{p(h_0)}{p(h_1)} \tag{2.20}$$

can be calculated by simply multiplying the prior odds with the Bayes factor.

[3]The index 01 indicates that the hypothesis h_0 (numerator) is compared to the hypothesis h_1 (denominator). Naturally, this comparison might equivalently be the other way round with

$$BF_{10} := \frac{f(x|h_1)}{f(x|h_0)} \tag{2.18}$$

and

$$BF_{10} = \frac{1}{BF_{01}}\,. \tag{2.19}$$

Note, that it is not possible to employ improper prior parameter distributions if the Bayes factor value is of interest. With improper prior distributions, the marginal densities $f(x|h_0)$ or $f(x|h_1)$ (equations (2.11) or (2.12)) are unbounded, such that the Bayes factor cannot be calculated properly. However, if the posterior odds are of interest, it is possible to employ an improper prior density $\pi(\theta)$, as long as it yields a proper posterior density $\pi(\theta)$. The posterior probabilities of the hypotheses can then be derived (analogue to equations (2.4) and (2.5)) by (compare equations (2.13) and (2.14))

$$p(h_0|x) = \int_{\Theta_0} \pi(\theta|x)\, d\theta \tag{2.21}$$

$$p(h_1|x) = \int_{\Theta_1} \pi(\theta|x)\, d\theta\,, \tag{2.22}$$

which form the posterior odds.

2.3 Interpretation

Altogether, there are two different formulas that involve the Bayes factor:

- After plugging in equations (2.11) and (2.12) into equation (2.17) one obtains the depiction of the Bayes factor

$$BF_{01} = \frac{\int_{\Theta_0} f(x|\theta) \cdot \pi_0(\theta)\, d\theta}{\int_{\Theta_1} f(x|\theta) \cdot \pi_1(\theta)\, d\theta} \tag{2.23}$$

  as a ratio of marginal densities.

- After reordering equation (2.20) one obtains the depiction of the Bayes factor

$$BF_{01} = \frac{p(h_0|x)}{p(h_1|x)} \Big/ \frac{p(h_0)}{p(h_1)} \tag{2.24}$$

  as the ratio of the posterior odds to the prior odds.

Based on these two formulas, two different interpretations of the Bayes factor can typically be found (cp. also Morey et al., 2016; Rouder et al., 2018a):

- *Comparison of prior predictive performances.* The marginal densities $f(x|h_0)$ (equation (2.11)) and $f(x|h_1)$ (equation (2.12)) integrate the density of the data $f(x|\theta)$ over the within-hypothesis prior parameter densities $\pi_0(\theta)$ and $\pi_1(\theta)$, respectively. In that sense, the marginal density of the observed data x is a weighted average,

in which the weights are given in the hypothesis-based prior parameter distribution. As the parametric density $f(x|\theta)$ provides values that are higher for data values x that are considered to be more likely for the given parameter value θ, the marginal densities $f(x|h_0)$ and $f(x|h_1)$ provide values that are higher for data values x that are considered to be more likely on average for the given averaging weights within $\pi_0(\theta)$ and $\pi_1(\theta)$, respectively. As two hypotheses h_0 and h_1 are considered, two such values, namely $f(x|h_0)$ and $f(x|h_1)$, are provided for the actually observed data set x. The hypothesis with the higher value has considered the actually observed data x to be more likely on (weighted) average than the other hypothesis. These values $f(x|h_0)$ and $f(x|h_1)$ (the marginal densities) are typically referred to as the prior predictive likelihoods of the data x for each hypothesis h_0 and h_1, respectively. By relating these two values with each other (via ratio as in equation (2.23)), one might compare the prior predictive performance of one hypothesis h_0 with the prior predictive performance of the other hypothesis h_1: *The prior predictive performance of hypothesis h_0 is BF_{01}-times better than the prior predictive performance of hypothesis h_1.*

- *Evidence.* The prior odds $p(h_0)/p(h_1)$ state how much higher the probability $p(h_0)$ of hypothesis h_0 is compared to the probability $p(h_1)$ of hypothesis h_1 *before* the actual data x were observed. In that sense, they are typically interpreted as the degrees of belief in the respective hypotheses before the scientific investigation was conducted. Analogously, the posterior odds $p(h_0|x)/p(h_1|x)$ state how much higher the probability $p(h_0|x)$ of hypothesis h_0 is compared to the probability $p(h_1|x)$ of hypothesis h_1 *after* the actual data x were observed. In that sense, they are typically interpreted as the degrees of belief in the respective hypotheses after the scientific investigation was conducted. As the Bayes factor is a multiplicative factor that describes how the prior odds change into the posterior odds by seeing the data (equation (2.20)) it is interpreted as quantifying how the prior beliefs in the hypotheses change into the posterior beliefs in the hypotheses by seeing the data. In that, the Bayes factor relates the posterior odds to the prior odds (equation (2.24)) and extracts the influence of the latter on the former (as the prior odds are in the denominator). Therefore, it is argued that the Bayes factor itself is cleansed of the influence of how the prior probabilities of the hypotheses were chosen. This can also be seen in equation (2.17), which allows to calculate the Bayes factor based on the within-hypothesis prior parameter distributions ($\pi_0(\theta)$, $\pi_1(\theta)$) and without the prior probabilities of the hypotheses ($p(h_0)$, $p(h_1)$). Accordingly, the Bayes factor BF_{01} is interpreted as quantification of the evidence w.r.t. the hypotheses h_0 and h_1, as it states how beliefs about the hypotheses change by observing the data x, but is unaffected by the prior beliefs in the hypotheses: *The data x are evidence favoring hypothesis h_0 BF_{01}-times as much as hypothesis h_1.*

While these two interpretations seem to be in line with the mathematical foundation of Bayes factors, their use in the greater research context must be assessed. Only because a

Bayes factor value was calculated and its interpretation as comparison of prior predictive performances or as evidence is mathematically correct, it is – in general – not correct to derive any arbitrary conclusion that seems to be deducible from a statement about the comparison of prior predictive performances or from an evidence statement on an intuitive, i.e. non-mathematical, level. Science is about critical self-reflection in the context of knowledge discovery and urges to scrutinize every step taken in the scientific process. It is of little value, if scientists put a lot of effort (i.e. time and money) into a thorough investigational setup and a soundly founded statistical analysis, only to use their results in a way that does not correspond to what would be correct in a scientific and mathematical-logical way. In that, after delineating the involvement of Bayes factors in the more comprehensive decision theoretic framework, which is able to formally elaborate on how to properly use scientific results, both interpretations of the Bayes factor will be finally be evaluated w.r.t. its practical usefulness (Chapter 8).

Chapter 3

Defining Practical Relevance

In the course of the investigation about a proper use of statistical results, such as Bayes factors, it is easy to stumble upon the concept of practical significance (e.g. Kirk, 1996). Conventionally, the omnipresent framework of NHST yields results that are either statistically significant or not statistically significant. But by mistaking the mathematical term *statistically significant* as "plain-English significant" and exaggerating it to "highly significant or very highly significant, important, big!" (Cohen, 1994, p. 1001 (both quotes)) statistically significant effects or results are frequently depicted as very impactful, although some are rather practically negligible. In that sense, one needs to distinguish between the statistical significance and the practical relevance of an effect or a result. Although Kirk (1996) worked within the frequentist framework and coined the term practical significance, the main considerations about practical significance do also apply to the Bayesian framework, in which the term *significance* is typically avoided. Therefore, the term *practical relevance* is employed within this dissertation thesis.

Interestingly, over two decades have passed since Kirk (1996) urged to consider the practical significance in addition to the statistical significance, yet not formal definition of it was available. Merely stating that effect sizes are "measures" (see e.g. Ellis and Steyn, 2003) or "indices" (see e.g. Thompson, 2002; Hojat and Xu, 2004) of practical significance which indicate if results are "meaningful" (see e.g. Vaske, 2002) or "useful" (see e.g. Kirk, 1996) seemed to be sufficient. However, in order to grasp the concept of practical relevance on a mathematical level and integrate it into the existing methodologies, its formal definition is mandatory.

Searching the literature about practical significance and considering the cases in which a statistically significant, but not practically relevant result was obtained, two main characteristics emerged (Schwaferts and Augustin, 2021d)

- There was an (sometimes implicitly assumed) underlying decision that had to be

guided, creating the context in which the practical relevance of an effect or of the result could be evaluated.

- There was – by and large – an implicit agreement on how one would decide in this underlying decision problem for each possible effect magnitude.

Both these characteristics are central components of statistical decision theory (e.g. Berger, 1985; Robert, 2007), rendering statistical decision theory the framework to rely on for formalizing the notion of practical relevance.

Put the other way round, the practical relevance of a result naturally depends on what it is used for, and on a formal level such a use is a decision problem. Analogously, Berger and Wolpert (1988, p.55) reason that "no matter what is meant by inference, if it is to be of any value, then somehow it must be used, or acted upon, and this does indeed lead back to the decision-theoretic framework."

3.1 Practically Relevant Effects

Accordingly, the decision theoretic framework will be depicted in the following, in which the notion of practical relevance can be defined.

In a decision, one has to choose between several different actions. In the context of hypothesis comparisons, in which there are typically only two hypotheses being contrasted against each other, only two actions shall be considered. Denote these actions as a_0 and a_1, which are comprised in the action space $\mathcal{A} = \{a_0, a_1\}$. The objective in the decision theoretic framework is now to decide between these two actions.

Following the notation of Chapter 2, the parameter θ is of interest. Typically, the parameter formalizes an effect and is such that the absence of an effect is represented by $\theta = 0$. This shall be assumed[1] in the following. Of the two actions a_0 and a_1, denote that action which is appropriate in the absence of an effect, i.e. if $\theta = 0$, as a_0.

Deciding for one of the actions $a \in \mathcal{A}$ if a certain parameter value $\theta \in \Theta$ is true has certain consequences, and the "badness" of these consequences is quantified in a loss function

$$L : \Theta \times \mathcal{A} \to \mathbb{R}_0^+ : (\theta, a) \mapsto L(\theta, a)\,. \tag{3.1}$$

[1] If there are also nuisance parameters, one might employ a parameter density $f(x|\theta, \varphi)$, in which θ is the effect parameter of interest and φ is the (vector of) nuisance parameter(s). If the absence of an effect is indicated by a different parameter value than $\theta = 0$, one might – without loss of generality – transform the parameter θ accordingly.

Naturally, the exact specification of such a loss function depends on the applied research context and the decision problem. In that, its specification is an applied problem. Theoretical elaborations in the framework of statistical decision theory typically assume the existence of such a loss function. However, in general, its specification constitutes a serious issue[2] for applied scientists: There are a lot of aspects to consider in a decision context and many of them are less clear, characterized by missing information, and all these shall be consolidated into a single precise numerical entity. Not to mention that such a consolidation should be done for every parameter value θ, of which there are – frequently – infinitely many (as typically Θ is a continuous subset of $\mathbb{R}$), and for every action $a \in \mathcal{A}$. Accordingly, it is expected that the willingness of applied researchers to use a decision-theoretic framework will be at the mercy of the fact that a loss function must be specified. Nevertheless, the primary goal of this elaboration is to formally define, and thus better understand, the notion of practical relevance, and it appears that decision theoretic concepts, such as actions and a loss function, are required to do so. The mere fact that its feasibility for applied scientists might be low, does not mean that such decision theoretic characteristics are irrelevant in a statistical analysis that aims at producing practically relevant results. Instead, it seems that a loss consideration (i.e. an involvement of the "badness" of the consequences of potential actions) is – although implicitly – indeed somehow present when elaborating on the practical relevance of an effect. In that, in the following it is first assumed that such a loss function is available, to develop the concepts of practical relevance formally, and then tried to loosen the strict requirements for applied scientists w.r.t. its specification (Chapters 4 and 6).

The lower the loss value of an action $a \in \mathcal{A}$ for a given parameter value $\theta \in \Theta$, the better is the action. Accordingly, if each parameter value $\theta \in \Theta$ is considered separately, one is required to decide for the action with lower loss:

- a_0 is preferred over a_1, if $L(a_0, \theta) < L(a_1, \theta)$.
- a_1 is preferred over a_0, if $L(a_1, \theta) < L(a_0, \theta)$.
- Both actions are considered equivalent (w.r.t. the quantification of the "badness" of their consequences) and there is no preference, if $L(a_1, \theta) = L(a_0, \theta)$.

As a_0 denotes the action that is appropriate if the effect is absent, i.e. if $\theta = 0$, its loss

[2]Although there are elaborations on how to elicit loss functions, utilities, or preferences (cp. e.g. Abdellaoui, 2000; Chajewska et al., 2000), which might then be used in a decision theoretic context, many of these elaborations are located within the field of economics, in which the utility might be quantified more easily by applying monetary considerations. Yet, in many other fields it might not be that easy. In this context, it shall be noted that even money itself might not have a utility which is linear in its amount, as there might be a variety of different conceptions of utility (cp. e.g. Diener and Oishi, 2000). Further, there is also a high degree of variability in preference elicitation, even if the same subject was asked repeatedly (cp. e.g. Froberg and Kane, 1989), emphasizing the need to deal with vague, partial, and ambiguous loss information.

$L(a_0, 0)$ is smaller than $L(a_1, 0)$. Similarly, other effects θ which also prefer a_0 over a_1 lead to the same decision as an absence of an effect, so they cannot be practically relevant effects. Accordingly, practically relevant effects are those effects that prefer a_1 over a_0 (reprinted from Schwaferts and Augustin, 2021d):

Definition 1 (Practical Relevance of an Effect) *Within this framework, an effect θ is practically relevant (or practically significant) w.r.t. the actions a_0, a_1, and the corresponding loss function L, if a_1 is preferred over a_0, i.e. if*

$$L(\theta, a_1) < L(\theta, a_0)\,, \tag{3.2}$$

else the effect θ is negligible (or practically zero) w.r.t. these actions and this loss function.

This definition emphasizes that the notion of the practical relevance of an effect is embedded into a specific decision theoretic context. The practical relevance (or negligibility) of an effect depends on the specific decision problem and the specific loss function. With different actions or a different loss function, different parameter values would be practically relevant (or negligible).

Further, it appears that the definition of a practically relevant (or practically negligible) effect does not involve an observed data set and this does indeed match with the intuition: A researcher is able to state which action $a \in \mathcal{A}$ is to be preferred if certain parameter values θ are true before the data were observed. There is merely low certainty about which parameter is true without the observed data.

3.2 Practical Relevance w.r.t. Hypotheses

In the context of the hypothesis specification as in equation (2.3), NHST specifies the null hypothesis such that it contains only the zero effect, i.e. $\Theta_0 = \{0\}$, and the alternative hypothesis such that it contains all other effects, i.e. $\Theta_1 = \{\theta \in \Theta | \theta \neq 0\}$. The critique against NHST, that it might yield statistically significant but not practically signification results (e.g. Cohen, 1994), relates on a formal level to the fact that h_1 hypothesizes not only practically relevant but also practically negligible effects (for a – potentially implicitly assumed – given decision problem and loss function). To tackle this critique, Θ_0 should not contain practically relevant effects and Θ_1 should not contain practically negligible effects. If so, one might say that the hypotheses h_0 and h_1 incorporate the notion of practical relevance (w.r.t. the given decision problem and loss function) (reprinted from Schwaferts and Augustin, 2021d) (illustrated in Figure 4.2).

Definition 2 (Practical Relevance w.r.t. Hypotheses) *Within this framework, two hypotheses about an effect θ (equation 2.3)* ***completely*** *incorporate the notion of practical*

*relevance (or practical significance) w.r.t. two associated actions a_0, a_1 and the corresponding loss function L, if Θ_1 contains **all** practically relevant effects and Θ_0 contains **all** negligible effects, i.e.*

$$\forall \theta \in \Theta : L(\theta, a_0) \leq L(\theta, a_1) \Rightarrow \theta \in \Theta_0 \tag{3.3}$$

$$\forall \theta \in \Theta : L(\theta, a_1) < L(\theta, a_0) \Rightarrow \theta \in \Theta_1 \,. \tag{3.4}$$

*These hypotheses (equation 2.3) **partially** incorporate the notion of practical relevance (or practical significance) w.r.t. these actions and this loss function, if Θ_1 contains **only** practically relevant effects and Θ_0 contains **only** negligible effects, i.e.*

$$\forall \theta \in \Theta_0 : L(\theta, a_0) \leq L(\theta, a_1) \tag{3.5}$$

$$\forall \theta \in \Theta_1 : L(\theta, a_1) < L(\theta, a_0) \,. \tag{3.6}$$

This definition, as elaborated on in the contribution (Schwaferts and Augustin, 2021d), distinguishes between hypotheses that completely incorporate the notion of practical relevance and hypotheses that only partially incorporate the notion of practical relevance. However, hypotheses that incorporate the notion of practical relevance only partially (and not completely) do not constitute a partition of the parameter space, especially the condition $\Theta_0 \cup \Theta_1 = \Theta$ does not hold, a case that is rather irrelevant for further elaborations about the practical relevance of Bayes factors. Therefore, in the remainder of this dissertation – unless otherwise stated – hypotheses are assumed to incorporate the notion of practical relevance completely whenever they are said to incorporate the notion of practical relevance.

Although connected with each other, it is important to consider the practical relevance (or negligibility) of an effect and the notion of practical relevance w.r.t. hypotheses as two disjoint concepts. In the context of the practical relevance of an effect, the parameter space Θ was basically separated into two subset, each one only containing parameter values that favor the same action. Parameter values within one of these sets are then labeled as practically relevant, parameter values within the other set as practically negligible. The assignment of these labels might seem somehow arbitrary: If these two labels were swapped, the definitions of the practical relevance and of the practical negligibility of an effect would still have the same mathematical structures. In that sense, these labels were simply selected in a manner to match with the intuition that a zero-effect is practically negligible and to match with the employment of term practical significance for effects essentially different from zero. Of importance here is that the parameter space was separated into two disjoint subset due to the nature of the decision problem and the loss function. The notion of practical relevance w.r.t. hypotheses regards whether the hypotheses are specified in line with these two disjoint subsets or not. If so, one might call the specification of the hypotheses as a practically relevant specification, if not, one might say that the specification of the hypotheses lacks practical relevance. In that sense, the notion of practical relevance w.r.t. hypotheses refers to the match of a specification with the underlying decision problem.

In such a way, one might assess the complete statistical analysis: If the practical relevance of a result is of interest, there is – at least implicitly assumed – an underlying decision problem. Necessary quantities for this analysis are specified in a practically relevant manner (w.r.t. this decision problem), if their specification matches with the underlying decision problem. If all quantities that are required by the statistical analysis are specified in a practically relevant manner and the statistical analysis is performed correctly, then the results are of practical value, as they are able to inform the underlying decision problem.

Accordingly, in order to evaluate the practical relevance of Bayes factors, it is mandatory to depict Bayes factors in the context of statistical decision theory.

Chapter 4

Hypothesis-Based Bayesian Decision Theory

4.1 Formal Framework

Analogue to the structure of Chapter 2, the general framework of Bayesian decision theory shall be outlined first, before its adaptation towards Bayes factors is addressed.

Following the notation of Chapter 2, the objective is to decide between the actions within $\mathcal{A} = \{a_0, a_1\}$ on the basis of a loss function L as in equation (3.1). By observing the data x, the prior density $\pi(\theta)$ was updated via Bayes rule (equation (2.1)) to the posterior density $\pi(\theta|x)$. It is now possible to calculate the expected posterior loss

$$\rho : \mathcal{A} \to \mathbb{R}_0^+ : a \mapsto \rho(a) = \int_\Theta L(\theta, a) \cdot \pi(\theta|x)\, d\theta \tag{4.1}$$

for each action $a \in \mathcal{A}$. The conditional Bayes decision principle (cp. e.g. Berger, 1985, p. 16) in the context of Bayesian decision theory states that an action is considered as optimal action

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, \rho(a) \tag{4.2}$$

if it has minimal expected posterior loss.

As delineated in the previous chapter, this loss function L might also be used to separate the parameter space Θ into two disjoint (partitioning) parameter sets Θ_0 and Θ_1, forming the hypotheses as in equation (2.3).

Hypothesis-based decision theory, however, is based on a loss function

$$L_H : \mathcal{H} \times \mathcal{A} \to \mathbb{R}_0^+ : (h, a) \mapsto L_H(h, a) \tag{4.3}$$

that is defined on the hypothesis space $\mathcal{H} = \{h_0, h_1\}$ instead of the parameter space Θ (as in equation (3.1)). In this simple setting with only two hypotheses and two actions, this hypothesis-based loss function L_H consists of only four values. In addition, deciding for a_0 if h_0 is true and for a_1 if h_1 is true is typically considered to be a correct decision. Without loss of generality, respective loss values might then be set to zero, i.e. $L_H(h_0, a_0) = L_H(h_1, a_1) = 0$, rendering the loss function L_H to be in regret form. The other two loss values quantify the "badness" of the consequences of the type-I-error ($k_1 := L(h_0, a_1)$) and the type-II-error ($k_0 := L(h_1, a_0)$), if deciding correctly has loss zero. These two values are assumed to be non-zero. In order to determine the optimal action with this hypothesis-based loss function L_H, only the ratio of these loss values is necessary. In that, define

$$k := \frac{k_1}{k_0} = \frac{L(h_0, a_1)}{L(h_1, a_0)}\,, \tag{4.4}$$

quantifying how bad the type-I-error is compared to the type-II-error.

As elaborated within Chapter 2, assessing the hypotheses in the light of the data x allows to calculate the posterior probabilities $p(h_0|x)$ and $p(h_1|x)$ of the hypotheses (equations (2.13) and (2.14)).

The expected posterior loss $\rho(a)$ (equation (4.1)) can now be calculated without a complicated integration via

$$\rho(a) = L(h_1, a) \cdot p(h_1|x) + L(h_0, a) \cdot p(h_0|x)\,. \tag{4.5}$$

In order to determine the optimal action, the minimal expected posterior loss has to be found (equation (4.2)). As there are only two actions, the minimum can be found by considering the ratio of expected posterior losses

$$r := \frac{\rho(a_1)}{\rho(a_0)} = \frac{L(h_0, a_1) \cdot p(h_0|x)}{L(h_1, a_0) \cdot p(h_1|x)} \tag{4.6}$$

$$= k \cdot \frac{p(h_0|x)}{p(h_1|x)} \tag{4.7}$$

and checking whether it is above or below 1. In that, the set $\mathcal{A}^*$ of optimal actions is[1]

$$\mathcal{A}^* = \begin{cases} \{a_0\} & \text{if } r > 1 \\ \{a_1\} & \text{if } r < 1 \\ \{a_0, a_1\} & \text{if } r = 1 \end{cases}\,. \tag{4.8}$$

If $r = 1$, then both actions are considered to be optimal, as they are expected to yield exactly the same "badness" of their consequences, and might be treated as practically equivalent, such that any of these action actions might selected arbitrarily.

[1]This equation (4.8) was taken from (Schwaferts and Augustin, 2021c), in which the conditions are erroneously reversed. The version of the equation depicted here is correct and the version reported in (Schwaferts and Augustin, 2021c) is incorrect. Nevertheless, the final formula for the optimal actions in the imprecise case within (Schwaferts and Augustin, 2021c) is correct.

4.2 Simplification Assumption

Formally, the parameter-based loss function L (equation (3.1)) and the hypothesis-based loss function L_H (equation (4.3)) are two different mathematical objects, yet it is possible to depict the hypothesis-based loss function L_H in terms of a parameter based loss function L, which is constant on the parameter values within the parameter sets Θ_0 and Θ_q that define the hypotheses (illustrated in Figure 4.1).

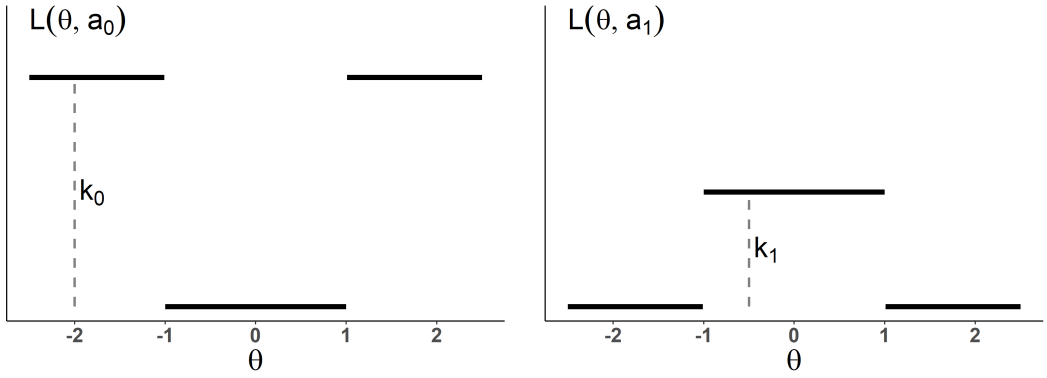

Figure 4.1: Simplified Loss Function. Assume a hypothesis-based loss function L_H in regret form is given by a specification of the values k_0 and k_1. Depicting this loss function in dependence of the parameter θ, the corresponding parameter-based loss function L (equation (3.1)) is constant within the sets Θ_0 and Θ_1, respectively. Within this example, these are $\Theta_0 = [-1, 1]$ and $\Theta_1 = (-\infty, -1) \cup (1, \infty)$. The plots depict the loss value $L(\theta, a)$ (y-axis) in dependence of the parameter θ (x-axis) and the actions a_0 (left plot) or a_1 (right plot). This figure was taken from (Schwaferts and Augustin, 2021b).

Naturally, this hypothesis-based loss function L_H contains way less information than the parameter-based loss function L (as illustrated in Figure 4.2). By using L to obtain parameter sets Θ_0 and Θ_1, which are then used in the context of L_H, information is lost. And the question arises of why a researcher should do so?

Obviously, if a parameter-based loss function L is available, then the optimal action a^* can be found easily (equation (4.2)), and there is no need to use such a hypothesis-based simplification. However, such a specific loss function L is rarely available in an unambiguous specification. So there is the need – for researchers being interested in the practical relevance of their obtained results – to deal with decision problems in which information about the loss function is scarce. So while such scarce information about the consequences of potential actions might be insufficient for the researcher to specify the full parameter-based loss function L unambiguously, it might be adequate to inform about the parameter sets Θ_0, Θ_1 and the value k, allowing a less ambiguous specification of the hypothesis-based loss function L_H.

Now, in order to guarantee that results obtained with such a simplified hypothesis-based loss function L_H do actually inform the decision problem of interest – which is assumed to be characterized by the unknown parameter-based loss function L – it is mandatory to understand the connection between L and L_H. And this connection is via hypotheses that incorporate the notion of practical relevance w.r.t. the underlying decision problem. If these hypotheses are used within the hypothesis-based loss function L_H, latter becomes as close as possible to the underlying decision problem. Then the researcher has to specify the value k in a way that best captures the available (potentially partial) information about the consequences of the type-I-error and the type-II-error.

Accordingly, by using the hypothesis-based loss function L_H instead of the full parameter-based loss function L an assumption was made, namely that it is possible to simplify the rather complicated loss L into the simpler form of L_H. This assumption shall be referred to as *simplification assumption* (illustrated in Figure 4.2). Naturally, by the nature of an assumption, this simplification assumption might be (and most likely is) incorrect, introducing a potential error that might lead to incorrect results, i.e. false decision. In that, there might be the case that – for a given data set x – the optimal action a^* might be different whether the full (but unknown) loss function L or the simplified loss function L_H were employed. However, it might be expected that the better L_H matches with L the less severe is the error due to the simplification assumption, emphasizing the importance of using hypotheses that incorporate the notion of practical relevance.

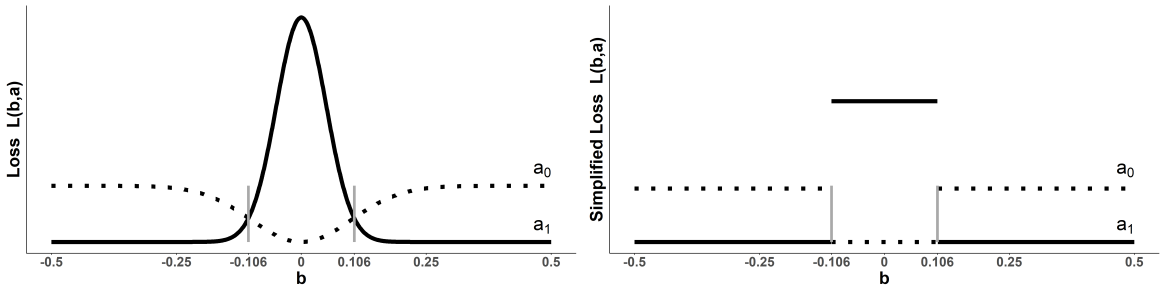

Figure 4.2: Example: Simplification Assumption. Assume the parameter of interest is the bias b of a presumably fair coin in a gamble, and the actions refer to whether (a_1) or not (a_0) to accuse the person who offers the gamble of cheating. Further, assume the parameter-based loss function within the left plot is given. According to the definitions of practical relevance (Chapter 3), hypotheses that incorporate the notion of practical relevance are then given by $\Theta_0 = [-0.106, 0.106]$ and $\Theta_1 = [-0.5, -0.106) \cup (0.106, 0.5]$. A simplified loss function in accordance with these hypotheses is constant within the respective parameter sets (right plot). These plots illustrate that the simplification of the loss function by using hypotheses is an assumption, namely that it is possible to depict the actual loss (left plot) by a loss function that is constant within the hypotheses. Finding these (constant) loss values in the context of the hypothesis-based loss function might be difficult and somewhat erroneous. Within this example, the loss values for this simplification (right plot) were chosen arbitrarily. This example and the left part of the figure were taken from (Schwaferts and Augustin, 2021d).

By no means is this potential error due to the simplification assumption a reason to refrain from a hypothesis-based Bayesian decision theoretic account! The aim here is to mathematically understand the process involved in the context of using results of hypothesis-based statistical analyses in a practical context. Within this process, it appears that this simplification assumption was made, and now – being stated explicit – its implications and potential errors become aware. As there is an underlying decision problem involved (or implicitly assumed) when dealing with the practical relevance of research results, a hypothesis-based statistical analysis will (usually silently) make this simplification assumption. Accordingly, instead of considering the simplification assumption as a malign part of the framework of hypothesis-based Bayesian decision theory, it should rather be considered as an integral part of every hypothesis-based statistical analysis whose results are used in a practical context.

4.3 Decision Theory: Bayes Factors

With the hypothesis-based Bayesian decision theoretic framework outlined (Section 4.1), it is possible to delineate its relation with Bayes factors.

As elaborated (equation (2.20)), Bayes factors are used to update the prior odds $p(h_0)/p(h_1)$ to the posterior odds $p(h_0|x)/p(h_1|x)$, which are used to calculate the ratio r of expected posterior losses (equation (4.7)). Accordingly, it might be formulated as

$$r = k \cdot BF_{01} \cdot \frac{p(h_0)}{p(h_1)}\,, \tag{4.9}$$

emphasizing its dependence on the Bayes factor BF_{01}. The set of optimal actions $\mathcal{A}^*$ can then be found analogously via equation (4.8).

Accordingly, it appears that in order to guide a decision in the context of Bayes factors, three quantities are necessary: The prior odds, the Bayes factor itself, and the value k representing the hypothesis-based loss function. How do each of those these three quantities relate to the underlying decision problem?

To investigate this question in greater detail, consider the case of Bayesian decision theory without hypotheses (as depicted in the beginning of Section 4.1) first: The prior density $\pi(\theta)$ was updated to the posterior density $\pi(\theta|x)$, which is then used to find the optimal action based on the loss function L (Figure 4.3, top scheme). The prior and the posterior distribution are central parts of Bayesian statistics and do not necessarily require a decision theoretic account. In that, they can be classified as non-decision-theoretic quantities. The loss function, in turn, is inevitably tied to the underlying decision problem and therefore a decision-theoretic quantity (Figure 4.3, light-and-dark gray coloring).

After introducing hypotheses into the Bayesian decision theoretic account and simplifying the loss function L towards the hypothesis-based loss function L_H, the situation is as follows: Again, the prior density $\pi(\theta)$ was updated to the posterior density $\pi(\theta|x)$, however, then assessed w.r.t. the hypotheses to obtain the posterior odds, which, in turn, are used to derive the optimal action based on the simplified loss L_H. As hypotheses need to be related to the underlying loss function L to incorporate the notion of practical relevance (Chapter 3), they are tied to the underlying decision problem and, therefore, actually a decision-theoretic quantity. In that, as the posterior odds depend on the hypotheses, they are also a decision-theoretic quantity (Figure 4.3, middle scheme).

In the context of calculating a Bayes factor, the prior density $\pi(\theta)$ is decomposed into the prior odds and the within-hypothesis prior parameter distributions, all depending on a hypothesis specification. The Bayes factor is then based only on the hypotheses and the within-hypothesis prior parameter distributions, leaving out the prior odds (Figure 4.3, bottom scheme). As hypotheses, which incorporate the notion of practical relevance, are considered to be a decision-theoretic quantity, so are the prior odds and the within-hypothesis prior parameter distributions: Without an underlying decision problem, it is not possible to specify the hypotheses such that they incorporate the notion of practical relevance, and therefore it is not possible to derive the prior odds and the within-hypothesis prior parameter distributions such that they are in accordance with the practical purpose of the study.

In summary, if the practical relevance of Bayes factors is of interest, Bayes factors are considered to be a decision-theoretic quantity, which cannot be reasonably (i.e. being in line with the practical purpose of the investigation) calculated without considering the underlying decision problem. Respective hypotheses are derived by considering the underlying decision problem and separating the parameter space according to the notion of practical relevance (Chapter 3).

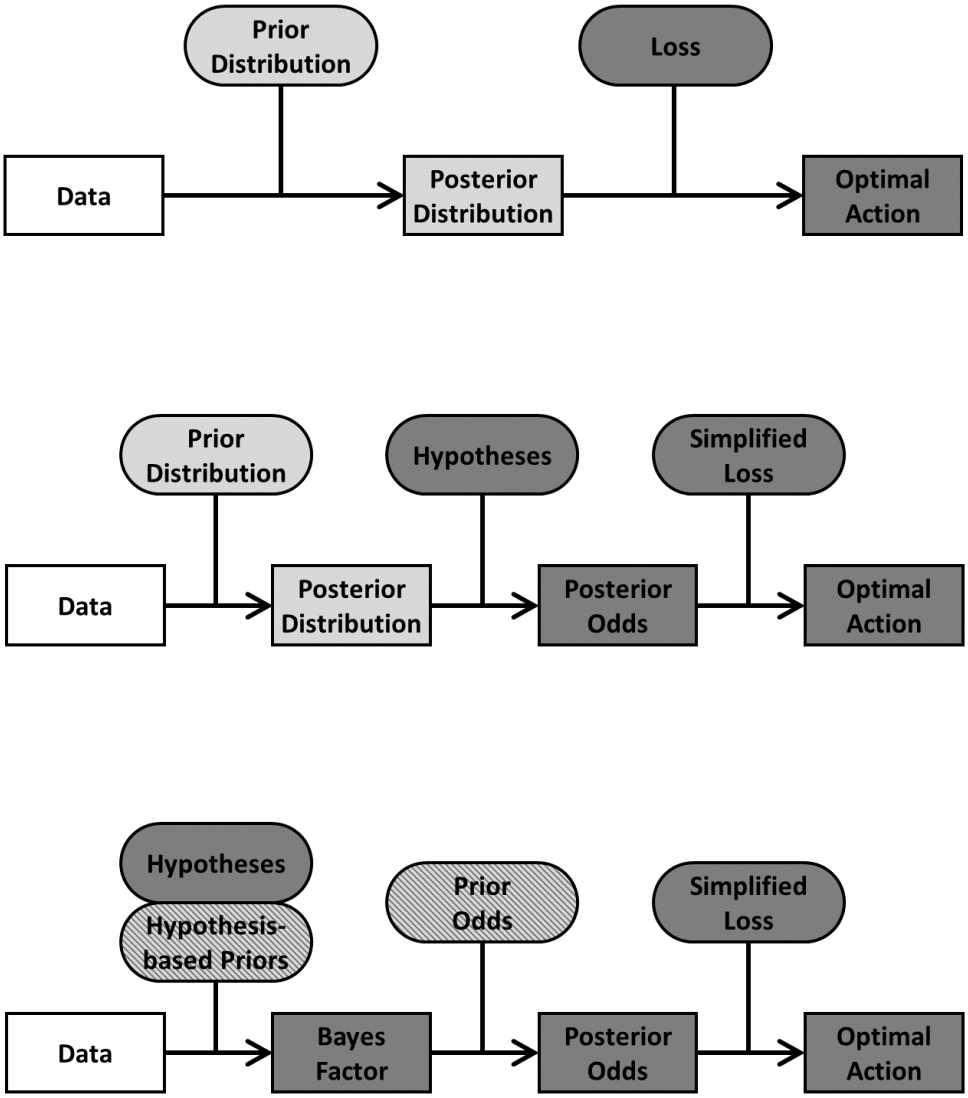

Figure 4.3: Guiding Decisions with Bayes Factors. These schemes illustrate the steps involved in deriving optimal actions from observed data. Round boxes depict additional specifications that are required besides the data and squared boxes depict the quantities that can be calculated from the data with the respective additional quantities. A light gray shading of these boxes states that these quantities can be specified and derived without considering a certain (underlying) decision problem, and a dark gray shading states that these quantities can only be reasonable specified or calculated when considering the underlying decision problem. Depicted are schemes for Bayesian decision theory (top), hypothesis-based Bayesian decision theory (middle) and Bayes factor based Bayesian decision theory (bottom). The dark-and-light gray shading of the hypothesis-based priors and the prior odds indicates that it is actually a specification that is independent of the underlying decision problem if taken together as overall prior, but separated in the context of the underlying decision problem by considering the hypotheses. In that sense, both these quantities cannot be reasonably specified without considering the underlying decision problem.

Chapter 5

Statistical Hypotheses

Derived within the previous chapters and in line with the current employment of hypotheses in frequentist statistics, statistical hypotheses are considered to be subsets of the parameter space.

In the literature of Bayes factors, however, hypotheses are depicted as mathematical objects that consist of *both* a subset of the parameter space *and* the respective within-hypothesis prior parameter distribution, which together represent a to-be-evaluated theory of interest (cp. e.g. Vanpaemel, 2010; Vanpaemel and Lee, 2012; Morey et al., 2016; Rouder et al., 2018a; Dienes, 2019; Tendeiro and Kiers, 2019). In this regard, it is argued, that the (within-hypothesis) prior distributions have "empirical content" (e.g. in Vanpaemel and Lee, 2012, p. 1052).

This conception is in disagreement with the notion of statistical hypotheses as being only subsets of the parameter space, and might complicate the employment of a hypothesis-based Bayesian decision theoretic account. In that, this disagreement should be addressed further and it appears that the – seemingly unrelated – concept of updating consistency of Bayes factors allows to find a solution.

5.1 Updating Consistency

Assume that, besides the data set x, there is an additional data set y which arose from the same investigational setup, such that it is distributed according to the same parametric density.

As the prior density $\pi(\theta)$ is assumed to formalize all available knowledge about the parameter before any data were observed, the posterior density $\pi(\theta|x)$ formalizes all available

knowledge about the parameter after the data x were observed. Accordingly, this posterior density $\pi(\theta|x)$ might be employed as a new prior density for the assessment of the new data set y, leading to the final posterior density $\pi(\theta|y, x)$. Naturally, it should not matter if the prior distribution was subsequently updated twice with both data sets x and y in two steps or if both data sets were merged first $z = (x, y)$ and then used to update the initial prior distribution at once (Figure 5.1):

$$\pi(\theta|y, x) = \pi(\theta|z) \tag{5.1}$$

If equation (5.1) holds Bayesian updating is called consistent, else Bayesian updating is called inconsistent (cp. Rüger, 1998, p. 190).

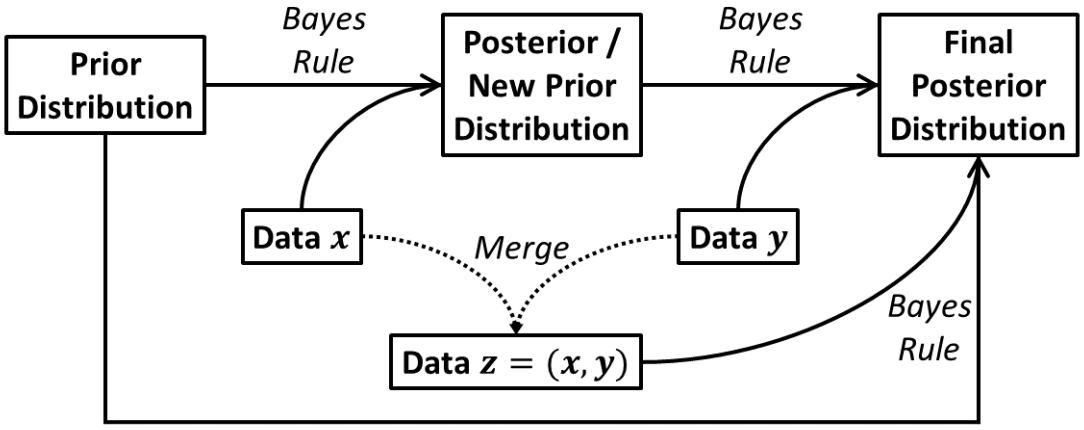

Figure 5.1: Consistent Bayesian Updating. Bayesian updating is consistent if it does not matter whether two separate data sets (following the same parametric sampling distribution) are considered separately (top path) or merged first and the considered at once (bottom path). This figure was taken from (Schwaferts and Augustin, 2021e).

When considering the decomposed prior density (equation (2.8)) as in the context of Bayes factors, subsequently updating works as follows (Figure 5.2):

Denote the Bayes factor[1] (equation (2.23)) that is based on the first data set x as BF^x and the Bayes factor (again equation (2.23)) that is based on the second data set y as BF^y. Please note, that now BF^y was calculated without considering that the previous data set x was already observed.

[1]Within this chapter the index 01 of the Bayes factor BF will be omitted to keep the notation simple and easily readable. Still, hypothesis h_0 (numerator) is compared to hypothesis h_1 (denominator).

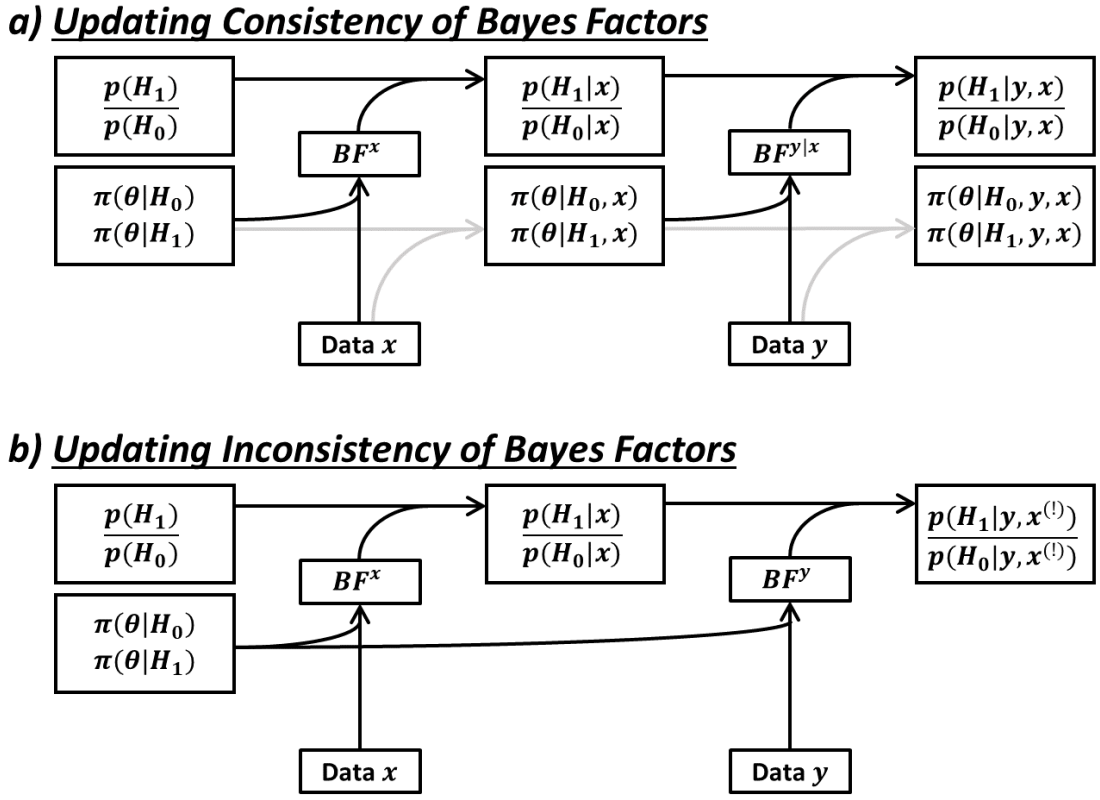

Figure 5.2: Consistent Updating with Bayes Factors. Scheme a) depicts consistent updating and scheme b) depicts inconsistent updating with Bayes factors. Updating the prior distribution, which was decomposed w.r.t. to the hypotheses (equation (2.8)), has to consider the update of both the prior odds and the within-hypothesis prior parameter distributions, although the Bayes factors is only required within the update of the prior odds (a, left). Nevertheless, the update of the within-hypothesis prior distributions has to be considered to conduct a consistent calculation of the Bayes factor of a second data set y (a, right). Ignoring the update of the within-hypotheses prior parameter distributions leads to a Bayes factor value of the second data set y that results in inconsistent Bayesian updating (b). This figure was taken from (Schwaferts and Augustin, 2021a).

After observing the first data set x, the prior odds get updated by the Bayes factor BF^x to the posterior odds (this is the same formula as equation (2.20), only adapting the notation with BF^x)

$$\frac{p(h_0|x)}{p(h_1|x)} = BF^x \cdot \frac{p(h_0)}{p(h_1)}\,. \tag{5.2}$$

In addition, also the within-hypothesis prior parameter densities $\pi_0(\theta)$ and $\pi_1(\theta)$ get updated via Bayes rule to their posterior densities $\pi_0(\theta|x)$ (equation (2.9)) and $\pi_1(\theta|x)$ (equation (2.10)) as depicted in Chapter 2 (Figure 5.2, top). However, by using these within-hypothesis posterior densities $\pi_0(\theta|x)$ and $\pi_1(\theta|x)$ as new within-hypothesis prior densities

for the calculation of the Bayes factor based on the second data set y, the resulting Bayes factor

$$BF^{y|x} = \frac{f(y|h_0, x)}{f(y|h_1, x)} = \frac{\int_{\Theta_0} f(y|\theta) \cdot \pi_0(\theta|x)\, d\theta}{\int_{\Theta_01} f(y|\theta) \cdot \pi_1(\theta|x)\, d\theta} \tag{5.3}$$

is, in general, different to the Bayes factor BF^y obtained with the initial within-hypothesis prior densities $\pi_0(\theta)$ and $\pi_1(\theta)$:

$$BF^{y|x} \overset{in\ general}{\neq} BF^y\,. \tag{5.4}$$

It is this Bayes factor $BF^{y|x}$ that allows to update the previous posterior odds to the final posterior odds

$$\frac{p(h_0|y, x)}{p(h_1|y, x)} = BF^{y|x} \cdot \frac{p(h_0|x)}{p(h_1|x)} \tag{5.5}$$

consistently. By using BF^y instead of $BF^{y|x}$ in equation (5.5), updating with Bayes factors is inconsistent (Figure 5.2, bottom). Again, the previous within-hypothesis posterior densities $\pi_0(\theta|x)$ and $\pi_1(\theta|x)$ get updated via Bayes rule to the final within-hypothesis posterior densities

$$\pi_0(\theta|y, x) = \frac{f(y|\theta) \cdot \pi_0(\theta|x)}{f(y|h_0, x)} \tag{5.6}$$

$$\pi_1(\theta|y, x) = \frac{f(y|\theta) \cdot \pi_1(\theta|x)}{f(y|h_1, x)}\,, \tag{5.7}$$

where

$$f(y|h_0, x) = \int_{\Theta_0} f(y|\theta) \cdot \pi_0(\theta|x)\, d\theta \tag{5.8}$$

$$f(y|h_1, x) = \int_{\Theta_1} f(y|\theta) \cdot \pi_1(\theta|x)\, d\theta \tag{5.9}$$

are respective marginal densities.

In summary, it appears that the update of the within-hypothesis prior densities $\pi_0(\theta)$ and $\pi_1(\theta)$ to the within-hypothesis posterior densities $\pi_0(\theta|x)$ and $\pi_1(\theta|x)$ must not be neglected in an analysis with Bayes factors, even if only one data set x is present and a Bayes factor value can be calculated without considering this update. Mathematically, this update occurs, a fact often neglected (as e.g. recognized by Rouder and Morey, 2011) or depicted wrongly (e.g. Tendeiro and Kiers, 2019, p. 776, footnote 2 therein) in the literature about Bayes factors.

5.2 Big Picture

Being aware of this update of the within-hypothesis prior distributions, it is now possible to assess the nature of statistical hypotheses. To do so, the big picture of statistical inference

(cp. Kass, 2011) shall be regarded in the context of Bayes factors.

This big picture of statistical inference (Figure 5.3) distinguishes between a real world and a theoretical world, in which the real world objects are formalized as mathematical objects. Interpreting mathematical objects from the theoretical world leads to their counterpart in the real world. In the context of a statistical analysis, the phenomenon of interest in the real world is represented as the parameter of the corresponding parametric sampling distribution in the theoretical world. The quality of this representation depends on the quality of the investigational setup.

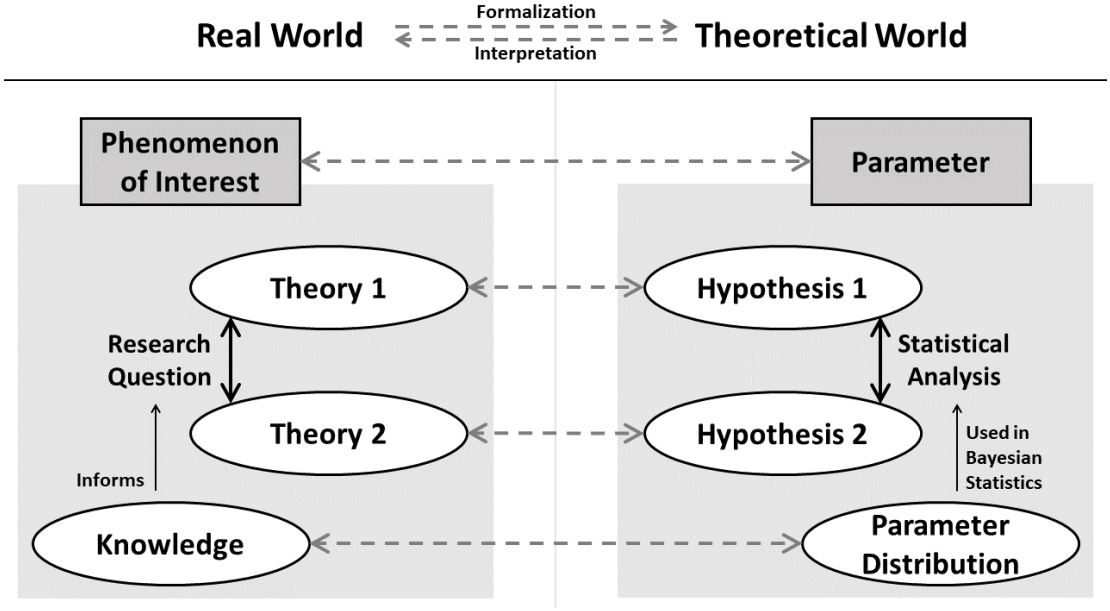

Figure 5.3: Bayes Factors: Big Picture. The big picture of statistical inference (Kass, 2011) differentiates between a real world (left side) and a theoretical world (right side). This scheme depicts how the essential theoretical quantities in the context of Bayes factors are related to their real world interpretations. This figure was taken from (Schwaferts and Augustin, 2021a).

In the context of Bayes factors, on the theoretical side, there is a Bayesian parameter distribution and – due to the comparative nature of Bayes factors (or hypothesis-based analyses in general) – two mathematical objects that should be contrasted against each other. These objects are referred to as (statistical) hypotheses (but are sometimes also denoted as models (cp. esp. Tendeiro and Kiers, 2019, p. 775, footnote 1 therein)). The interpretation of the Bayesian (prior or posterior) parameter distribution is knowledge[2] about the phenomenon of interest (before or after the data were observed). The hypotheses in the theoretical world formalize two theories about the phenomenon of interest in the real world that should be contrasted against each other by the research question. Due to their contrasting nature, the theories will be referred to as theoretical positions. Essentially, they

[2]Cp. footnote 1 on page 11 of this thesis.

are two yet uninvestigated conjectures about the phenomenon of interest. In a Bayesian analysis, the prior distribution will get updated to the posterior distribution that is said to contain all relevant knowledge about the parameter (i.e. phenomenon of interest in the real world) after the data are observed, so it is this posterior knowledge that is able to answer the research question in the real world.

In such a contrasting research question, the theoretical positions should be compared against each other. The desired outcome is a statement about the relative plausibility of the theoretical positions. There are two theoretical positions about the phenomenon of interest, which is more plausible? Which one is favored by the data? It is not: There are two theoretical positions about the phenomenon of interest, how do both theoretical positions change by seeing the data? If so, the results would be two new theoretical positions, not an assessment of the relative plausibility of the initially stated theoretical positions. However, it were the initially stated theoretical positions which were of scientific interest and contrasted in the research question. So, *the initially stated theoretical positions must not change by seeing the data, only their plausibility should.* This requirement is in line with the interpretation of Bayes factors as a quantification of evidence: Evidence does not change a theory, only its plausibility. Theories with low evidence might be dismissed for the sake of a different theory, but this is a different step that comes after evidence quantification. How to quantify evidence and how to use evidence are two separate aspects. Bayes factors *per se* only concern the quantification of evidence.

Now, assume that – as frequently stated in the literature about Bayes factors – hypotheses (i.e. the mathematical objects that are contrasted by Bayes factors) are represented by *both* sets of parameters *and* within-hypothesis prior parameter distributions, such that the theoretical positions are formalized by both these mathematical objects. It was shown in the Section 5.1 that the within-hypothesis prior parameter distributions do get updated by seeing the data. If these within-hypothesis prior parameter distributions formalize the theoretical positions of interest, after seeing the data the within-hypothesis posterior distribution formalize different theoretical positions, which are, in general, not of interest for the research question. It appears that a proper and useful interpretation of Bayes factors as evidence quantification is no longer possible, if the within-hypothesis prior distributions are also used to formalize theoretical positions. In that, they cannot be constituents of those mathematical objects (hypotheses) that are contrasted with each other by the Bayes factor.

Further, it appears that both updating consistency of Bayes factors and a proper interpretation of Bayes factors as evidence quantification can be maintained if hypotheses are only defined by sets of parameters and not by prior distributions (Schwaferts and Augustin, 2021a). In that, this nature of statistical hypotheses is in line with the derivation of statistical hypotheses as subsets of the parameter space in the context of the simplification of a loss function (Chapter 4).

Naturally, the question arises of how to specify the prior distribution, if not by the content of the theoretical positions of interest. Sorting back to the interpretation of Bayesian parameter distributions as knowledge about the phenomenon of interest, the answer to this question is: The prior distribution specifies what *is known* about the phenomenon of interest before conducting the investigation, the hypotheses specify what *should be assessed* about the phenomenon of interest in the context of the research question. *Former is actual knowledge, latter are some hypothetical conjectures.*

Elaborating this difference between the real-world counterparts of both hypotheses and prior distributions, it appears that there is a clear guidance on the structure of the specification of the prior parameter distribution. In Section 2.1 two different ways to specify the prior parameter distribution were depicted: Either as an overall prior parameter density $\pi(\theta)$ or as the within-hypothesis prior parameter densities $\pi_0(\theta)$ and $\pi_1(\theta)$ together with the prior odds $p(h_0)/p(h_1)$. Former specification is independent of the hypotheses and latter specification is dependent on the hypotheses. However, by combining the within-hypothesis prior parameter densities $\pi_0(\theta)$ and $\pi_1(\theta)$ with the prior odds $p(h_0)/p(h_1)$ to the overall prior parameter density $\pi(\theta)$ (equation (2.8)), there is no dependence on the hypotheses anymore. In that, using latter specification of the prior situation with a correct relation to the real world seems to be nearly impossible: The within-hypothesis prior parameter densities $\pi_0(\theta)$, $\pi_1(\theta)$ and the prior odds $p(h_0)/p(h_1)$ should be specified such that they contain the actual knowledge about the phenomenon of interest in dependence on some hypothetical conjectures, but by merging them to the overall prior parameter density $\pi(\theta)$ its dependence on these hypothetical conjectures has to be gone, containing only the actual knowledge about the phenomenon of interest as if there were no hypothetical conjectures. This seems a remarkable procedure. In that sense, it suggests itself to simply specify the overall prior density $\pi(\theta)$, representing the actual knowledge about the phenomenon of interest, and the hypotheses, representing the hypothetical conjectures of interest, in two independent steps.

Chapter 6

Generalizations with Imprecise Specifications

After consolidating that the disagreement about the nature of statistical hypotheses between the previous decision theoretic considerations (Chapter 4) and the current recommendations in the literature about Bayes factors (primarily elaborated in Vanpaemel, 2010; Vanpaemel and Lee, 2012) does indeed arise from an erroneous conception within the literature, the focus of this thesis can be brought back on how to improve the practical relevance of Bayes factors.

The general insight from the development of the definition of the notion of practical relevance (Chapter 3) was that all essential quantities that are used within a statistical analysis need to be specified such that they match best with the underlying (potentially implicitly assumed) decision context. Put the other way round: By misspecifying essential quantities results are obtained that do inform past the practical purpose of the investigation. In that sense, these results lack practical relevance.

It was elaborated that it is important to specify the hypotheses such that they incorporate the notion of practical relevance (Section 3.2) and to specify the simplified loss function L_H via the value k (equation (4.4)) such that it captures the available information about the consequences of the type-I-error and the type-II-error. Besides, also the prior parameter distribution needs to be specified to perform a hypothesis-based Bayesian decision theoretic analysis. Also this quantity needs to be specified such that it properly captures the relevant knowledge about the parameter.

Typically, information about the prior distribution, the hypotheses, and the value k in the context of the simplified loss function L_H might be scarce, vague, and partial. It might rather be seen as an exception if an applied scientist is able to unambiguously specify these quantities as precise values without pretending a level of certainty that is actually

not available.

In line with the basic understanding of the framework of imprecise probabilities (see e.g. Walley, 1991; Augustin et al., 2014), researchers should be allowed to specify these quantities (prior, hypotheses, loss) in a less precise, i.e. imprecise, way, such that the scarcity, vagueness, and incompleteness of the available information about these quantities is captured within the specification more accurately. So, instead of requiring these quantities to be precise single-valued entities, they should rather be interval- or set-valued entities.

In that, the prior distribution might be generalized to be a set of prior distributions, the parameter sets in the context of the hypotheses might be generalized to be sets of parameter sets, and the value k of the hypothesis-based loss function L_H might be generalized to be a set of reasonable values for k. With these imprecise specifications, the applied researcher might be more comfortable to determine these essential quantities in accordance with the available scarce, vague, and partial information. How to deal with these imprecise quantities in the hypothesis-based Bayesian decision theoretic framework will be outlined in the following. These elaborations are published within (Schwaferts and Augustin, 2019, 2021c), and a special case of Bayes factors with an imprecisely specified prior distribution is published within (Ebner et al., 2019).

6.1 Imprecise Prior Distribution

While the previous chapters omitted an explicit notation for the prior parameter *distribution* and employed only a notation for the prior parameter *density* $\pi(\theta)$, former will come in handy for the elaboration about the generalization into the framework of imprecise probabilities. In that, denote the prior parameter distribution as π_θ which has the density $\pi(\theta)$.

The set of prior distributions that are all considered to be reasonable in the actual research context shall be denoted by Π_θ. This set is referred to as imprecise prior parameter distribution and constitutes an entity on its own that represents the prior (partial) knowledge about the phenomenon of interest. After observing the data x, this imprecise prior parameter distribution Π_θ gets updated to the imprecise posterior parameter distribution

$$\Pi_{\theta|x} = \{\pi_{\theta|x} | \pi_\theta \in \Pi_\theta\}, \tag{6.1}$$

where each posterior parameter distribution $\pi_{\theta|x}$ was obtained separately from one of the prior parameter distributions $\pi_\theta \in \Pi_\theta$ via Bayes rule (equation (2.1) (in this regard cp. also the generalized Bayes rule (Walley, 1991) and sensitivity considerations in the context of robust statistics (Ríos Insua and Ruggeri, 2012)).

6.2 Imprecise Hypotheses

As elaborated in Chapters 3 and 4, hypotheses are specified such that they incorporate the notion of practical relevance, if the parameter values θ within each hypothesis would favor to the same action, respectively. There might, however, be some parameter values $\theta \in \Theta$ for which the applied researcher is not able to determine the preferred action, especially for those parameter values close to the borders between the hypotheses. Loosening this strict requirement that every parameter value has to be in either of the hypotheses, and allowing that parameter values might be in both, either, or none of the hypotheses, the specification of the hypotheses might be closer to the actual research question.

Formally, this leads to sets

$$[\Theta]_0 := \{\Theta_0 \subset \Theta \mid \Theta_0 \text{ reasonable under } H_0\} \tag{6.2}$$
$$[\Theta]_1 := \{\Theta_1 \subset \Theta \mid \Theta_1 \text{ reasonable under } H_1\} \,, \tag{6.3}$$

of parameter sets for the imprecise hypotheses

$$H_0 : \theta \in [\Theta]_0 \quad \text{vs.} \quad H_1 : \theta \in [\Theta]_1 \,. \tag{6.4}$$

Similarly, these sets $[\Theta]_0$ and $[\Theta]_1$ of parameter sets are considered as entities on their own and as representing the theoretical positions that are contrasted in the research question.

The imprecise posterior probabilities

$$P(H_0|x) = \left\{ p(h_0|x) \,\middle|\, \Theta_0 \in [\Theta]_0, \pi_{\theta|x} \in \Pi_{\theta|x} \right\} \tag{6.5}$$
$$P(H_1|x) = \left\{ p(h_1|x) \,\middle|\, \Theta_1 \in [\Theta]_1, \pi_{\theta|x} \in \Pi_{\theta|x} \right\} \,. \tag{6.6}$$

of the imprecise hypotheses H_0 and H_1 are then derived element-wise by considering every combination of each posterior distribution $\pi_{\theta|x} \in \Pi_{\theta|x}$ and each parameter set $\Theta_0 \in [\Theta]_0$ or $\Theta_1 \in [\Theta]_1$, respectively.

These imprecise posterior probabilities then form the imprecise posterior odds

$$\left[\frac{P(H_0|x)}{P(H_1|x)}\right] := \left\{ \frac{p(h_0|x)}{p(h_1|x)} \,\middle|\, p(h_0|x) \in P(H_0|x), p(h_1|x) \in P(H_1|x) \right\} \,, \tag{6.7}$$

having the supremum

$$\overline{P} := \sup \left[\frac{P(H_0|x)}{P(H_1|x)}\right] \tag{6.8}$$

and the infimum

$$\underline{P} := \inf \left[\frac{P(H_0|x)}{P(H_1|x)}\right] \,. \tag{6.9}$$

6.3 Imprecise Loss

In order to generalize the hypothesis-based loss function L_H such that imprecise specifications are allowed, the value k needs to be considered to be a set-valued quantity K instead. This set K contains all precise values k that are considered to be in line with the available (potentially partial) information about the consequences of the type-I-error and the type-II-error. This set K together with the imprecise hypotheses H_0 and H_1 delineate the imprecise hypothesis-based loss function. Denote the supremum and infimum of this set K as $\overline{K} := \sup K$ and $\underline{K} := \inf K$, respectively. An illustration of such an imprecise hypothesis-based loss function, depicted terms of a parameter-based loss function, is shown in Figure 6.1.

With an element-wise combination of every $k \in K$ and every precise posterior odds $\frac{p(h_0|x)}{p(h_1|x)} \in \left[\frac{P(H_0|x)}{P(H_1|x)}\right]$ within the imprecise posterior odds, the imprecise ratio of expected posterior losses can be calculated:

$$R := \left\{ r = k \cdot \frac{p(h_0|x)}{p(h_1|x)} \;\middle|\; k \in K, \frac{p(h_0|x)}{p(h_1|x)} \in \left[\frac{P(H_0|x)}{P(H_1|x)}\right] \right\} . \tag{6.10}$$

With the supremum and infimum

$$\overline{R} := \sup R = \overline{K} \cdot \overline{P} \tag{6.11}$$

$$\underline{R} := \inf R = \underline{K} \cdot \underline{P} , \tag{6.12}$$

of this imprecise ratio R of expected posterior losses, the set $\mathcal{A}^*$ of optimal actions can be determined by:

$$\mathcal{A}^* = \begin{cases} \{\,\} & \text{if } \underline{R} < 1 < \overline{R} \\ \{a_0\} & \text{if } 1 \leq \underline{R}, 1 < \overline{R} \\ \{a_1\} & \text{if } \underline{R} < 1, \overline{R} \leq 1 \\ \{a_0, a_1\} & \text{if } \underline{R} = \overline{R} = 1 \end{cases} . \tag{6.13}$$

If $\underline{R} < 1 < \overline{R}$, there are possibilities within the imprecise specifications that results in a_0 to be optimal and possibilities within the imprecise specifications that results in a_1 to be optimal. In that, information is insufficient to unambiguously guide the decision and the decision should be withheld. Therefore, no action can be considered as optimal and the set $\mathcal{A}^*$ of optimal actions is empty. If either $1 \leq \underline{R}, 1 < \overline{R}$ or $\underline{R} < 1, \overline{R} \leq 1$, then all possibilities within the imprecise specifications yield the same action a_0 or a_1, respectively, to be optimal and the decision can be guided unambiguously. The condition $\underline{R} = \overline{R} = 1$ arises only if the ratio of expected posterior losses is actually precise. In that sense, equation (6.13) can be seen as a generalization of equation (4.8).

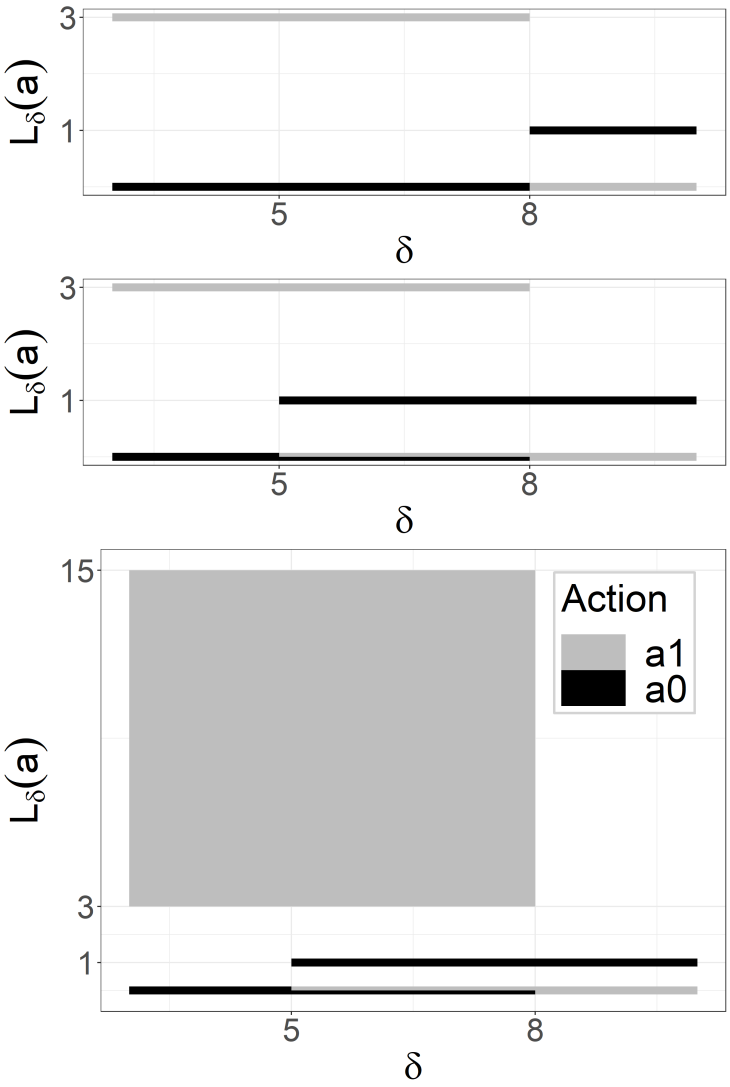

Figure 6.1: Simplified Loss Function within the Imprecise Framework. The top plot depicts a precise simplified loss function with $k = 3$ on the parameter δ of interest for the precise hypotheses defined by $\Theta_0 = (-\infty, 8]$ and $\Theta_1 = (8, \infty)$ for both actions a_0 (black) and a_1 (gray). The middle plot depicts this loss function with an imprecise hypothesis specification by $H_0 : \delta \in [\Delta]_0 = \left\{\Delta_0 = (-\infty, \tilde{\delta}] \,\middle|\, \tilde{\delta} \in [5, 8]\right\}$ and $H_1 : \delta \in [\Delta]_1 = \left\{\Delta_1 = [\tilde{\delta}, \infty) \,\middle|\, \tilde{\delta} \in [5, 8]\right\}$. Note the overlapping lines within $\delta = [5, 8]$. The bottom plot does consider an imprecisely specified loss function with $K = [3, 15]$ in addition to the imprecise hypotheses H_0 and H_1. This figure was taken from (Schwaferts and Augustin, 2021c).

If the decision should be withheld, i.e. if $\underline{R} < 1 < \overline{R}$, then additional information is required to eliminate the uncertainty that prohibits an unambiguous conclusion. By collecting additional data, additional information about the phenomenon of interest, about the differentiation between the contrasted theoretical positions, or about the consequences

of the type-I-error or the type-II-error, the imprecise ratio R of expected posterior losses might be narrowed down to exclude the number 1, allowing to guide the decision unambiguously.

Chapter 7

Framework for Applications

Now, after generalizing Bayes factors to both the decision theoretic framework and the framework of imprecise probabilities, a simple and straightforward framework shall be outlined that can be applied in empirical studies.

On the one hand, it appears that applying a comprehensive decision theoretic account in an empirical context typically fails due to requirement of specifying the loss function. On the other hand, it appears that applying the framework of imprecise probabilities in an empirical context might happen to fail due to potentially complicated calculations.

However, by using a hypothesis-based analysis, i.e. by accepting the simplification assumption (Section 4.2), and using the imprecise specification with K, the requirements for the specification of the loss function are extremely mitigated. Further, it appears that complicated calculations in the context of the imprecise hypothesis-based Bayesian decision theoretic account are primarily centered around the parameter distributions and the calculation of the posterior odds. By using a precise prior distribution and precise hypotheses, but an imprecise loss function L_H with an interval-valued specification of K, calculations are hardly more complicated than in the precise case. In that, a framework might be outlined that uses an imprecise loss function in an otherwise precise setting, therefore mitigating the requirements on the loss function as much as possible, but still keeping calculations feasible.

Certainly, including imprecise specifications of the prior distribution and the hypotheses is expected to provide more reliable results and is recommended if reasonable (i.e. if available relevant information is scarce, vague, and partial) and feasible. How to do so was elaborated on in Chapter 6. Yet, it should be considered that the specification of a prior distribution and of the hypotheses is an integral part of every analysis that is based on Bayes factors and the status quo is to employ precise specifications thereof. Further, the status quo is also to refrain from considering a loss function – or, more generally, the consequences of

certain decisions that are based on Bayes factor results. In order to improve the practical relevance of Bayes factors, one has to start at this status quo and ask, which improvement in methodology is the most suitable to achieve this goal. The suitability of an methodological improvement does, of course, include its ease in applications. And by merely including an imprecise loss function into the analysis with Bayes factors (and not imprecise prior distributions or hypotheses), computations will still have a similar level of feasibility, but an essential step was taken towards an improvement of the practical relevance of Bayes factors: The practical relevance of Bayes factors (or any result) simply depends on what they are used for (Chapter 3). Formally, this use is a decision and a decision cannot be guided without considering (the badness of) its consequences. Yet without considering a loss function, the consequences of the decision are ignored at all, and the practical relevance cannot be assessed on a formal level. So, to establish a connection between the practical purpose and the Bayes factors, a loss consideration is mandatory. And the simplified hypothesis-based loss function with imprecise K allows to include such a loss consideration into the analysis without overloading the applied scientist with too many specification requirements (in contrast to a full precise parameter-based loss function L).

7.1 Framework

This hypothesis-based Bayesian decision theoretic framework with imprecise loss function is outlined in detail within the contribution (Schwaferts and Augustin, 2021b).

In essence, the Bayesian analysis of the data x follows the typical Bayesian structure (as outlined in Chapter 2): In the context of the scientific investigation that is described by the parametric sampling distribution $f(x|\theta)$, the applied scientist has to specify the prior density $\pi(\theta)$ and the hypotheses h_0, h_1 (equation (2.3)) to be in accordance with the prior knowledge and the theoretical positions of interest, respectively. The prior density gets updated via Bayes rule (equation (2.1)) to the posterior density $\pi(\theta|x)$, which in turn allows to calculate the posterior probabilities of the hypothesis $p(h_0|x)$, $p(h_1|x)$, forming the posterior odds.

Considering the intended use of the scientific investigation, actions a_0, a_1 need to be stated and their consequences in dependence of the respective hypotheses h_0, h_1 assessed. This leads to the specification of the ratio k of loss values for the type-I-error and the type-II-error. This specification might be provided as an interval[1] $K = [\underline{K}, \overline{K}]$, for which the

[1]While the elaborations in Chapter 6 used a set-valued specification of K, this framework is depicted with an interval-valued specification of K, because it is to expect that applied scientists might be able to consider an upper and a lower bound of this ratio, rating every value in between as plausible as well. If there are certain values of k in between these bounds, they might be excluded leading to a set-valued (and not interval-valued) specification. As the resulting set $\mathcal{A}^*$ of optimal actions (equation (6.13)) depends only on the supremum and the infimum of the set K (which are also the bounds $\underline{K}$ and $\overline{K}$, if K is an

applied scientists has to specify the maximal and minimal reasonable values for k.

The supremum and the infimum of the imprecise ratio R of expected posterior losses (equation (6.10)) can now be calculated by (compare equations (6.11) and (6.12))

$$\overline{R} = \overline{K} \cdot \frac{p(h_0|x)}{p(h_1|x)} \tag{7.1}$$

$$\underline{R} = \underline{K} \cdot \frac{p(h_0|x)}{p(h_1|x)}\,, \tag{7.2}$$

due to the precise nature of the posterior odds. The optimal action can then, again, be found by simply comparing these two values (equation (7.1) and equation (7.2)) with the value 1, as in equation (6.13).

7.2 Step-By-Step Guide

A step-by-step guide for this framework intends to provide an orientation for applied scientist who are interested in increasing the practical relevance of their hypothesis-based Bayesian analyses. This guide is provided in (Schwaferts and Augustin, 2021b) and mostly adopted verbatim here:

Step 1: Actions. First of all, the researcher needs to specify the actions. It is recommended to explicitly state and report these actions, e.g.[2] by

a_0: do not administer aspirin to prevent myocardial infarction

a_1: administer aspirin to prevent myocardial infarction

If the researcher has difficulties stating the actions, maybe there is no decision to guide and a descriptive analysis might suffice.

Step 2: Sampling Distribution. Next, the researcher should provide a detailed description of the investigation and how it is characterized (i.e. the sampling distribution). It is recommended to also explicitly state the employed parameter θ and its interpretation. This is the basis for specifying the hypotheses.

Step 3: Prior Distribution. In the Bayesian setting, it is possible to include prior knowledge about the phenomenon of interest into the analysis. In that, the researcher has to specify a prior distribution on the parameter. It is recommended to fully report the

interval), this will not affect the results.

[2]The example here (as well as the example in Step 6) refers to Bartolucci et al. (2011).

available prior knowledge about the parameter θ and why this leads to the prior density $\pi(\theta)$.

It is recommended at this step of the analysis to also state all other possible prior densities that are in accordance with the available prior knowledge, as these serve as basis for a subsequent sensitivity analysis or an analysis with an imprecisely specified prior distribution (as outlined in Chapter 6).

Naturally, also non-informative priors might be specified and they might also be improper (as long as they lead to proper posterior distributions).

Step 4: Assumption. If the researcher is unable to specify the loss function L, then a simplification as in Section 4.2 might be a solution. This simplification is an assumption on the loss function, namely that the loss function is constant within each of two parameter sets. If this assumption is not appropriate, it might lead to errors (which are inherent to every hypothesis-based analysis in the context of a practical purpose) and the researcher needs to be aware of this consequence. It is recommended to explicitly report that this assumption was made. Transparency is one of the basic principles in science (cp. Gelman and Hennig, 2017).

Step 5: Hypotheses. Now, the researcher has to consider each possible parameter value θ and assess which action should be preferred if this parameter value would be true. All parameters for which a_0 or a_1 should be preferred are comprised within the sets Θ_0 or Θ_1, respectively. Certainly, there are parameter values that define the border between both sets Θ_0 and Θ_1. It is recommended to explicitly state what these values mean in real life and why they define reasonable borders between Θ_0 and Θ_1.

Step 6: Errors. Deciding for a_1 if $\theta \in \Theta_0$ is the type-I-error and deciding for a_0 if $\theta \in \Theta_1$ is the type-II-error. Both errors should be delineated, as they serve as basis for specifying the ratio k. It is recommended to explicitly state these errors and their consequences, e.g. by

> Type-I-error: administer aspirin to prevent myocardial infarction, but the effect is negligible. Consequence: patients unnecessarily suffer side effects of aspirin.
>
> Type-II-error: do not administer aspirin to prevent myocardial infarction, although it would have an effect. Consequence: some patients suffer a myocardial infarction, which could have been prevented.

Of course, this is only a schematic illustration and in real empirical studies these elaborations will be more comprehensive.

Step 7: Loss Magnitude. The researcher has to imagine that the "badness" of deciding correctly is 0. In this context, the researcher has to determine how much worse the type-

I-error is compared to the type-II-error. This is the value k. As a precise value for k is difficult to determine, it might be easier to specify a range $[\underline{K}, \overline{K}]$ of plausible values for k. It is recommended to report all considerations that lead to this specification.

Step 8: Investigation. Now, the investigation can be conducted and it is recommended to preregister[3] the previous specifications, the design of the experiment, and the planned (decision theoretic) analysis of the data (cp. Nosek et al., 2018; Klein et al., 2018). Registered reports[4] even allow to obtain a peer-review prior to collecting the data.

Step 9: Posterior Distribution. The observed data are used to obtain the posterior distribution as well as the posterior probabilities of the hypotheses $p(h_0|x)$ and $p(h_1|x)$. There are countless references on how to do this (e.g. Gelman et al., 2013; Kruschke, 2015).

Step 10: Optimal Action. The researcher has to calculate $\underline{R}$ and $\overline{R}$ as in equations (7.2) and (7.1) to find the optimal action as in equation (6.13).

For $\underline{R} < 1 < \overline{R}$, the decision should be withheld, because the data or the information about the decision problem are not sufficient to unambiguously guide the decision. In this case, a reasonable strategy might be to collect more data or to gather more information about the decision problem, especially about the consequences of the errors, to narrow down $[\underline{K}, \overline{K}]$. However, it is recommended to transparently report that a decision was withheld at first and which subsequent steps were taken to obtain more information.

Step 11: Publish Data. Of course, other researchers might need the data to guide their decisions. It is to expect that they have different prior knowledge and that their decisions employ different hypotheses. Without having access to the data set (but only to the reported analysis), it might be difficult, or even impossible, for them to guide their decisions properly, emphasizing the importance of open science[5].

7.3 Comparison

In relation to Bayes factors, this hypothesis-based Bayesian decision theoretic framework with imprecise loss function has two new characteristics: First, there is a direct relation with the practical purpose of the scientific investigation on a formal level. Second, scarcity, vagueness, or incompleteness about the consequences of respective decisions might yield unambiguous results that do not pretend a level of precision which is not available and advise to withhold the decision (until more information or data were obtained).

[3] Study designs can be preregistered e.g. at `www.cos.io/initiatives/prereg`.

[4] Information about registered reports can be found e.g. at `www.cos.io/rr`.

[5] Comprehensive information about open science are provided e.g. by the LMU Open Science Center: `www.osc.uni-muenchen.de`.

In the current literature about hypothesis-based Bayesian analyses in context of applied sciences, there is another methodology that also stands out with exactly these characteristics: The so called HDI+ROPE decision rule (Kruschke, 2015, 2018).

In the context of this decision rule, it is assumed that the researcher is interested in a certain parameter null value θ^* and whether to accept or to reject this value for practical purposes. To do so, the researcher needs to specify a region of practical equivalence[6] (ROPE) around this null value θ^*, which is a "range of parameter values that are equivalent to the null value for practical purposes" (Kruschke, 2018, p. 272). After observing the data x, the 95%-highest density interval (HDI) of the posterior density $\pi(\theta|x)$ is calculated and compared to the ROPE:

- If the HDI falls completely inside the ROPE, then accept the null value for practical purposes.
- If the HDI falls completely outside the ROPE, then reject the null value for practical purposes.
- Else, withhold a decision.

In order to compare this decision rule with the hypothesis-based Bayesian decision theoretic framework with imprecise loss, the decision theoretic foundation of the HDI+ROPE decision rule was derived in contribution (Schwaferts and Augustin, 2020). It appears that the outcome of withholding a decision in the context of the HDI+ROPE decision rule cannot be derived from an imprecise specification. In that, scarcity, vagueness, and incompleteness of the available information about the essential quantities do not seem to be included into the analysis. The indecisiveness seems to arise solely from observed data sets that do not contain enough information. Further, it appears that the underlying loss function behind this decision rule is artificial: Its values are set such that certain decision will be derived under certain circumstances, but its values are not primarily in line with the actual underlying decision problem. In that, it seems that the HDI+ROPE decision rule includes the practical purpose only in the specification of the ROPE, but not in formalizing the "badness" of consequences within the loss function. In that, an important link between the practical purpose and the final decision (result of the analysis) is not available in the HDI+ROPE decision rule: Practical consequences of incorrect decisions were not included into the formal analysis.

As a result of this comparison of the decision theoretic foundation of the HDI+ROPE decision rule with the hypothesis-based decision theoretic framework with imprecise loss, it appears that latter is closer to the practical purpose of the investigation and is better

[6]Using the notation of the present thesis, this region of practical equivalence appears to relate to the parameter set Θ_0 of the hypothesis h_0.

capable of including scarcity, vagueness, and incompleteness of the available information about the consequences of incorrect decision into the analysis.

Consequently, it is recommended to prefer the hypothesis-based Bayesian decision framework with imprecise loss for a Bayesian analysis in the context of the practical purpose of a scientific investigation.

Chapter 8

Conclusion

A formal definition of the notion of practical relevance was established (Chapter 3), outlining that formal considerations about the practical relevance of effects or about the practical relevance w.r.t. hypotheses require an underlying decision problem (i.e. a formal description of the practical purpose of the scientific study). In that sense, Bayes factors were depicted within the greater context of hypothesis-based Bayesian decision theory (Chapter 4) and generalized into the framework of imprecise probabilities by allowing an imprecise specification of the prior distribution, the hypotheses, and the hypothesis-based loss function (Chapter 6). In order to condense these generalizations into a simple and straightforward framework for applications, only an imprecise loss function was employed in the hypothesis-based Bayesian decision theoretic framework, ensuring feasible calculations and a relation of the analysis to the underlying decision problem (Chapter 7).

Yet, besides developing these twofold generalizations of Bayes factors into the decision theoretic framework and the framework of imprecise probabilities, one fundamental but important question was continuously addressed in the course of this thesis: What is a statistical hypothesis? The answer to this question is multifaceted and requires elaborations from several different points of view:

- From a definition point of view, statistical hypotheses (in the context of a parametric sampling distribution) are *subsets of the parameter space*. There are different conceptions (primarily elaborated in Vanpaemel, 2010; Vanpaemel and Lee, 2012) about this mathematical nature of hypotheses (namely, as sets of parameters together with prior distributions), yet it was outlined within this thesis that these conceptions are problematic (Chapter 5).
- From a point of view about the employment of hypotheses in statistical analyses, statistical hypotheses are *mathematical objects with a contrasting nature*. In the context of most statistical analyses, there are (at least) two hypotheses that are

compared against each other.

- From an interpretation point of view, statistical hypotheses are the *formalization of theories*[1] that are statistically evaluated. These are hypothetical conjectures and might be true or false. The purpose of the scientific investigation is to derive statements about these conjectures, such as whether they are in line with the data, whether they might be falsified, or which probabilities they might be assigned with. The statistical analysis does not change these conjectures themselves, it only provides their assessments.

- From a decision theoretic point of view, statistical hypotheses are a *component in the simplification of a loss function.* In that sense, the employment of hypotheses corresponds to accepting the simplification assumption about the shape of the loss function (Section 4.2). With more information to specify the non-simplified parameter-based loss function, a hypothesis-based analysis is not necessary.

Finally, one last aspect needs to be brought into a conclusion: The evaluation of the interpretations of Bayes factors. In Section 2.3, the two interpretations of Bayes factors as comparison of prior predictive performances and as evidence were outlined. Assume an empirical scientist conducts an investigation and calculates a Bayes factor. And now? What is it that the empirical scientist can do with the Bayes factor? To answer this question, consider both interpretations:

- *Comparison of prior predictive performances.* The Bayes factor states how much higher the prior predictive performance of one hypothesis is compared to the other hypothesis. However, it is not possible to use the prior predictive likelihood of that hypothesis with the higher prior predictive performance for future prediction. As illustrated in the context of updating consistency (Section 5.1), the prior gets updated by seeing the data to the posterior. In that, the prior predictive likelihoods are already outdated when calculating a Bayes factor. Consistent Bayesian predictions use posterior predictive likelihoods for future predictions, once data were already observed. Consequently, although interpreted w.r.t. prior predictive performances, Bayes factors cannot be usefully employed when future predictions are of interest.

- *Evidence.* The Bayes factor states how much more the data favor one hypothesis over the other hypothesis. In that sense, the data are interpreted as evidence and the Bayes factor quantifies this evidence. Evidence changes beliefs, as delineated in equation (2.20). Interestingly, there is no other formula than this equation that states how to use Bayes factors. Accordingly, there is no other use for Bayes factors than to change the beliefs in the hypotheses! The question of what to do with Bayes factors leads

[1] Within this thesis, the term theoretical position was employed to emphasize both its contrasting nature and its interpretation as theory.

to the question of what to do with the posterior odds. In general, posterior odds are neither 0 nor ∞, and there is still a non-zero belief in both hypotheses. So falsifying and, therefore, dismissing one of the hypotheses cannot be derived solely from the posterior odds in a mathematically correct manner. Accepting one hypothesis and dismissing the other hypothesis is thus a subsequent step that comes after calculating the posterior odds. In effect, it is a decision about the hypotheses. And it is decision theory that elaborates on how to guide it properly.

In summary, the Bayes factor is a quantity that appears within the determination of the optimal action in a hypothesis-based Bayesian decision problem, and even its interpretation as evidence emphasizes its involvement in guiding decisions. While some decisions might be of mere scientific interest (e.g. which of the theoretical positions to dismiss and which to pursue in further scientific investigations), others might be related to a practical purpose. By including information about this practical purpose into the analysis in the potentially vague form it is available, a researcher is able to improve the practical relevance of Bayes factors.

Bibliography

Abdellaoui M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46(11):1497–1512. URL http://dx.doi.org/10.1287/mnsc.46.11.1497.12080.

Augustin T., Coolen F.P., de Cooman G., and Troffaes M.C.M., editors (2014). *Introduction to Imprecise Probabilities*. John Wiley, Chichester. URL http://dx.doi.org/10.1002/9781118763117.

Bartolucci A.A., Tendera M., and Howard G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *The American Journal of Cardiology*, 107(12):1796–1801. URL http://dx.doi.org/10.1016/j.amjcard.2011.02.325.

Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition. URL http://dx.doi.org/10.1007/978-1-4757-4286-2.

Berger J.O. and Wolpert R.L. (1988). The likelihood principle. *Lecture Notes-Monograph Series*, 6:iii–160.2 (discussion: 160.3–199). URL http://www.jstor.org/stable/4355509.

Berkson J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536. URL http://dx.doi.org/10.2307/2279690.

Chajewska U., Koller D., and Parr R. (2000). Making rational decisions using adaptive utility elicitation. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI)*, pages 363–369. Austin, TX.

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

Diener E. and Oishi S. (2000). Money and happiness: Income and subjective well-being across nations. In *Culture and Subjective Well-Being*, pages 185–218. MIT Press, Cambridge, MA.

Dienes Z. (2019). How do i know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, 2(4):364–377. URL http://dx.doi.org/10.1177/2515245919876960.

Dienes Z. and Mclatchie N. (2018). Four reasons to rrefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 25(1):207–218. URL http://dx.doi.org/10.3758/s13423-017-1266-z.

Ebner L., Schwaferts P., and Augustin T. (2019). Robust Bayes factor for independent two-sample comparisons under imprecise prior information. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 167–174. PMLR. URL http://proceedings.mlr.press/v103/ebner19a.html.

Ellis S.M. and Steyn H.S. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): Research note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, 12(4):51–53. URL http://dx.doi.org/10.4300/JGME-D-12-00156.1.

Froberg D.G. and Kane R.L. (1989). Methodology for measuring health-state preferences—ii: Scaling methods. *Journal of Clinical Epidemiology*, 42(5):459–471. URL http://dx.doi.org/10.1016/0895-4356(89)90136-4.

Gelman A. and Hennig C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180:967–1033. URL http://dx.doi.org/10.1111/rssa.12276.

Gelman A., Stern H.S., Carlin J.B., Dunson D.B., Vehtari A., and Rubin D.B. (2013). *Bayesian Data Analysis*. Chapman & Hall. URL http://dx.doi.org/10.1201/9780429258411.

Gigerenzer G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606. URL http://dx.doi.org/10.1016/j.socec.2004.09.033.

Gönen M., Johnson W.O., Lu Y., and Westfall P.H. (2005). The Bayesian two-sample t test. *The American Statistician*, 59:252–257. URL http://dx.doi.org/10.1198/000313005X55233.

Gronau Q.F., Ly A., and Wagenmakers E.J. (2019). Informed Bayesian t-tests. *The American Statistician*. URL http://dx.doi.org/10.1080/00031305.2018.1562983.

Hilbert S., McAssey M., Bühner M., Schwaferts P., Gruber M., Goerigk S., and Taylor P.C.J. (2019). Right hemisphere occipital rTMS impairs working memory in visualizers but not in verbalizers. *Scientific Reports*, 9(1):1–8. URL http://dx.doi.org/10.1038/s41598-019-42733-6.

Hojat M. and Xu G. (2004). A visitor's guide to effect sizes – statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education*, 9(3):241–249. URL http://dx.doi.org/10.1023/B:AHSE.0000038173.00909.f6.

Ioannidis J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Jaynes E.T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press. URL http://dx.doi.org/10.1017/CBO9780511790423.

Jeffreys H. (1961). *Theory of Probability*. Oxford University Press, Oxford, third edition.

Kass R.E. (2011). Statistical inference: The big picture. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 26(1):1–9. URL http://dx.doi.org/10.1214/10-STS337.

Kass R.E. and Raftery A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795. URL http://dx.doi.org/10.2307/2291091.

Kirk R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5):746–759. URL http://dx.doi.org/10.1177/0013164496056005002.

Kirk R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61(2):213–218. URL http://dx.doi.org/10.1177/00131640121971185.

Klein O., Hardwicke T.E., Aust F., Breuer J., Danielsson H., Mohr A.H., IJzerman H., Nilsonne G., Vanpaemel W., Frank M.C., and Frank M.C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, 4(1):20(1–15). URL http://dx.doi.org/10.1525/collabra.158.

Kruschke J.K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan*. Academic Press, New York. URL http://dx.doi.org/10.1016/B978-0-12-405888-0.09999-2.

Kruschke J.K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280. URL http://dx.doi.org/10.1177/2515245918771304.

Kruschke J.K. and Liddell T.M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206. URL http://dx.doi.org/10.3758/s13423-016-1221-4.

Ly A., Verhagen J., and Wagenmakers E.J. (2016). Harold jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32. URL http://dx.doi.org/10.1016/j.jmp.2015.06.004.

Morey R.D., Romeijn J.W., and Rouder J.N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18. URL http://dx.doi.org/10.1016/j.jmp.2015.11.001.

Morey R.D. and Rouder J.N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4):406–419. URL http://dx.doi.org/10.1037/a0024377.

Nosek B.A., Ebersole C.R., DeHaven A.C., and Mellor D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606. URL http://dx.doi.org/10.1073/pnas.1708274114.

Ríos Insua D. and Ruggeri F., editors (2012). *Robust Bayesian Analysis*. Springer Science & Business Media. URL http://dx.doi.org/10.1007/978-1-4612-1306-2.

Robert C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, second edition. URL http://dx.doi.org/10.1007/0-387-71599-1.

Rouder J.N., Haaf J.M., and Aust F. (2018a). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, 85(1):41–56. URL http://dx.doi.org/10.1080/03637751.2017.1394581.

Rouder J.N., Haaf J.M., and Vandekerckhove J. (2018b). Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1):102–113. URL http://dx.doi.org/10.3758/s13423-017-1420-7.

Rouder J.N. and Morey R.D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18(4):682–689. URL http://dx.doi.org/10.3758/s13423-011-0088-7.

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

Rüger B. (1998). *Test-und Schätztheorie: Band I: Grundlagen*. De Gruyter Oldenbourg.

van de Schoot R., Winter S.D., Ryan O., Zondervan-Zwijnenburg M., and Depaoli S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2):217–239. URL http://dx.doi.org/10.1037/met0000100.

Schwaferts P. and Augustin T. (2019). Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 338–345. PMLR. URL http://proceedings.mlr.press/v103/schwaferts19a.html.

Schwaferts P. and Augustin T. (2020). Bayesian decisions using regions of practical equivalence (ROPE): Foundations. Technical Report 235, Ludwig-Maximilians-University Munich, Department of Statistics. URL http://dx.doi.org/10.5282/ubm/epub.74222.

Schwaferts P. and Augustin T. (2021a). Bayes factors can only quantify evidence w.r.t. sets of parameters, not w.r.t. (prior) distributions on the parameter. URL http://arxiv.org/abs/2110.09871.

Schwaferts P. and Augustin T. (2021b). How to guide decisions with bayes factors. URL http://arxiv.org/abs/2110.09981.

Schwaferts P. and Augustin T. (2021c). Imprecise hypothesis-based Bayesian decision making with composite hypotheses. In A. Cano, J. De Bock, E. Miranda, and S. Moral, editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, page 280–288. PMLR. URL https://proceedings.mlr.press/v147/schwaferts21a.html.

Schwaferts P. and Augustin T. (2021d). Practical relevance: A formal definition. URL http://arxiv.org/abs/2110.09837.

Schwaferts P. and Augustin T. (2021e). Updating consistency in Bayes factors. Technical Report 236, Ludwig-Maximilians-University Munich, Department of Statistics. URL http://dx.doi.org/10.5282/ubm/epub.75073.

Schwaferts C., Schwaferts P., von der Esch E., Elsner M., and Ivleva N.P. (2021). Which particles to select, and if yes, how many? subsampling methods for Raman microspectroscopic analysis of very small microplastic. *Analytical and Bioanalytical Chemistry*, 413(14):3625–3641. URL http://dx.doi.org/10.14459/2021mp1596628.

Tendeiro J.N. and Kiers H.A.L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological methods*, 24(6):774–795. URL http://dx.doi.org/10.1037/met0000221.

Thompson B. (2002). 'statistical','practical,' and 'clinical': How many kinds of significance do counselors need to consider? *Journal of Counseling & Development*, 80(1):64–71. URL http://dx.doi.org/10.1002/j.1556-6678.2002.tb00167.x.

Vanpaemel W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54(6):491–498. URL http://dx.doi.org/10.1016/j.jmp.2010.07.003.

Vanpaemel W. and Lee M.D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, 19(6):1047–1056. URL http://dx.doi.org/10.3758/s13423-012-0300-4.

Vaske J.J. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife*, 7(4):287–300. URL http://dx.doi.org/10.1080/10871200214752.

Walley P. (1991). *Statistical Reasoning With Imprecise Probabilities*. Chapman & Hall, London.

Part B

Contributions

Contribution 1

Schwaferts & Augustin (2021d): Practical Relevance: A Formal Definition. (Preprint)

Practical Relevance: A Formal Definition

Patrick Schwaferts Thomas Augustin

patrick.schwaferts@stat.uni-muenchen.de
thomas.augustin@stat.uni-muenchen.de

Ludwig-Maximilians-Universität Munich
Department of Statistics
Methodological Foundations of Statistics and its Applications
Ludwigsstraße 33, 80539 Munich, Germany

Abstract

There is a general agreement that it is important to consider the practical relevance of an effect in addition to its statistical significance, yet a formal definition of practical relevance is still pending and shall be provided within this paper. It appears that an underlying decision problem, characterized by actions and a loss function, is required to define the notion of practical relevance, rendering it a decision theoretic concept. In the context of hypothesis-based analyses, the notion of practical relevance relates to specifying the hypotheses reasonably, such that the null hypothesis does not contain only a single parameter null value, but also all parameter values that are equivalent to the null value on a practical level. In that regard, the definition of practical relevance is also extended into the context of hypotheses. The formal elaborations on the notion of practical relevance within this paper indicate that, typically, a specific decision problem is implicitly assumed when dealing with the practical relevance of an effect or some results. As a consequence, involving decision theoretic considerations into a statistical analysis suggests itself by the mere nature of the notion of practical relevance.

Keywords: Practical Significance, Practical Relevance, Nil Hypothesis, Decision Theory, Reproducibility Crisis, Null Hypothesis Significance Testing

1 Introduction

More than twenty years have passed since Kirk (1996) urged to consider the practical relevance of research results in addition to their statistical significance with his paper entitled "Practical Significance: A Concept Whose Time Has Come," yet no formal definition of this concept is currently available. Instead, merely stating that effect sizes are "measures" (see e.g. Ellis and Steyn, 2003) or "indices" (see e.g. Thompson, 2002; Hojat and Xu, 2004) of practical significance which indicate if results are "meaningful" (see e.g. Vaske, 2002) or "useful" (see e.g. Kirk, 1996) seemed to be sufficient. This, however, is by no means a proper mathematical incorporation of the notion of practical relevance (or practical significance) within the frameworks of statistical methodologies. In that, this paper attempts to provide a formal definition of practical relevance.

There are two different lines of research that lead to a definition of practical relevance, which

appear to be closely related. The first one is mentioned above and directly concerned with the practical relevance of an effect (see e.g. Kirk, 1996). The second one is based on the criticism (see e.g. Cohen, 1994) that null hypotheses within the omnipresent approach of null hypothesis significance testing (NHST) typically hypothesize a single parameter value representing a zero effect size, yet this is not of interest as it does not matter if the effect is literally zero, but only practically zero (e.g. a parameter value of, say, 0.01 is not exactly zero but might be practically equivalent to zero in most cases). In that regard, currently promoted statistical methods use reasonably specified null hypotheses, considering smallest effect sizes of interest in equivalence tests (see e.g. Lakens, 2017; Lakens et al., 2018), regions of practical equivalence (ROPE) around the null value in Bayesian decision rules (see e.g. Kruschke, 2015, 2018), or interval-valued null hypotheses in the context of Bayes factors (see e.g. Morey and Rouder, 2011; Hoijtink et al., 2019; Heck et al., 2020).

By evaluating both of these lines of research, it seems that practical relevance can only be described by referring to a (potentially implicitly assumed) decision problem in which one of two actions should be chosen, one being associated with an effect that is practically zero and one being associated with an effect that is practically relevant (i.e. non-zero). In that regard, the context of decision making is necessary to formalize this situation and to provide a definition of practical relevance. In accordance with both of these lines of research, two different definitions of practical relevance might be distinguished, one referring to effects (Section 2) and one referring to hypotheses (Section 3). The implications of these definitions for applied sciences will be discussed (Section 4).

2 Practical Relevance of an Effect

2.1 Context

Practical significance is typically introduced in research papers by stating that effect sizes are measures of it which indicate the importance of the result. However, there is a general agreement that effect sizes are not synonymous with practical significance (see e.g. Vaske, 2002; Peeters, 2016) and that there is more to practical significance than the mere size of the effect. In this regard, Kirk (2001, p. 213, line breaks added) states that "[r]esearchers want to answer three basic questions:

(a) Is an observed effect real or should it be attributed to chance?

(b) If the effect is real, how large is it? and

(c) Is the effect large enough to be useful?"

The first question (a) might be answered by assessing the uncertainty within the observed effect, which is conventionally done by conducting a statistical test, although recently different approaches are promoted, such as Bayesian statistics (see e.g. van de Schoot et al., 2017) or estimation methods that acknowledge the available uncertainty, such as confidence intervals (see e.g. Cumming, 2014). The second question (b) might be answered by the effect size estimate, however, the answer to the third question (c) is more difficult. The usefulness of an effect naturally depends on what it is used for.

Consider the following stereotypical example that is aptly illustrated by Sullivan and Feinn (2012, p. 279):

> A commonly cited example of this problem is the Physicians Health Study of aspirin to prevent myocardial infarction (MI). [(Bartolucci et al., 2011)] In more than 22000 subjects over an average of 5 years, aspirin was associated with a reduction in MI (although not in overall cardiovascular mortality) that was highly statistically significant: $P < .00001$. The study was terminated early due to the conclusive evidence, and aspirin was recommended for general prevention. However, the effect size was very small: a risk difference of 0.77% with $r^2 = .001$ – an extremely small effect size. As a result of that study, many people were advised to take aspirin who would not experience benefit yet were also at risk for adverse effects. Further studies found even smaller effects, and the recommendation to use aspirin has since been modified.

Similar examples can be found easily, e.g. analogously about whether a medication should be administered or not in the context of a certain disease (Baicus and Caraiola, 2009), about the gain in knowledge and the ability to think critically of university students in the context of deciding about different teaching methods (Peeters, 2016), or fictitiously about assessing IQ differences between two arbitrary groups of students that might lead to decisions about where to erect new schools for talented students (Thompson, 1993).

All those examples have two characteristics in common. First, a decision has to be be guided (e.g. administer aspirin to prevent MI or not), such that the usefulness of the reported effect can be assessed w.r.t. this decision, creating the framework to answer question (c). Second, there is agreement on the lack of practical relevance of the reported effect, such that it seems reasonable to decide as if no effect was present. Accordingly, a way to determine how to decide for each possible effect is implicitly employed, which allows to implicitly answer question (c) in the framework of the decision of interest.

Both a decision and a way to decide for each possible effect are central components of statistical decision theory (see e.g. Berger, 1985; Robert, 2007). Therefore, it suggests itself to employ decision theoretic concepts in order to define the practical relevance of an effect.

2.2 Formal Definition

Assume the observed data y are modelled as realization of the random variable Y with parametric density $f(y|\theta, \varphi)$, where θ is the effect parameter of interest (e.g. a standardized or non-standardized difference in means between two groups or a correlation between two features) and φ is a vector of nuisance parameters being not of interest. Without loss of generality, the effect parameter θ is such that $\theta = 0$ indicates the absence of an effect (else available parameters might be transformed accordingly).

A decision should be guided between two actions a_0 and a_1, where a_0 should denote the action that is appropriate if the effect is absent, i.e. if $\theta = 0$.

For each possible effect $\theta \in \Theta$ within the parameter space Θ, each of both actions has certain consequences and the "badness" of these consequences is quantified by a loss function

$$L : \Theta \times \mathcal{A} \to \mathbb{R}_0^+ : (\theta, a) \mapsto L(\theta, a)\,, \tag{1}$$

where $\mathcal{A} = \{a_0, a_1\}$ is the action space. So, e.g. $L(0, a_0)$ and $L(0, a_1)$ quantify how bad it would be to decide for a_0 and a_1, respectively, if $\theta = 0$ is true.

The loss function might be seen as representing the consequences of the applied decision on a formal level. In order to obtain such a loss function for a certain decision problem, one needs to think about the consequences of deciding for a_0 or a_1 in dependence of the parameter value θ and then find a function L that mathematically represents the badness of those consequences. Naturally, information about these consequences might be vague, such that specifying a loss function unambiguously might be difficult, as many different loss functions might be in accordance with the available vague information. However, the fact that it might be difficult to specify such a loss function unambiguously does not prohibit that an appropriate mathematical formulation of the notion of practical relevance is embedded within a decision theoretic context. The aim of this paper is to derive the concept of practical relevance on a formal level to gain a better mathematical understanding of it. Therefore, assume for now that such a loss function is available.

For each effect θ the action with smaller loss should be preferred. Therefore, one of the following holds for each effect θ:

- If $L(a_0, \theta) < L(a_1, \theta)$, then a_0 is preferred over a_1.
- If $L(a_1, \theta) < L(a_0, \theta)$, then a_1 is preferred over a_0.
- If $L(a_0, \theta) = L(a_1, \theta)$, then there is no preference between a_0 and a_1.

Due to choosing a_0 to be appropriate in the absence of an effect, $L(0, a_0)$ is smaller than $L(0, a_1)$. Intuitively, other effects θ that also prefer a_0 over a_1 cannot be practically relevant as they lead to the same decision as the absence of an effect. Consequently, it must be those effects θ that prefer a_1 over a_0 which are practically relevant.

Summing up, these considerations lead to the definition of the practical relevance of an effect:

Definition 1 (Practical Relevance of an Effect)**.** *Within this framework, an effect θ is practically relevant (or practically significant) w.r.t. the actions a_0, a_1, and the corresponding loss function L, if a_1 is preferred over a_0, i.e. if*

$$L(\theta, a_1) < L(\theta, a_0)\,, \tag{2}$$

else the effect θ is negligible (or practically zero) w.r.t. these actions and this loss function.

The terms practical relevance and practical significance shall be used interchangeably, because although these considerations arise from the frequentist literature about the practical significance of an effect, they also apply to the Bayesian context in which the term "significance" is typically avoided.

2.3 Example

An artificial coin flipping example shall be used as illustration. Person A offers a gamble to person B: Person A will flip a presumably fair coin 10 times. If the number of heads is between 4 and 6, then person B wins, else person A wins. However, person B manages to check the coin in advance. To do so, person B plans to flip the coin several times to estimate its probability of heads π and calculate the coins' bias $b := \pi - 0.5 \in [-0.5, 0.5]$, which represents the effect parameter. Depending on the outcome, person B might think about accusing person A of cheating. In that, the possible actions of person B are:

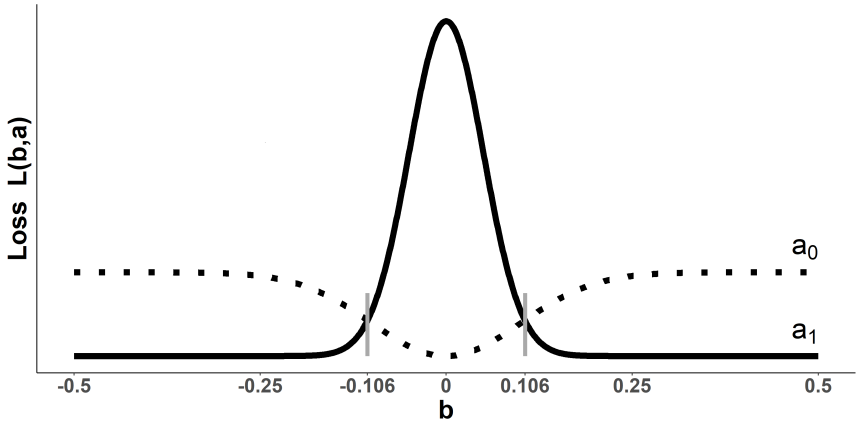

Figure 1: Example: Loss Function. The loss value (y-axis) is depicted for each possible bias b (x-axis) and each of both actions a_0 (dotted line) and a_1 (solid line). Both lines cross at bias values $b = -0.106$ and $b = 0.106$. Within this example, this loss function was arbitrarily chosen and is treated as given.

a_0: not accuse person A of cheating

a_1: accuse person A of cheating

For each potential bias b, person B assesses how bad each of both actions would be, and reasons that this badness might be represented mathematically by the loss function depicted in Figure 1 (As mentioned, specifying such a loss function unambiguously in an applied context is a difficult task. Therefore, for this example, this loss function shall be treated as given.).

With this loss function, it appears that bias values b within $B_0 = [-0.106, 0.106]$ are negligible and bias values within $B_1 = [-0.5, -0.106) \cup (0.106, 0.5]$ are practically relevant w.r.t. to the decision of person B.

2.4 Discussion

As made explicit by this definition, the practical relevance or negligibility of an effect involves certain actions and a specific loss function. For different actions or with a different loss function, different effects might be practically relevant or negligible. In that, it is recommended to explicitly state the actions and describe corresponding consequences in an applied context.

In order to keep the formal definitions simple, an effect θ for which both actions have the same loss, i.e. $L(a_0, \theta) = L(a_1, \theta)$, is arbitrarily treated as negligible. Nevertheless, it might be equally valid to treat it as practically relevant, yet this will hardly be of importance in applied investigations.

The decision theoretic concepts employed within this definition are the actions themselves and the loss function. As latter allows to determine which action should be preferred for each effect θ, these concepts are exactly those needed to answer question (c) about whether the effect θ is useful (i.e. practically relevant) or not (Kirk, 2001). Naturally, the observed data y have not been involved so far, as determining which potential effects θ are practically

relevant is possible (and actually necessary) before collecting the data.

Although only few decision theoretic concepts (actions and loss function) are necessary to define the practical relevance (or negligibility) of an effect, decision theory is an extensive framework for deciding in the face of uncertainty. In addition to well-founded work on how to include data in this decision process, it is possible to include a variety of different external information, e.g. within the loss function, within a prior distribution or by choosing a certain decision paradigm or decision principle (see e.g. Berger, 1985).

The reliance on decision theoretic concepts has already been anticipated by those scientists dealing with practical significance. For example, Pintea (2010, p. 103) emphasized the importance of the decision for which a research result is used: "An increasing number of authors underline the gap between researchers who only report the statistical significance of their results and practitioners who need relevant information for their decisions in clinical, counseling, educational, and organizational practice." Also the necessity to include the usefulness or value of each effect into a statistical analysis, which can be achieved by a loss function, was highlighted e.g. by Thompson (1993, p. 365): "If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating p[-value]s, and so p[-value]s cannot be blithely used to infer the value of research results."

3 Practical Relevance in the Context of Hypotheses

3.1 Context

An additional characteristic of the examples, that lead to the definition of a practically relevant effect, is that a conventional null hypothesis significance test leads to a questionable result. Null hypothesis significance testing (NHST) typically involves a null hypothesis that hypothesizes only a single parameter value representing a zero effect – such that this null hypothesis is frequently referred to as nil hypothesis (see e.g. Cohen, 1994) – and a general alternative hypothesis that hypothesizes all other possible parameter values. For over 80 years, this approach has been subject to the critique of being not of interest (see e.g. Berkson, 1938; Cohen, 1994; Gigerenzer, 2004). In that, the conventional NHST approach might lead to a conclusion which favors the alternative hypothesis (commonly interpreted as presence of an effect) even if the observed effect is negligible (as e.g. in the aspirin example). This is because also negligible effects are hypothesized within such a general alternative hypothesis.

Similar issues might arise in the context of Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995; Gönen et al., 2005; Rouder et al., 2009), a Bayesian alternative to frequentist hypothesis tests. Frequently, Bayes factors are also calculated with sharp null hypotheses (see e.g. Jeffreys, 1961; Rouder et al., 2009, 2018a,b; Lakens et al., 2020), such that corresponding alternative hypotheses might also contain negligible effects. However, it has to be noted that there are exceptions, which consider interval-valued null hypotheses (see e.g. Morey and Rouder, 2011; Hoijtink et al., 2019; Heck et al., 2020).

Critics claim that – using the terminology of this paper – the null hypothesis should hypothesize all negligible effects, not only the zero effect, and that the alternative hypothesis should hypothesize practically relevant effects (see e.g. Berger, 1985; Morey and Rouder, 2011; Lakens et al., 2018; Kruschke, 2018; Blume et al., 2019). As the definition of a

practically relevant or a negligible effect requires the presence of two actions and a loss function, so does meeting this claim. In that, appropriately specified hypotheses are to be defined within the context of decision theory.

3.2 Formal Definition

Continuing with the previous notation, hypotheses about the effect θ are subsets Θ_0 and Θ_1 of the parameter space Θ:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \,. \tag{3}$$

In NHST, $\Theta_0 = \{0\}$ contains only the zero effect and $\Theta_1 = \{\theta \in \Theta | \theta \neq 0\}$ contains all other effect values. However, for a given decision between the actions a_0 and a_1 and a given loss function L, this Θ_1 (representing the general alternative) typically contains also effects θ that are negligible w.r.t. the given decision problem, resulting in the critique outlined above.

Instead, Θ_0 should contain no practically relevant effects and Θ_1 should contain no negligible effects, yet this is possible in two different ways:

- All effects within the subsets Θ_0 and Θ_1 are negligible and practically relevant, respectively.
- All negligible and practically relevant effects are within the subsets Θ_0 and Θ_1, respectively.

In the first case, there might still be (negligible or practically relevant) effects left which are not contained in either hypothesis, i.e. $\Theta = \Theta_0 \cup \Theta_1$ need not hold, and in the second case, any effect is contained in one of the hypotheses, i.e. $\Theta = \Theta_0 \cup \Theta_1$ holds. Accordingly, in the former case the hypotheses might incorporate the notion of practical relevance only partially, in the latter case even completely, leading to the following definition[1]:

Definition 2 (Practical Relevance w.r.t. Hypotheses)**.** *Within this framework, two hypotheses about an effect θ (equation 3)* ***completely*** *incorporate the notion of practical relevance (or practical significance) w.r.t. two associated actions a_0, a_1 and the corresponding loss function L, if Θ_1 contains* ***all*** *practically relevant effects and Θ_0 contains* ***all*** *negligible effects, i.e.*

$$\forall \theta \in \Theta : L(\theta, a_0) \leq L(\theta, a_1) \Rightarrow \theta \in \Theta_0 \tag{4}$$

$$\forall \theta \in \Theta : L(\theta, a_1) < L(\theta, a_0) \Rightarrow \theta \in \Theta_1 \,. \tag{5}$$

[1]Within this definition, hypotheses are separated in a mathematically exact way based on the loss function: Even for very small differences between the loss values of both actions, an effect value is placed within one of the hypotheses if its loss value is smaller than the other loss value. In this regard, an idealized precise underlying loss function is assumed and dealt with in an numerical exact way, leading to a precise boundary between both hypotheses. In contrast, a more applied and less idealized view of the loss function might be to allow a rather vague boundary between the hypotheses, consisting of a range of different effect values. Consequently, all these effects, which characterize this vague boundary, cannot differentiate between both hypotheses and might be (arbitrarily) referred to negligible. With regard to its interpretation, this emphasizes a rather subjective nature of the loss function: The badness of the consequences of the different actions are *perceived* as somehow equivalent for these ranges of effects.

These hypotheses (equation 3) ***partially*** *incorporate the notion of practical relevance (or practical significance) w.r.t. these actions and this loss function, if* Θ_1 *contains* ***only*** *practically relevant effects and* Θ_0 *contains* ***only*** *negligible effects, i.e.*

$$\forall \theta \in \Theta_0 : L(\theta, a_0) \leq L(\theta, a_1) \tag{6}$$

$$\forall \theta \in \Theta_1 : L(\theta, a_1) < L(\theta, a_0)\,. \tag{7}$$

3.3 Discussion and Example

Naturally, if hypotheses incorporate the notion of practical relevance completely they also do so partially. Continuing the previous coin flipping example, the hypotheses

$$H_0 : b \in [-0.106, 0.106] \quad \text{vs.}$$

$$H_1 : b \in [-0.5, -0.106) \cup (0.106, 0.5] \tag{8}$$

incorporate the notion of practical relevance both completely and partially w.r.t. the underlying actions and loss function.

However, this implication does not hold in the reverse direction. For example, the (arbitrarily chosen) hypotheses

$$H_0 : b \in \{0\} \quad \text{vs.} \quad H_1 : b \in \{0.3\} \tag{9}$$

incorporate the notion of practical relevance only partially, but not completely, w.r.t. the underlying actions and loss function.

Incorporating the notion of practical relevance into hypotheses only partially, and not completely, is equivalent to a restriction on the parameter space. The union of both of these hypotheses constitutes a new restricted parameter space $\{0, 0.3\}$, in which they (equation (9)) incorporate the notion of practical relevance completely. Accordingly, this implies that all other parameter values $[-0.5, 0.5] \setminus \{0, 0.3\}$ are irrelevant within the conducted analysis. This is a strong claim and needs to be justified. Therefore, it is recommended to primarily employ a parameter space that is meaningful in the context of the sampling distribution and then derive hypotheses that incorporate the notion of practical relevance completely w.r.t. this parameter space and the underlying decision problem.

4 In Practice

This paper offers a formal definition of the concept of practical relevance (or practical significance) for both effects and hypotheses. It appears that a proper definition of this concept depends on an underlying decision problem. Without such a decision problem, it is neither possible to assess the practical relevance of an effect nor to specify practically relevant hypotheses, as their practical relevance naturally depends on what they are used for.

The main goal of this elaboration is to understand the notion of practical relevance on a formal level. This is necessary to include it into a statistical analysis in a mathematically correct manner. Without it, the discussion about the practical relevance of an observed effect or of some research results is of mere qualitative nature. The researcher interprets

the observed effect and the results, and integrates them into the broader research context. If this context relates to a practical problem, there will be a judgment about the practical relevance of the results. Yet, this judgment might be biased and debatable. Humans are prone to fallacies, and critical self reflection is the foundation of science. By considering the concept of practical relevance of a formal level, a mathematically and logically correct evaluation of the practical relevance of an observed effect or of some research results is possible.

Just because decision theoretic concepts were employed in the definitions provided within this paper does not mean that these concepts are dispensable in the absence of a formal representation of the notion of practical relevance. As outlined (Section 2.1), elaborations in the context of practical relevance did indeed – at least implicitly – employ decision theoretic concepts, yet in an informal way. Accordingly, the elaborations within this paper aim only to set out these implicit decision theoretic considerations.

While it is easier to see that there is an underlying decision problem in an applied scientific investigation, it might not be that apparent in the context of foundational scientific work. Typically, an attempt to describe the statistical analysis of foundational investigations in terms of statistical decision theory yields loss functions that appear to be quite artificial, employing default loss values (see e.g. Berger, 1985). Potential actions might be rather unspecific, such as e.g. "dismiss hypothesis H_0", "do not dismiss hypothesis H_0", or "follow the line of research that is in accordance with hypothesis H_0". In such a context, it might be argued that it is beneficial to refrain from making decisions (e.g. Rouder et al., 2018b, p. 110), putting an emphasis on describing the observed data, allowing others to use it in their specific context. Also, there might be scientific research situations in which a sharp null hypothesis might be reasonable (see e.g. Heck et al., 2020). If there are good reasons to do so, a sharp null hypothesis, that might not relate to a practical context, might naturally be employed. Science is very versatile and no single method or consideration does apply to every scientific context. Yet, for all those scientific investigations that are actually interested in the practical implications of their results, the definitions given within this paper might come into play.

Naturally, the question arises, whether it is necessary to have a fully specified loss function L to assess the practical relevance of an effect or of a result. The unambiguous specification of such a loss function might be seen as an unsolvable task within an applied context. Information about the consequences of respective actions is expected to be scarce, vague, ambiguous, and partial, yet it should be condensed into a single quantitative entity. To make things worse, this should be done for all possible parameter values θ (of which there are frequently infinitely many) and for all actions. Consequently, it appears that the willingness of applied scientists to employ a decision theoretic account goes down to the necessity of having a loss function.

Of course, if such a loss function L is fully available, then a decision theoretic analysis can be performed (as e.g. outlined in Berger, 1985; Robert, 2007). However, as can be seen in the definitions within this paper, not all information from a loss function is required to determine the practical relevance of an effect or to specify hypotheses such that they incorporate the notion of practical relevance completely. In that, it is possible to assess the practical relevance without fully knowing the loss function. Instead, the researcher has to merely gather all available information about the consequences of the respective actions and separate the parameter space according to the preference for each action. In detail,

the procedure is as follows:

- Think about what your research should be used for.
- Explicitly state and describe the actions a_0 and a_1 of the decision problem that represents the purpose of your research.
- After specifying the (parametric) statistical model, gather all available information about the consequences of deciding for a_0 and a_1 in dependence of the parameter value.
- Using this information, determine for each parameter value which of both actions should be preferred over the other.

This leads to two parameter sets that define hypotheses that completely incorporate the notion of practical relevance. In that, hypotheses were specified reasonably w.r.t. the practical purpose of the investigation.

In general, this is considered to be an applied, not a statistical, problem: It is the applied scientists who have the knowledge about the practical context of interest, such that it is them who can best specify the hypotheses reasonably (see e.g. Kirk, 1996, 2001; Morey and Rouder, 2011; Lakens, 2017; Lakens et al., 2018; Kruschke, 2018). Usually, the statisticians, who develop a statistical method, do not know the specific research context. Even further, any recommendation about default hypothesis specifications cannot match with all the different contexts a statistical method can be applied in. In many scientific fields, comprehensive guidelines on how to specify hypotheses reasonably are the exception rather than the rule. In that sense, the procedure above is a general guideline that might help applied scientists to specify their hypotheses reasonably, without being restricted to the characteristics of a specific field of applied science.

In the field of methodologies, there are plenty of methods that require statistical hypotheses to be specified reasonably w.r.t. the underlying practical purpose. Examples are equivalence tests (see e.g. Lakens, 2017; Lakens et al., 2018) in the frequentist setting, and Bayes factors (see e.g. Morey and Rouder, 2011; Hoijtink et al., 2019; Heck et al., 2020) or decision rules that consider the region of practical equivalence (ROPE) around a null value (see e.g. Kruschke, 2015, 2018) in the Bayesian setting. The present elaboration might be helpful for these methodologies. Yet, it should be noted that the mentioned methodologies, which employ reasonably specified hypotheses, do not yield an optimal action in the context of a practical decision problem as their result: Equivalence tests result in classic tests decisions about rejecting or not rejecting a hypothesis, Bayes factors quantify evidence, and ROPE-based decision rules accept or reject a parameter null value (or withdraw the decision). In order to find the optimal action in the context of a practical decision, further loss considerations with regard to this practical decision need to be performed.

Accordingly, another alternative is to use a hypothesis-based decision theoretic analysis. Then, however, it is necessary to provide additional quantitative specifications in the context of the loss function, than merely specifying the hypotheses reasonably. Yet, it appears that – in the context of hypothesis-based decision theory with only two hypotheses and two actions – it is only one loss value that needs to be specified, relating the consequences of the type-I-error (decide for a_1 if H_0 is true) to the consequences of the type-II-error (decide for a_0 if H_1 is true). Further, this loss value might also be allowed to be interval-valued (in this regard, cp. Walley, 1991; Augustin et al., 2014), such that a range of plausible values

might be specified, representing a rather robust loss specification. In that sense, the specification requirements of the loss function are extremely mitigated, allowing its employment in applied empirical investigations. How to derive the optimal action in a Bayesian setting with such a robust loss specification is elaborated on elsewhere (Schwaferts and Augustin, 2019, 2021, 2020).

5 Discussion

Critics might argue that the definitions of practical relevance, as provided within this paper, merely shift the difficulty of specifying hypotheses reasonably w.r.t. a practical purpose to the difficulty of specifying a loss function. Yet, it needs to be distinguished between

- *[Understanding]* formulating these definitions in an attempt to understand the notion of practical relevance on a formally exact level (which is the purpose of this paper),
- *[Development of Methodologies]* developing statistical methodologies for applied scientists (which might be motivated by the elaborations within this paper and can be located within the frameworks depicted in (Schwaferts and Augustin, 2019, 2021, 2020)), and
- *[Promotion of Methodologies]* claiming that certain statistical methodologies are appropriate in a variety of different contexts.

It is not the formulation of these definitions that shifts the difficulty to the specification of the loss function. Formulating these definitions merely generates understanding. If there is a practical purpose, then there is an underlying decision problem. A statistical analysis that wants to derive conclusions about the practical relevance of an effect or of the results needs to consider this underlying decision problem. In this context, the difficulty of specifying a loss function has – although mainly hidden – always been there, and current methodologies that try to circumvent loss considerations might be suboptimal in assessing the practical relevance of the observed effect or of the obtained results. Accordingly, it is not a shift in difficulty, but a disclosure of where the difficulty truly is. As outlined (Section 4), the direction of the development of appropriate methodologies might be indicated, but their development is another issue. Naturally, of importance for this development is how to deal with the difficulties in the specification of the loss function. Yet, ignoring loss considerations at all cannot yield results that are related to the practical purpose of the study. In that, the claim that such methodologies without loss considerations yield practically relevant results should be treated with caution. Although these methodologies might appear to be applied more easily because of their ignorance to loss considerations, this cannot be an argument to promote them for scientific investigations with a practical purpose.

References

Augustin T., Coolen F.P., de Cooman G., and Troffaes M.C.M., editors (2014). *Introduction to Imprecise Probabilities*. John Wiley & Sons, Chichester. URL `http://dx.doi.org/10.1002/9781118763117`.

Baicus C. and Caraiola S. (2009). Effect measure for quantitative endpoints: Statistical versus clinical significance, or 'how large the scale is?'. *European Journal of Internal Medicine*, 20(5):e124–e125. URL `http://dx.doi.org/10.1016/j.ejim.2008.10.002`.

Bartolucci A.A., Tendera M., and Howard G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *The American Journal of Cardiology*, 107(12):1796–1801. URL http://dx.doi.org/10.1016/j.amjcard.2011.02.325.

Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition. URL http://dx.doi.org/10.1007/978-1-4757-4286-2.

Berkson J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536. URL http://dx.doi.org/10.2307/2279690.

Blume J.D., Greevy R.A., Welty V.F., Smith J.R., and Dupont W.D. (2019). An introduction to second-generation p-values. *The American Statistician*, 73(sup1):157–167. URL http://dx.doi.org/10.1080/00031305.2018.1537893.

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

Cumming G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1):7–29. URL http://dx.doi.org/10.1177/0956797613504966.

Ellis S.M. and Steyn H.S. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): Research note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, 12(4):51–53. URL http://dx.doi.org/10.4300/JGME-D-12-00156.1.

Gigerenzer G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606. URL http://dx.doi.org/10.1016/j.socec.2004.09.033.

Gönen M., Johnson W.O., Lu Y., and Westfall P.H. (2005). The Bayesian two-sample t test. *The American Statistician*, 59:252–257. URL http://dx.doi.org/10.1198/000313005X55233.

Heck D.W., Boehm U., Böing-Messing F., Bürkner P.C., Derks K., Dienes Z., Fu Q., Gu X., Karimova D., Kiers H., et al. (2020). A review of applications of the Bayes factor in psychological research. URL http://dx.doi.org/10.31234/osf.io/cu43g.

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

Hojat M. and Xu G. (2004). A visitor's guide to effect sizes – statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education*, 9(3):241–249. URL http://dx.doi.org/10.1023/B:AHSE.0000038173.00909.f6.

Jeffreys H. (1961). *Theory of Probability*. Oxford University Press, Oxford, third edition.

Kass R.E. and Raftery A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795. URL http://dx.doi.org/10.2307/2291091.

Kirk R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5):746–759. URL http://dx.doi.org/10.1177/0013164496056005002.

Kirk R.E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61(2):213–218. URL http://dx.doi.org/10.1177/00131640121971185.

Kruschke J.K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan*. Academic Press, New York. URL http://dx.doi.org/10.1016/B978-0-12-405888-0.09999-2.

Kruschke J.K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280. URL http://dx.doi.org/10.1177/2515245918771304.

Lakens D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362. URL http://dx.doi.org/10.1177/1948550617697177.

Lakens D., McLatchie N., Isager P.M., Scheel A.M., and Dienes Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, 75(1):45–57. URL http://dx.doi.org/10.1093/geronb/gby065.

Lakens D., Scheel A.M., and Isager P.M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269. URL http://dx.doi.org/10.1177/2515245918770963.

Morey R.D. and Rouder J.N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4):406–419. URL http://dx.doi.org/10.1037/a0024377.

Peeters M.J. (2016). Practical significance: Moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning*, 8(1):83–89. URL http://dx.doi.org/10.1016/j.cptl.2015.09.001.

Pintea S. (2010). The relevance of results in clinical research: Statistical, practical, and clinical significance. *Journal of Cognitive and Behavioral Psychotherapies*, 10(1):101–114.

Robert C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, second edition. URL http://dx.doi.org/10.1007/0-387-71599-1.

Rouder J.N., Haaf J.M., and Aust F. (2018a). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, 85(1):41–56. URL http://dx.doi.org/10.1080/03637751.2017.1394581.

Rouder J.N., Haaf J.M., and Vandekerckhove J. (2018b). Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1):102–113. URL http://dx.doi.org/10.3758/s13423-017-1420-7.

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

van de Schoot R., Winter S.D., Ryan O., Zondervan-Zwijnenburg M., and Depaoli S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2):217–239. URL http://dx.doi.org/10.1037/met0000100.

Schwaferts P. and Augustin T. (2019). Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 338–345. PMLR. URL http://proceedings.mlr.press/v103/schwaferts19a.html.

Schwaferts P. and Augustin T. (2020). Bayesian decisions using regions of practical equivalence (ROPE): Foundations. Technical Report 235, Ludwig-Maximilians-University Munich, Department of Statistics. URL http://dx.doi.org/10.5282/ubm/epub.74222.

Schwaferts P. and Augustin T. (2021). Imprecise hypothesis-based Bayesian decision making with composite hypotheses. In A. Cano, J. De Bock, E. Miranda, and S. Moral, editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, page 280–288. PMLR. URL https://proceedings.mlr.press/v147/schwaferts21a.html.

Sullivan G.M. and Feinn R. (2012). Using effect size – or why the p value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282. URL http://dx.doi.org/10.4300/JGME-D-12-00156.1.

Thompson B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education*, 61(4):361–377. URL http://dx.doi.org/10.1080/00220973.1993.10806596.

Thompson B. (2002). 'Statistical','practical,' and 'clinical': How many kinds of significance do counselors need to consider? *Journal of Counseling & Development*, 80(1):64–71. URL http://dx.doi.org/10.1002/j.1556-6678.2002.tb00167.x.

Vaske J.J. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife*, 7(4):287–300. URL http://dx.doi.org/10.1080/10871200214752.

Walley P. (1991). *Statistical Reasoning With Imprecise Probabilities*. Chapman & Hall, London.

Contribution 2

Schwaferts & Augustin (2021e): Updating Consistency in Bayes Factors. (Technical Report)

Patrick Schwaferts
Thomas Augustin

Updating Consistency in Bayes Factors

Technical Report Number 236, 2021
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Updating Consistency in Bayes Factors

Patrick Schwaferts Thomas Augustin

Ludwig-Maximilians-Universität Munich
Department of Statistics
Methodological Foundations of Statistics and its Applications
Ludwigsstraße 33, 80539 Munich, Germany

Abstract

When it comes to extracting information from data by means of Bayes rule, it should not matter if all available data are considered at once or if Bayesian updating is performed subsequently with partitions of the data. This property is called updating consistency. However, in the context of Bayes factors, a prominent Bayesian tool that is used for comparing hypotheses, some researchers illustrated that updating consistency might not be given. Therefore, this technical report addresses the updating consistency of Bayes factors and shows its existence. In that, it serves as mathematical basis for the evaluation of the origin of putative updating inconsistencies. In addition, results about updating mixture priors are brought into the terminology commonly employed in the context of Bayes factors, as these were used in the elaboration about updating consistency. The depicted results imply that a necessary condition for updating consistency is to consider and report not only the Bayes factor value alone but also the posterior distributions as outcome of the analysis.

Keywords: Bayesian Statistics, Bayes Factor, Sequential Updating, Updating Consistency, Mixture Prior, Spike-and-Slab Prior

1 Introduction

Within the context of Bayesian statistics, the knowledge about a phenomenon of interest that is available prior to an investigation is typically formalized as a (subjective) prior probability distribution. Once a respective investigation has been performed, the obtained data are used to update this initial prior distribution via Bayes rule, yielding a posterior distribution. This posterior is said to reflect all relevant knowledge which is available after the investigation (see Figure 1, top-left). In that, the posterior distribution might be treated as prior distribution for a subsequent statistical analysis of a newly obtained data set (based on the same investigational setup). Naturally, the final posterior distribution after sequentially updating twice (see Figure 1, top-right) should be identical to the posterior distribution that is obtained by merging both data sets first and then updating the initial prior distribution at once (see Figure 1, bottom). If so, the Bayesian updating procedure is called "consistent" [cp. Rüger, 1998, p. 190].

However, the most prominent Bayesian method for hypothesis comparisons employed in psychological research - the so called Bayes factor [see e.g. Kass and Raftery, 1995, Gönen

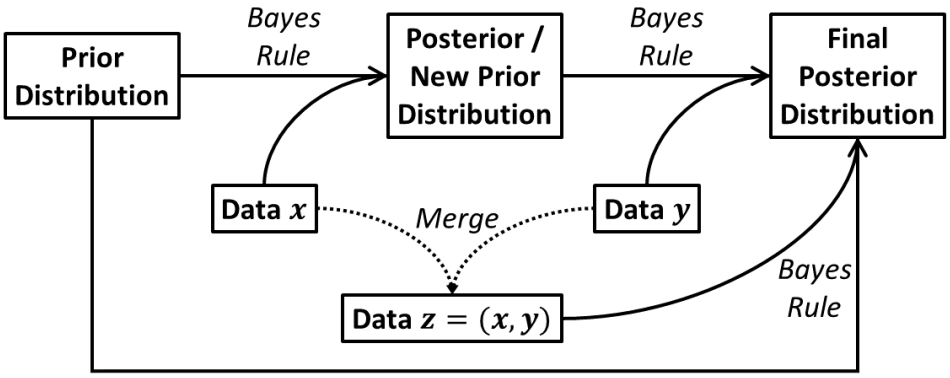

Figure 1: Consistent Bayesian Updating.

et al., 2005, Rouder et al., 2009] - might be characterized by an inconsistent Bayesian updating, as already indicated by Rouder and Morey [2011]. In contrast to specifying only one prior distribution that reflects the available knowledge prior to the investigation, an analysis with Bayes factors allows the specification of a prior distribution for each employed hypothesis, which is said to reflect its content [see e.g. Vanpaemel, 2010, Vanpaemel and Lee, 2012, Morey et al., 2016, Rouder et al., 2018a]. Although more than just a single prior distribution is employed, together with a distribution on the hypotheses themselves it is possible to merge all these hypothesis-based priors to an overall mixture distribution [see e.g. Rouder et al., 2018b]. By considering this mixture prior distribution, its updating might be assessed w.r.t. consistency, such that the origin of putative updating inconsistencies in the context of Bayes factors might be evaluated.

Accordingly, Bayes factors shall be outlined in Section 2 before considering the updating of the corresponding mixture prior in Section 3. These considerations are used to show that updating with Bayes factors is consistent (Section 4), but also that inconsistent updating might occur easily (Section 4.3). Implications about the minimal requirement of what is considered as outcome of an analysis with Bayes factors of a single data set are depicted in Section 5.

This technical report intends to depict the mathematical background of updating consistency in the context of Bayes factors in greater detail. Special emphasize will be given to explain mathematical transformations step by step with numerous references to previous definitions and equations. In addition, as all data, parameter, and hypotheses are random quantities, which are related to each other, Bayes rule is always applied meticulously, allowing clarity about which quantities are conditioned on.

2 Bayes Factors

Assume the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$ are modeled as realizations of independent and identically distributed (iid) random quantities $X_i \overset{iid}{\sim} P_{X_i|\theta}$ with parametric density $f(x_i|\theta)$

for all $i = 1, \ldots, n$ and a parameter value $\theta \in \Theta$, such that $X \sim P_{X|\theta}$ with density

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)\,. \tag{1}$$

Statistical hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \tag{2}$$

contrast two subsets Θ_0 and Θ_1 of the parameter space Θ. Frequently, the null hypothesis is sharp and consists of only a single parameter value θ_0, i.e. $\Theta_0 = \{\theta_0\}$.

In the Bayesian setting, there is a prior distribution (P_θ°, see below) on the parameter θ. In the context of Bayes factors, however, a prior distribution is typically provided separately for each hypothesis: The prior distribution $P_\theta^{(1)}$ with density

$$\pi_1(\theta) := \pi(\theta|H_1) \tag{3}$$

is restricted to the (parameter subset specified within the) alternative hypothesis H_1, and the prior distribution $P_\theta^{(0)}$ with density

$$\pi_0(\theta) := \pi(\theta|H_0) \tag{4}$$

is restricted to the (parameter subset specified within the) null hypothesis H_0. If the null hypothesis is sharp, the corresponding prior distribution $P_\theta^{(0)}$ is degenerate with all probability mass on θ_0 .

In addition to $P_\theta^{(1)}$ and $P_\theta^{(0)}$, a prior distribution on the hypotheses themselves needs to be specified by

$$\rho := p(H_0) \quad \text{and} \quad p(H_1) = 1 - p(H_0) = 1 - \rho\,, \tag{5}$$

yielding the so called prior odds $p(H_1)/p(H_0)$.

The density of $P_{X|\theta}$ is assumed to be related to the hypotheses H_1 and H_0 only via the parameter value θ, i.e.

$$f(\boldsymbol{x}|H_1, \theta) = f(\boldsymbol{x}|H_0, \theta) = f(\boldsymbol{x}|\theta)\,. \tag{6}$$

The marginal density of the data $\boldsymbol{x}$ might be calculated w.r.t. each hypothesis

$$f(\boldsymbol{x}|H_1) \overset{marg.}{=} \int f(\boldsymbol{x}|H_1, \theta) \cdot \pi(\theta|H_1)\, d\theta \overset{\substack{eq.\\(6)\\(3)}}{=} \int f(\boldsymbol{x}|\theta) \cdot \pi_1(\theta)\, d\theta \tag{7}$$

$$f(\boldsymbol{x}|H_0) \overset{marg.}{=} \int f(\boldsymbol{x}|H_0, \theta) \cdot \pi(\theta|H_0)\, d\theta \overset{\substack{eq.\\(6)\\(4)}}{=} \int f(\boldsymbol{x}|\theta) \cdot \pi_0(\theta)\, d\theta \tag{8}$$

and the Bayes factor based on data $\boldsymbol{x}$ w.r.t. the hypotheses H_0 and H_1 is defined as the ratio of these marginal densities

$$BF_{10}^{\boldsymbol{x}} := \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \overset{\substack{eq.\\(7)\\(8)}}{=} \frac{\int f(\boldsymbol{x}|\theta) \cdot \pi_1(\theta)\, d\theta}{\int f(\boldsymbol{x}|\theta) \cdot \pi_0(\theta)\, d\theta}\,, \tag{9}$$

with a said interpretation of the data $\boldsymbol{x}$ being $BF_{10}^{\boldsymbol{x}}$ times as much evidence for H_1 than for H_0 [see e.g. Morey et al., 2016]. In that regard, consider the discussion in Section 5. Analogously, its inverse

$$BF_{01}^{\boldsymbol{x}} := \frac{1}{BF_{10}^{\boldsymbol{x}}} \tag{10}$$

should quantify the evidence within the data favoring H_0 over H_1.

The prior odds can be updated by the Bayes factor to the posterior odds

$$\frac{p(H_1|\boldsymbol{x})}{p(H_0|\boldsymbol{x})} \overset{\substack{Bayes\\ rule}}{=} \frac{\frac{f(\boldsymbol{x}|H_1)\cdot p(H_1)}{f(\boldsymbol{x})}}{\frac{f(\boldsymbol{x}|H_0)\cdot p(H_0)}{f(\boldsymbol{x})}} \overset{\substack{eq.\\(9)}}{=} BF_{10}^{\boldsymbol{x}} \cdot \frac{p(H_1)}{p(H_0)}\,. \tag{11}$$

In that, the posterior probability of H_0 denoted by

$$\rho_{|\boldsymbol{x}} := p(H_0|\boldsymbol{x}) \tag{12}$$

can be calculated as

$$\begin{aligned}
&\frac{p(H_1|\boldsymbol{x})}{p(H_0|\boldsymbol{x})} = BF_{10}^{\boldsymbol{x}} \cdot \frac{p(H_1)}{p(H_0)} \quad \overset{\substack{eq.\\(5)\\(12)}}{\Leftrightarrow} \quad \frac{1-\rho_{|\boldsymbol{x}}}{\rho_{|\boldsymbol{x}}} = BF_{10}^{\boldsymbol{x}} \cdot \frac{1-\rho}{\rho} \\
\Leftrightarrow \quad & 1-\rho_{|\boldsymbol{x}} = BF_{10}^{\boldsymbol{x}} \cdot \frac{1-\rho}{\rho} \cdot \rho_{|\boldsymbol{x}} \quad \Leftrightarrow \quad 1 = BF_{10}^{\boldsymbol{x}} \cdot \frac{1-\rho}{\rho} \cdot \rho_{|\boldsymbol{x}} + \rho_{|\boldsymbol{x}} \\
\Leftrightarrow \quad & 1 = \rho_{|\boldsymbol{x}} \left[BF_{10}^{\boldsymbol{x}} \frac{1-\rho}{\rho} + 1 \right] \quad \Leftrightarrow \quad \rho_{|\boldsymbol{x}} = \frac{1}{BF_{10}^{\boldsymbol{x}} \frac{1-\rho}{\rho} + 1} \\
\Leftrightarrow \quad & \rho_{|\boldsymbol{x}} = \frac{\rho}{BF_{10}^{\boldsymbol{x}}(1-\rho) + \rho}\,.
\end{aligned} \tag{13}$$

3 Updating of Mixture Priors

Instead of treating the priors under both hypotheses separately, they might be merged to a single mixture prior distribution

$$P_\theta^\circ := \rho \cdot P_\theta^{(0)} + (1-\rho) \cdot P_\theta^{(1)}\,, \tag{14}$$

which has the density

$$\pi^\circ(\theta) = \rho \cdot \pi_0(\theta) + (1-\rho) \cdot \pi_1(\theta)\,. \tag{15}$$

With $P_\theta^{(0)}$ being degenerate this mixture prior is also referred to as spike-and-slab prior [see e.g. Rouder et al., 2018b], consisting of a spike-part $P_\theta^{(0)}$ and a slab-part $P_\theta^{(1)}$.

Theorem 1 (Updating of Mixture Priors)**.** *Updating the prior mixture distribution P_θ° using data $\boldsymbol{x}$ leads to the posterior distribution $P_{\theta|\boldsymbol{x}}^\circ$ with density*

$$\pi^\circ(\theta|\boldsymbol{x}) = \rho_{|\boldsymbol{x}} \cdot \pi_0(\theta|\boldsymbol{x}) + (1-\rho_{|\boldsymbol{x}}) \cdot \pi_1(\theta|\boldsymbol{x})\,, \tag{16}$$

where $\pi_0(\theta|\boldsymbol{x})$ as well as $\pi_1(\theta|\boldsymbol{x})$ are posterior densities of θ, which arise from updating the prior densities $\pi_0(\theta)$ as well as $\pi_1(\theta)$ separately, i.e.

$$\pi_0(\theta|\boldsymbol{x}) := \pi(\theta|H_0, \boldsymbol{x}) \overset{\substack{Bayes\\ rule}}{=} \frac{f(\boldsymbol{x}|H_0,\theta)\cdot\pi(\theta|H_0)}{f(\boldsymbol{x}|H_0)} \overset{\substack{eq.\\(6)\\(4)\\(8)}}{=} \frac{f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}\,, \tag{17}$$

$$\pi_1(\theta|\boldsymbol{x}) := \pi(\theta|H_1,\boldsymbol{x}) \overset{Bayes\ rule}{=} \frac{f(\boldsymbol{x}|H_1,\theta)\cdot\pi(\theta|H_1)}{f(\boldsymbol{x}|H_1)} \overset{\substack{eq.\\(6)\\(3)\\(7)}}{=} \frac{f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}\,. \tag{18}$$

Proof. Before calculating the density

$$\pi^\circ(\theta|\boldsymbol{x}) \overset{Bayes\ rule}{=} \frac{f(\boldsymbol{x}|\theta)\cdot\pi^\circ(\theta)}{f(\boldsymbol{x})} \tag{19}$$

of the posterior distribution $P^\circ_{\theta|\boldsymbol{x}}$ with

$$f(\boldsymbol{x}) \overset{marg.}{=} \int f(\boldsymbol{x}|\theta)\cdot\pi^\circ(\theta)\,d\theta\,, \tag{20}$$

consider the following first:

$$\begin{aligned} f(\boldsymbol{x}) &\overset{eq.\ (20)}{=} \int f(\boldsymbol{x}|\theta)\cdot\pi^\circ(\theta)\,d\theta \\ &\overset{eq.\ (15)}{=} \int f(\boldsymbol{x}|\theta)\,[\rho\cdot\pi_0(\theta)+(1-\rho)\cdot\pi_1(\theta)]\,d\theta \\ &= \int [\rho\cdot f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)]+[(1-\rho)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)]\,d\theta \\ &= \rho\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta+(1-\rho)\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta \end{aligned} \tag{21}$$

This can be transformed in two different ways:

$$\begin{aligned} & f(\boldsymbol{x}) \overset{eq.\ (21)}{=} \rho\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta+(1-\rho)\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta \\ \Leftrightarrow\quad & f(\boldsymbol{x})-\rho\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta=(1-\rho)\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta \\ \Leftrightarrow\quad & \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}-\rho=(1-\rho)\,\frac{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta} \\ \overset{eq.\ (9)}{\Leftrightarrow}\quad & \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}-\rho=(1-\rho)BF^{\boldsymbol{x}}_{10} \\ \Leftrightarrow\quad & f(\boldsymbol{x})=[(1-\rho)BF^{\boldsymbol{x}}_{10}+\rho]\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta \end{aligned} \tag{22}$$

or

$$\begin{aligned} & f(\boldsymbol{x}) \overset{eq.\ (21)}{=} \rho\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta+(1-\rho)\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta \\ \Leftrightarrow\quad & f(\boldsymbol{x})-(1-\rho)\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta=\rho\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \quad \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta} - (1-\rho) = \rho\,\frac{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta} \\
&\overset{\substack{eq.\\(9)\\(10)}}{\Leftrightarrow} \quad \frac{f(\boldsymbol{x})}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta} - (1-\rho) = \rho BF_{01}^{\boldsymbol{x}} \\
&\Leftrightarrow \quad f(\boldsymbol{x}) = [\rho BF_{01}^{\boldsymbol{x}} + (1-\rho)] \int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta\,. \qquad (23)
\end{aligned}$$

In addition, consider

$$\begin{aligned}
1-\rho_{|\boldsymbol{x}} &\overset{(13)}{=} 1 - \frac{\rho}{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho} = \frac{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho-\rho}{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho} = \frac{(1-\rho)BF_{10}^{\boldsymbol{x}}}{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho} \\
&= \frac{BF_{10}^{\boldsymbol{x}}(1-\rho)}{BF_{10}^{\boldsymbol{x}}\left[(1-\rho)+\frac{1}{BF_{10}^{\boldsymbol{x}}}\rho\right]} = \frac{(1-\rho)}{(1-\rho)+\frac{1}{BF_{10}^{\boldsymbol{x}}}\rho} \overset{(10)}{=} \frac{(1-\rho)}{(1-\rho)+BF_{01}^{\boldsymbol{x}}\rho} \\
&= \frac{(1-\rho)}{\rho BF_{01}^{\boldsymbol{x}}+(1-\rho)}\,. \qquad (24)
\end{aligned}$$

Now, the posterior density $\pi^\circ(\theta|\boldsymbol{x})$ can be calculated as

$$\begin{aligned}
\pi^\circ(\theta|\boldsymbol{x}) &\overset{\substack{Bayes\\rule}}{=} \frac{f(\boldsymbol{x}|\theta)\cdot\pi^\circ(\theta)}{f(\boldsymbol{x})} \\
&\overset{(15)}{=} \frac{f(\boldsymbol{x}|\theta)\,[\rho\cdot\pi_0(\theta)+(1-\rho)\cdot\pi_1(\theta)]}{f(\boldsymbol{x})} \\
&= \rho\frac{f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)}{f(\boldsymbol{x})} + (1-\rho)\frac{f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)}{f(\boldsymbol{x})} \\
&\overset{\substack{eq.\\(22)\\(23)}}{=} \frac{\rho}{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho}\cdot\frac{f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta} \\
&\qquad + \frac{1-\rho}{\rho BF_{01}^{\boldsymbol{x}}+(1-\rho)}\cdot\frac{f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta} \\
&\overset{\substack{eq.\\(17)\\(18)}}{=} \frac{\rho}{(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho}\,\pi_0(\theta|\boldsymbol{x}) + \frac{1-\rho}{\rho BF_{01}^{\boldsymbol{x}}+(1-\rho)}\,\pi_1(\theta|\boldsymbol{x}) \\
&\overset{\substack{eq.\\(13)\\(24)}}{=} \rho_{|\boldsymbol{x}}\cdot\pi_0(\theta|\boldsymbol{x}) + (1-\rho_{|\boldsymbol{x}})\cdot\pi_1(\theta|\boldsymbol{x})\,,
\end{aligned}$$

□

Certainly, this is not a new result as e.g. Mitchell and Beauchamp [1988] already employed spike-and-slab priors (which are a special case of mixture priors) and e.g. Rouder et al.

[2018b] depicted the priors in the context of Bayes factors by an overall spike-and-slab prior. However, these considerations explicitly utilize a notation typically employed in analyses with Bayes factors and are needed for further elaboration on the updating consistency of Bayes factors.

4 Updating Consistency

4.1 Framework

In order to assess Bayes factors w.r.t. updating consistency two different data sets are necessary. Accordingly, in addition to $\boldsymbol{x}$ and X as in Section 2, consider a second data set $\boldsymbol{y} = (y_1, \ldots, y_m)$ being independent of the previous one and modeled analogously, i.e. $Y_j \overset{iid}{\sim} P_{Y_j|\theta}$ with the same parametric density $f(y_j|\theta)$ for all $j = 1, \ldots, m$. Therefore, $Y \sim P_{Y|\theta}$ with density

$$f(\boldsymbol{y}|\theta) = \prod_{j=1}^{m} f(y_j|\theta)\,. \tag{25}$$

Analogue to equation (6), the density of $P_{Y|\theta}$ is also assumed to be related to the hypotheses H_1 and H_0 only via the parameter value θ, i.e.

$$f(\boldsymbol{y}|H_1, \theta) = f(\boldsymbol{y}|H_0, \theta) = f(\boldsymbol{y}|\theta)\,. \tag{26}$$

Define $Z := (X, Y)$ and $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y})$. As

$$f(\boldsymbol{z}|\theta) \overset{\substack{X,Y \\ ind.}}{=} f(\boldsymbol{y}|\theta) \cdot f(\boldsymbol{x}|\theta)\,, \tag{27}$$

the density of $\boldsymbol{z}$ is related to the hypotheses only via the parameter value θ as well:

$$f(\boldsymbol{z}|H_1, \theta) = f(\boldsymbol{z}|H_0, \theta) = f(\boldsymbol{z}|\theta)\,. \tag{28}$$

Analogue to the marginal densities of $\boldsymbol{x}$ (equations (7) and (8)), those of $\boldsymbol{y}$ and $\boldsymbol{z}$ are calculated as

$$f(\boldsymbol{y}|H_1) \overset{marg.}{=} \int f(\boldsymbol{y}|H_1, \theta) \cdot \pi(\theta|H_1)\, d\theta \overset{\substack{eq. \\ (26) \\ (3)}}{=} \int f(\boldsymbol{y}|\theta) \cdot \pi_1(\theta)\, d\theta \tag{29}$$

$$f(\boldsymbol{y}|H_0) \overset{marg.}{=} \int f(\boldsymbol{y}|H_0, \theta) \cdot \pi(\theta|H_0)\, d\theta \overset{\substack{eq. \\ (26) \\ (4)}}{=} \int f(\boldsymbol{y}|\theta) \cdot \pi_0(\theta)\, d\theta \tag{30}$$

$$f(\boldsymbol{z}|H_1) \overset{marg.}{=} \int f(\boldsymbol{z}|H_1, \theta) \cdot \pi(\theta|H_1)\, d\theta \overset{\substack{eq. \\ (28) \\ (3)}}{=} \int f(\boldsymbol{z}|\theta) \cdot \pi_1(\theta)\, d\theta \tag{31}$$

$$f(\boldsymbol{z}|H_0) \overset{marg.}{=} \int f(\boldsymbol{z}|H_0, \theta) \cdot \pi(\theta|H_0)\, d\theta \overset{\substack{eq. \\ (28) \\ (4)}}{=} \int f(\boldsymbol{z}|\theta) \cdot \pi_0(\theta)\, d\theta\,. \tag{32}$$

and the marginal densities of $\boldsymbol{y}$ w.r.t. to the posterior distributions of θ given the first data set $\boldsymbol{x}$ are

$$\begin{aligned} f(\boldsymbol{y}|H_1,\boldsymbol{x}) &\overset{marg.}{=} \int f(\boldsymbol{y}|H_1,\theta,\boldsymbol{x})\cdot\pi(\theta|H_1,\boldsymbol{x})\,d\theta \\ &\overset{\substack{X,Y\\ ind.}}{=} \int f(\boldsymbol{y}|H_1,\theta)\cdot\pi(\theta|H_1,\boldsymbol{x})\,d\theta \\ &\overset{\substack{eq.\\(26)\\(18)}}{=} \int f(\boldsymbol{y}|\theta)\cdot\pi_1(\theta|\boldsymbol{x})\,d\theta \end{aligned} \tag{33}$$

$$\begin{aligned} f(\boldsymbol{y}|H_0,\boldsymbol{x}) &\overset{marg.}{=} \int f(\boldsymbol{y}|H_0,\theta,\boldsymbol{x})\cdot\pi(\theta|H_0,\boldsymbol{x})\,d\theta \\ &\overset{\substack{X,Y\\ ind.}}{=} \int f(\boldsymbol{y}|H_0,\theta)\cdot\pi(\theta|H_0,\boldsymbol{x})\,d\theta \\ &\overset{\substack{eq.\\(26)\\(17)}}{=} \int f(\boldsymbol{y}|\theta)\cdot\pi_0(\theta|\boldsymbol{x})\,d\theta\,. \end{aligned} \tag{34}$$

The corresponding Bayes factor values are

$$BF_{10}^{\boldsymbol{y}} := \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)} \overset{\substack{eq.\\(29)\\(30)}}{=} \frac{\int f(\boldsymbol{y}|\theta)\cdot\pi_1(\theta)\,d\theta}{\int f(\boldsymbol{y}|\theta)\cdot\pi_0(\theta)\,d\theta} \tag{35}$$

$$BF_{10}^{\boldsymbol{z}} := \frac{f(\boldsymbol{z}|H_1)}{f(\boldsymbol{z}|H_0)} \overset{\substack{eq.\\(31)\\(32)}}{=} \frac{\int f(\boldsymbol{z}|\theta)\cdot\pi_1(\theta)\,d\theta}{\int f(\boldsymbol{z}|\theta)\cdot\pi_0(\theta)\,d\theta} \tag{36}$$

$$BF_{10}^{\boldsymbol{y}|\boldsymbol{x}} := \frac{f(\boldsymbol{y}|H_1,\boldsymbol{x})}{f(\boldsymbol{y}|H_0,\boldsymbol{x})} \overset{\substack{eq.\\(33)\\(34)}}{=} \frac{\int f(\boldsymbol{y}|\theta)\cdot\pi_1(\theta|\boldsymbol{x})\,d\theta}{\int f(\boldsymbol{y}|\theta)\cdot\pi_0(\theta|\boldsymbol{x})\,d\theta}\,. \tag{37}$$

4.2 Consistent Updating

Theorem 2 (Subsequent Updating with Bayes Factors)**.** *Based on the framework above, updating the prior odds $p(H_1)/p(H_0)$ using both $\boldsymbol{x}$ and $\boldsymbol{y}$ subsequently yields the posterior odds*

$$\frac{p(H_1|\boldsymbol{y},\boldsymbol{x})}{p(H_0|\boldsymbol{y},\boldsymbol{x})} = BF_{10}^{\boldsymbol{y}|\boldsymbol{x}}\cdot BF_{10}^{\boldsymbol{x}}\cdot\frac{p(H_1)}{p(H_0)}\,. \tag{38}$$

Proof.

$$\begin{aligned} \frac{p(H_1|\boldsymbol{y},\boldsymbol{x})}{p(H_0|\boldsymbol{y},\boldsymbol{x})} &\overset{\substack{Bayes\\rule}}{=} \frac{f(\boldsymbol{y}|H_1,\boldsymbol{x})}{f(\boldsymbol{y}|H_0,\boldsymbol{x})}\frac{p(H_1|\boldsymbol{x})}{p(H_0|\boldsymbol{x})} \\ &\overset{\substack{Bayes\\rule}}{=} \frac{f(\boldsymbol{y}|H_1,\boldsymbol{x})}{f(\boldsymbol{y}|H_0,\boldsymbol{x})}\frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)}\frac{p(H_1)}{p(H_0)} \end{aligned}$$

$$\overset{\substack{eq.\\(37)\\(9)}}{=} BF_{10}^{\boldsymbol{y}|\boldsymbol{x}} \cdot BF_{10}^{\boldsymbol{x}} \cdot \frac{p(H_1)}{p(H_0)}\,.$$

□

Theorem 3 (Consistent Updating with Bayes Factors)**.** *Based on the framework above, updating the prior odds $p(H_1)/p(H_0)$ with the corresponding Bayes factor values is consistent, i.e.*

$$\frac{p(H_1|\boldsymbol{z})}{p(H_0|\boldsymbol{z})} = \frac{p(H_1|\boldsymbol{y},\boldsymbol{x})}{p(H_0|\boldsymbol{y},\boldsymbol{x})}\,. \tag{39}$$

Proof. At first, consider

$$BF_{10}^{\boldsymbol{x}} = BF_{10}^{\boldsymbol{x}} \frac{\rho + (1-\rho)BF_{10}^{\boldsymbol{x}}}{\rho + (1-\rho)BF_{10}^{\boldsymbol{x}}} = \frac{BF_{10}^{\boldsymbol{x}}\left[\rho + (1-\rho)BF_{10}^{\boldsymbol{x}}\right]}{BF_{10}^{\boldsymbol{x}}\left[\frac{\rho}{BF_{10}^{\boldsymbol{x}}} + (1-\rho)\right]} = \frac{\rho + (1-\rho)BF_{10}^{\boldsymbol{x}}}{\frac{\rho}{BF_{10}^{\boldsymbol{x}}} + (1-\rho)}$$

$$\overset{\substack{eq.\\(10)}}{=} \frac{(1-\rho)BF_{10}^{\boldsymbol{x}} + \rho}{\rho\, BF_{01}^{\boldsymbol{x}} + (1-\rho)}\,. \tag{40}$$

Now, the Bayes factor value $BF_{10}^{\boldsymbol{z}}$ might be decomposed:

$$\begin{aligned}
BF_{10}^{\boldsymbol{z}} &\overset{\substack{eq.\\(36)}}{=} \frac{\int f(\boldsymbol{z}|\theta)\cdot\pi_1(\theta)\,d\theta}{\int f(\boldsymbol{z}|\theta)\cdot\pi_0(\theta)\,d\theta} \\
&\overset{\substack{X,Y\\ind.}}{=} \frac{\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}{\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta} \\
&= \frac{\frac{1}{f(\boldsymbol{x})}\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}{\frac{1}{f(\boldsymbol{x})}\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta} \\
&\overset{\substack{eq.\\(22)\\(23)}}{=} \frac{\frac{1}{\left[\rho BF_{01}^{\boldsymbol{x}}+(1-\rho)\right]\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}{\frac{1}{\left[(1-\rho)BF_{10}^{\boldsymbol{x}}+\rho\right]\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}\int f(\boldsymbol{y}|\theta)\cdot f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta} \\
&= \frac{(1-\rho)BF_{10}^{\boldsymbol{x}} + \rho}{\rho BF_{01}^{\boldsymbol{x}} + (1-\rho)} \cdot \frac{\int f(\boldsymbol{y}|\theta)\cdot\frac{f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_1(\theta)\,d\theta}\,d\theta}{\int f(\boldsymbol{y}|\theta)\cdot\frac{f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)}{\int f(\boldsymbol{x}|\theta)\cdot\pi_0(\theta)\,d\theta}\,d\theta} \\
&\overset{\substack{eq.\\(17)\\(18)}}{=} \frac{(1-\rho)BF_{10}^{\boldsymbol{x}} + \rho}{\rho BF_{01}^{\boldsymbol{x}} + (1-\rho)} \cdot \frac{\int f(\boldsymbol{y}|\theta)\cdot\pi_1(\theta|\boldsymbol{x})\,d\theta}{\int f(\boldsymbol{y}|\theta)\cdot\pi_0(\theta|\boldsymbol{x})\,d\theta} \\
&\overset{\substack{eq.\\(37)}}{=} \frac{(1-\rho)BF_{10}^{\boldsymbol{x}} + \rho}{\rho BF_{01}^{\boldsymbol{x}} + (1-\rho)} \cdot BF_{10}^{\boldsymbol{y}|\boldsymbol{x}} \\
&\overset{\substack{eq.\\(40)}}{=} BF_{10}^{\boldsymbol{x}} \cdot BF_{10}^{\boldsymbol{y}|\boldsymbol{x}}\,.
\end{aligned} \tag{41}$$

Therefore:

$$\frac{p(H_1|\boldsymbol{z})}{p(H_0|\boldsymbol{z})} \overset{\substack{Bayes\\ rule}}{=} \frac{f(\boldsymbol{z}|H_1)}{f(\boldsymbol{z}|H_0)} \cdot \frac{p(H_1)}{p(H_0)} \overset{\substack{eq.\\ (36)}}{=} BF_{10}^{\boldsymbol{z}} \cdot \frac{p(H_1)}{p(H_0)}$$

$$\overset{\substack{eq.\\ (41)}}{=} BF_{10}^{\boldsymbol{y}|\boldsymbol{x}} \cdot BF_{10}^{\boldsymbol{x}} \cdot \frac{p(H_1)}{p(H_0)} \overset{\substack{eq.\\ (38)}}{=} \frac{p(H_1|\boldsymbol{y},\boldsymbol{x})}{p(H_0|\boldsymbol{y},\boldsymbol{x})}\,.$$

□

4.3 Inconsistent Updating

Remark that in order to update consistently with Bayes factors, the Bayes factor value $BF_{10}^{\boldsymbol{y}|\boldsymbol{x}}$ of the second data set $\boldsymbol{y}$ need to be based on the posterior distributions $\pi_1(\theta|\boldsymbol{x})$ and $\pi_0(\theta|\boldsymbol{x})$ that incorporate the information of the previous data set $\boldsymbol{x}$.

However, using $BF_{10}^{\boldsymbol{y}}$ instead of $BF_{10}^{\boldsymbol{y}|\boldsymbol{x}}$ is erroneous and yields odds

$$\frac{p(H_1|\boldsymbol{y},\boldsymbol{x}^{(!)})}{p(H_0|\boldsymbol{y},\boldsymbol{x}^{(!)})} := BF_{10}^{\boldsymbol{y}} \cdot BF_{10}^{\boldsymbol{x}} \cdot \frac{p(H_1)}{p(H_0)}\,, \tag{42}$$

which are in general different to the posterior odds obtained by updating consistently, i.e.

$$\frac{p(H_1|\boldsymbol{y},\boldsymbol{x}^{(!)})}{p(H_0|\boldsymbol{y},\boldsymbol{x}^{(!)})} \overset{\substack{in\\ gen.}}{\neq} \frac{p(H_1|\boldsymbol{y},\boldsymbol{x})}{p(H_0|\boldsymbol{y},\boldsymbol{x})}\,. \tag{43}$$

This is due to ignoring the information about θ within the first data set $\boldsymbol{x}$ while calculating the Bayes factor value based on the second data set $\boldsymbol{y}$, and the superscript (!) indicates this loss of information.

Bayes factor updating inconsistencies might occur e.g. in the following scenario: Two different research teams are interested in the same research question and utilize the same hypotheses and employ the same prior distributions on the parameter of interest. Both teams conduct a scientific investigation with identical design and calculate a Bayes factor value independently of each other. As each value is said to describe the change in belief within the hypotheses, it is tempting (e.g. in a meta-analysis of both investigations) to utilize both Bayes factor values to calculate the final belief (posterior odds) within the initially stated hypotheses. This, however, is exactly the error displayed in equation (42).

5 Outcome of Analyses with Bayes Factors

In order to avoid updating inconsistencies, both the Bayes factor value $BF_{10}^{\boldsymbol{x}}$ and the posterior distributions $\pi_1(\theta|\boldsymbol{x})$ as well as $\pi_0(\theta|\boldsymbol{x})$ are required to perform the analysis (with Bayes factors) of the second data set $\boldsymbol{y}$ once the first data set $\boldsymbol{x}$ is available.

Accordingly, considering solely the Bayes factor value $BF_{10}^{\boldsymbol{x}}$ as the outcome of the first analysis (of data $\boldsymbol{x}$) is not sufficient. Also the updated posterior distributions $\pi_1(\theta|\boldsymbol{x})$ and $\pi_0(\theta|\boldsymbol{x})$ need to be considered and reported. This appears to be obvious in the face of the posterior mixture distribution described in theorem 1, which cannot be described by the Bayes factor value $BF_{10}^{\boldsymbol{x}}$ alone. This is summarized in the following theorem.

Theorem 4 (Outcome of Analyses with Bayes Factors)**.** *A necessary condition for updating consistency in Bayes factors is to consider and report both the Bayes factor value* $BF_{10}^{\boldsymbol{x}}$ *and the posterior distributions* $\pi_1(\theta|\boldsymbol{x})$ *as well as* $\pi_0(\theta|\boldsymbol{x})$ *as outcomes of the analysis (of the data set* $\boldsymbol{x}$*).*

These considerations about updating inconsistency in Bayes factors might also be relevant e.g. in the following case: It is argued that, in the context of Bayes factors, the shape of the prior distributions ($\pi_1(\theta)$ and $\pi_0(\theta)$) reflects the content of the hypotheses [see e.g. Vanpaemel, 2010, Vanpaemel and Lee, 2012, Morey et al., 2016, Rouder et al., 2018a], but by incorporating the information of the data $\boldsymbol{x}$ into these distributions by means of Bayes rule, these distributions change to $\pi_1(\theta|\boldsymbol{x})$ and $\pi_0(\theta|\boldsymbol{x})$, which might then reflect different contents. Although erroneous, it is tempting to treat the Bayes factor value $BF_{10}^{\boldsymbol{x}}$ as quantification of the evidence within the data $\boldsymbol{x}$ w.r.t. to the hypotheses that are described by the initial prior distributions $\pi_1(\theta)$ and $\pi_0(\theta)$, as these hypotheses were formulated to answer the research question of interest. By doing so, the change within the distributions of θ is discarded and inconsistent updating might occur.

6 Summary

With theorem 1 results about updating mixture distributions are brought into the notation typically involved in the context of Bayes factors. Theorem 2 describes the final posterior odds after considering two separate data sets subsequently and theorem 3 argues that this updating procedure is consistent. As elaborated in Section 4.3, updating inconsistencies occur by discarding information and an exemplary situation was provided, in which this might happen unintentionally. Theorem 4 provides a minimum requirement on what to consider and report as outcome of a statistical analysis with Bayes factors, and a context in which this might oppose other recommendations about Bayes factors was illustrated in Section 5. However, a thorough discussion of the occurrence and consequences of updating inconsistencies in applied Bayes factors was not intended within this technical report and is still pending. Yet, this report enables this discussion by providing the necessary mathematical foundations.

References

M. Gönen, W. O. Johnson, Y. Lu, and P. H. Westfall. The Bayesian two-sample t test. *The American Statistician*, 59(3):252–257, 2005.

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

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

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.

J. N. Rouder and R. D. Morey. A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18(4):682–689, 2011.

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(2): 225–237, 2009.

J. N. Rouder, J. M. Haaf, and F. Aust. From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, 85(1):41–56, 2018a.

J. N. Rouder, J. M. Haaf, and J. Vandekerckhove. Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1): 102–113, 2018b.

B. Rüger. *Test-und Schätztheorie: Band I: Grundlagen*. De Gruyter Oldenbourg, 1998.

W. Vanpaemel. Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54(6):491–498, 2010.

W. Vanpaemel and M. D. Lee. Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, 19(6):1047–1056, 2012.

Contribution 3

Schwaferts & Augustin (2021b): Bayes Factors can Only Quantify Evidence w.r.t. Sets of Parameters, not (Prior) Distributions on the Parameter. (Preprint)

Bayes Factors can only Quantify Evidence w.r.t. Sets of Parameters, not w.r.t. (Prior) Distributions on the Parameter

Patrick Schwaferts, Thomas Augustin
Ludwig-Maximilians-Universität Munich, Germany

Bayes factors are characterized by both the powerful mathematical framework of Bayesian statistics and the useful interpretation as evidence quantification. Former requires a parameter distribution that changes by seeing the data, latter requires two fixed hypotheses w.r.t. which the evidence quantification refers to. Naturally, these fixed hypotheses must not change by seeing the data, only their credibility should! Yet, it is exactly such a change of the hypotheses themselves (not only their credibility) that occurs by seeing the data, if their content is represented by parameter distributions (a recent trend in the context of Bayes factors for about one decade), rendering a correct interpretation of the Bayes factor rather useless. Instead, this paper argues that the inferential foundation of Bayes factors can only be maintained, if hypotheses are sets of parameters, not parameter distributions. In addition, particular attention has been paid to providing an explicit terminology of the big picture of statistical inference in the context of Bayes factors as well as to the distinction between knowledge (formalized by the prior distribution and being allowed to change) and theoretical positions (formalized as hypotheses and required to stay fixed) of the phenomenon of interest.

Introduction

Statistical hypotheses have always been sets of parameters in classic frequentist hypothesis tests. However, in the context of Bayes factors – a prominent Bayesian method for hypothesis comparisons (Gönen, Johnson, Lu, & Westfall, 2005; Jeffreys, 1961; Kass & Raftery, 1995; Rouder, Speckman, Sun, Morey, & Iverson, 2009) – it is argued that also the prior distribution on the parameter might be employed to represent the hypotheses (or "models") that should be contrasted against each other. This view was promoted primarily by Vanpaemel (2010) in an attempt to turn one of the fundamental issues of Bayes factors, namely its prior sensitivity (see e.g. Kass & Raftery, 1995; Kruschke, 2015; Liu & Aitkin, 2008; Sinharay & Stern, 2002), from a limitation to a feature. By now, this view can be found within many other publications, sometimes rather explicitly, sometimes only implicitly (see e.g. Dienes (2019, p. 364f), Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2018, p. 228), Heck et al. (2020, p. 5), Morey, Romeijn, and Rouder (2016, p. 16), Rouder, Morey, and Wagenmakers (2016), Tendeiro and Kiers (2019, p. 776, 780), Vanpaemel and Lee (2012)). Even the authors of this paper were previously influenced by this view (Ebner, Schwaferts, & Augustin, 2019). However, as will be outlined within this paper, the inferential foundation of Bayes factors is severely impaired when representing statistical hypotheses via parameter distributions. Instead, statistical hypotheses need to be sets of parameters only, even in Bayesian statistics.

In order to elaborate these considerations, an explicit terminology is outlined first, building on the framework by Kass (2011). Subsequently, updating consistency of Bayes factors is given a detailed account to determine conditions that lead to inconsistencies. Finally, the representation of hypotheses by sets of parameters and by prior distributions is assessed, respectively, showing that former does not suffer foundational issues, only latter does.

Author information:

Patrick Schwaferts: patrick.schwaferts@stat.uni-muenchen.de
Thomas Augustin: thomas.augustin@stat.uni-muenchen.de

Methodological Foundations of Statistics and its Applications
Department of Statistics
Ludwig-Maximilians-Universität Munich
Ludwigsstraße 33, 80539 Munich, Germany

Big Picture

Formalization and Interpretation

A comprehensive view of statistical inference distinguishes between the *real world* and a *theoretical world* (Kass, 2011), where latter contains mathematical formalizations of the relevant characteristics in the real world. Interpreting the components of the theoretical world leads to their counterparts in the real world. In that sense, both worlds can be connected by *formalization* and *interpretation* (see Figure 1). Based on this general view by Kass (2011), the pig picture of statistical inference in the context of the Bayes

factors shall be derived in detail in the following.

Typically, a researcher designs a scientific investigation to assess a phenomenon of interest. This scientific investigation leads to data that are described by a parametric sampling distribution, and the *parameter* (which lives in the theoretical world) should correspond to *phenomenon of interest* in the real world. If the correct interpretation of the parameter does not match with the phenomenon a researcher is interested in, then the design of the scientific investigation might be reconsidered. For the reminder of this paper, a proper correspondence between the parameter and the phenomenon of interest is assumed.

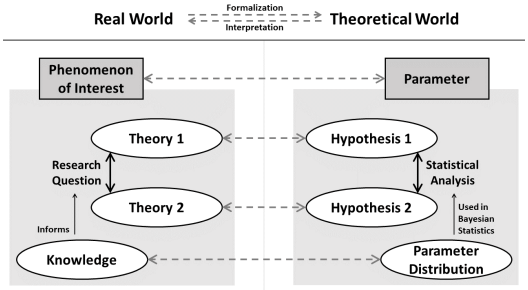

Figure 1. Big Picture of Statistical Inference in the Context of Bayes Factors. While the general view about the real world and the theoretical world is elaborated on by Kass (2011), the big picture of statistical inference in the context of Bayes factors is elaborated on in detail within this paper.

Within this fundamental framework, the big picture of statistical inference in the context of Bayes factors shall be established (as eventually depicted in Figure 1). This, however, is not easy, as relevant terms, as e.g. "theory", "model", or "hypothesis", have a multitude of different meanings and usages. In that, it is mandatory to explicitly define the employed terms such that their usage can be universally agreed on. Therefore, this elaboration should start at the very beginning, namely with two undeniable mathematical properties of Bayes factors.

Common Ground: The Bayes Factor

The Bayes factor (formulas below) is a quantity that is used within a *statistical analysis* and lives in the theoretical world. Two mathematical properties of Bayes factors cannot be denied:

- It is a Bayesian quantity, such that it requires a distribution on the parameter.
- It has a contrasting nature and contrasts two mathematical objects against each other (frequently referred to as hypotheses or models) (cp. e.g. Rouder et al., 2016, Element #2 on p. 16).

Put the other way: Without a parameter distribution or without contrasting two mathematical objects (hypotheses or models) there cannot be a Bayes factor. Accordingly, whenever a Bayes factor is employed it is safe to assume the existence of a parameter distribution and the existence of a contrast and its two contrasted objects. Although these two simple facts about Bayes factors might seem trivial, it is important to state them explicitly, as it is exactly these two properties that serve as the basis to derive the big picture of statistical inference in the context of Bayes factors. While it is expected that there is a common agreement upon these two facts, there might be different views about other concepts employed in the context of Bayes factor (e.g. about the nature of hypotheses). By starting the elaboration with these two facts that can be agreed on, it is possible to assess the origin of disagreements about other concepts.

Parameter Distribution and Knowledge

Parameter distributions (e.g. prior or posterior) live in the theoretical world and are typically interpreted as knowledge (see e.g. Jaynes, 2003) or uncertainty (see e.g. Kruschke, 2015) or degrees of belief (see e.g. Jeffreys, 1961) or information (see e.g. Berger, 1985) about the phenomenon of interest. Within this paper, the term knowledge shall be employed, as the exact label is not relevant for the elaborations below, only the fact that it is the interpretation of the parameter distribution. *Accordingly, define the term knowledge (about a phenomenon of interest) within this paper as the interpretation of a Bayesian parameter distribution.*

Hypotheses (or Models)

The mathematical objects in the theoretical world that are contrasted by the Bayes factor shall be referred to as (statistical) *hypotheses* (although the attribute "statistical" will e omitted as the term hypothesis is not employed in a non-statistical sense within this paper). Other publications (e.g. Rouder, Haaf, & Aust, 2018; Rouder et al., 2016) might state that the Bayes factor contrasts two *models* against each other, yet the formula of the Bayes factor is exactly the same as in those publications that contrast hypotheses against each other (cp. also Morey et al., 2016, p. 11). Therefore, these models are the same mathematical objects as the hypotheses within this paper (namely those mathematical objects that are contrasted against each other by the Bayes factor). Other authors (e.g. Kruschke & Liddell, 2018) use both terms (model and hypothesis) rather interchangeably (cp. also Tendeiro & Kiers, 2019, p. 775, esp. footnote 1). In the remainder of this paper, the term "model" shall be avoided, as it appears to have a variety of different other usages as well. *Accordingly, define the term hypothesis within this paper as one of the two mathematical objects that are contrasted against each other by the Bayes factor.*

To assess the nature of this mathematical object is the aim of this paper and it will be argued that it can only be a set of parameters and not a parameter (prior) distribution.

Theoretical Positions (or Theories) and Research Question

The Bayes factor contrasts two mathematical objects against each other in the theoretical world, and the same scheme applies to the real world after interpretation: There is a contrast between two *theoretical positions* about the phenomenon of interest in the real world. *In that sense, define the term theoretical position within this paper as the interpretation of a hypothesis*.

The respective *research question* contrasts these two theoretical positions against each other. Please note that, in general, the nature of potential research questions about the phenomenon of interest is extremely versatile. However, only those research questions can be answered by Bayes factors, that contrast two theoretical positions against each other. If a research question contrasts two theories against each other, which cannot be formalized as those mathematical objects that are contrasted by the Bayes factor, then the Bayes factor is not suitable to answer such a research question.

Typically, the term "theory" is employed instead of "theoretical position", and it is said that the Bayes factor compares two "theories" (cp. also Rouder, Haaf, & Aust, 2018, who use both terms). In this context, both terms (theory or theoretical position) denote the same, namely the interpretation of a hypothesis (i.e. the interpretation of the mathematical objects that are contrasted against each other by the Bayes factor). However, the term "theory" might be used in a multitude of different other ways as well, e.g. in a non-contrasting context or such that it cannot be formalized as a hypothesis in the context of Bayes factors. To avoid confusion and to emphasized its contrasting nature, only the term "theoretical position" shall be employed within this paper.

Summary Terminology

So far, the concepts of the big picture of statistical inference in the context of Bayes factor have been outlined and it should be emphasized that the terms "hypothesis", "theoretical position", and "knowledge" are used within this paper to facilitate an understandable elaboration. In fact, it might have been possible to merely use the descriptions "mathematical objects that are contrasted against each other by the Bayes factor" (hypotheses), "interpretation of these mathematical objects" (theoretical positions), and "interpretation of a parameter distribution" (knowledge). Together with the two above mentioned undeniable properties about Bayes factors, namely the existence of a contrast (of two mathematical objects) and the existence of a Bayesian parameter distribution, the employed concepts should have been explicitly outlined. Other publications might employ a different terminology, such that it is necessary to check which concepts are actually referred to by each term in each publication.

Statistical Inference with Bayes Factors

Statistical inference is the procedure of deriving conclusions from observed data. Naturally, there is a variety of different inferential approaches, each using different principles to extract information from the observed data. The elegance of Bayes factors might be attributed to the fact that they combine two different approaches to statistical inference in one single quantity: Bayesian learning and evidential quantification.

- Within the Bayesian approach to statistical inference, a parameter prior distribution gets updated to a parameter posterior distribution by including the information from the observed data via Bayes rule. Conclusions are then derived solely from the parameter distribution.
- Within the evidential approach to statistical inference (cp. e.g. Berger & Wolpert, 1988; Blume, 2011; Royall, 1997, 2004), the information within the data are used to quantify evidence w.r.t. two different fixed theoretical positions about a phenomenon of interest. Assume two theoretical positions A and B are of interest (and specified in the research question) and assume the evidence within the data is quantified to be 5, then the evidential interpretation is: After observing the data the credibility of A over B is 5-times higher than before the data were observed (see e.g. Morey et al., 2016)

On the one hand, while Bayesian statistics is able to answer also different research questions, by using Bayes factors the nature of potential research questions is limited to those that contrast theoretical positions, thus allowing a useful and intuitive interpretation in the context of evidential quantification. On the other hand, while the framework of evidential quantification might consider a variety of different contrasting statistical analysis, by using Bayes factors the statistical analysis is restricted to the Bayesian framework, providing a thorough and powerful mathematical foundation (see e.g. Berger, 1985; Jeffreys, 1961).

Knowledge vs. Theoretical Positions

Accordingly, to consider the inferential foundation of Bayes factors comprehensively, both the knowledge about the phenomenon of interest (Bayesian inference) and the theoretical positions about the phenomenon of interest (evidential inference) need to be distinguished. These concepts are fundamentally different! While Bayesian learning allows (or even requires) the knowledge itself to be altered, theoretical positions in the context of evidential quantification stay fixed, only their credibility may change. Without such a clear

distinction, the inferential foundation of Bayes factors might break apart:

- If the theoretical positions themselves (not only their credibility) are allowed to change by observing the data, then the useful and intuitive interpretation as evidence quantification is lost. Assume two theoretical positions A and B are of interest (and specified in the research question) but change by seeing the data to the theoretical positions C and D, respectively, and assume the evidence within the data is quantified to be 5, then the correct but useless interpretation is: The credibility of C over D after the data were observed is 5-times higher than the credibility of A over B before the data were observed.

- If the knowledge (and thus the parameter distribution) is forced to stay fixed although some data were observed, then Bayes rule is not applied, leading to updating inconsistencies (outlined in detail below).

Accordingly, a clear distinction between knowledge (formalized as parameter distributions and being allowed to change) and theoretical positions (formalized as statistical hypotheses which stay fixed) about a phenomenon of interest is mandatory. In that, the framework depicted here (Figure 1) does account for the fundamental different nature of knowledge and theoretical positions about a phenomenon of interest.

Interestingly, on a side note, it might be stated that – by its nature – also the prior knowledge is able to inform the research question. Typically, prior knowledge is insufficient to answer the research question adequately, justifying the necessity to conduct a scientific investigation. However, the structure of how to answer the research question with the available knowledge is independent of whether data were observed or not: In a Bayesian context, it is a parameter distribution from which the answer to the research question is derived, and this way of deriving answers might work both for the prior and the posterior distribution.

One-to-One Correspondence

Ideally, the correspondence between the concepts of the real world with those in the theoretical world (gray dashed arrows in Figure 1) should be one-to-one, mathematically described by a bijective mapping. Without such a bijective mapping, chosen formalizations might be arbitrary or the interpretation of the results might inform past the research question. While the correspondence between the phenomenon of interest and the parameter depends on the quality of the experimental design, and the mapping between parameter distributions and knowledge is typically assumed to be bijective in the Bayesian setting (two different distributions represent two different bodies of knowledge), of interest for this elaboration is the relation between theoretical positions and hypotheses. Consider two cases:

- There is a bijective mapping between theoretical positions and hypotheses. Then, different hypotheses represent different theoretical positions.

- A variety of different hypotheses formalize one single theoretical position. When forced to commit oneself to one of those hypotheses (as typically required by the statistical analysis), this choice might intuitively be called instantiation: The theoretical position is formally instantiated by one of the hypotheses. This terminology is frequently employed it the literature that (potentially implicitly) assume hypotheses to be represented by distributions (see e.g. Morey et al. (2016, p. 13), Rouder, Haaf, and Aust (2018, p. 2), Vanpaemel (2010, p. 491), Vanpaemel and Lee (2012, p. 1054)), suggesting that such a non-bijective relation might be implicitly assumed. Other publications do also employ such a non-bijective relation without using the term instantiation (e.g. Dienes, 2019, Box 3 on p. 369).

In that, these two types of relations between theoretical positions and hypotheses might describe the ideal and the actual situation, respectively, and will be used below for elaborating issues inherent to representing hypotheses by parameter distributions.

Updating Consistency

In Bayesian statistics, it is Bayes rule and not another principle, which states how a prior distribution gets updated to a posterior distribution (i.e. how information is extracted from the observed data; see Figure 2, top-left). If the prior distribution reflects all available knowledge about the phenomenon of interest *before* the investigation is conducted, then the corresponding posterior distribution reflects all available knowledge about the phenomenon of interest *after* the investigation is conducted. Disagreeing would imply that Bayes rule is not able to extract all information within the observed data that is relevant for the phenomenon of interest (and no Bayesian would do so). As a consequence, the posterior distribution might be employed as a prior distribution for the Bayesian analysis of a data set obtained in a subsequent investigation with the same design (see Figure 2, top-right). Naturally, the final posterior distribution after subsequently updating twice should be identical to the posterior distribution obtained by merging both data sets first and then updating the initial prior distribution at once (see Figure 2, bottom). If so Bayesian updating is *consistent* (cp. Rüger, 1998, p. 190), else it is *inconsistent*.

In general, updating with Bayes factors is consistent (Schwaferts & Augustin, 2021) (see Figure 3a). Assume the observed data $x = (x_1, \ldots, x_n)$ consists of independent and

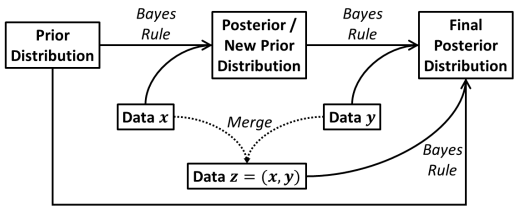

Figure 2. Consistent Bayesian Updating. Updating subsequently with two independent data sets x and y (top path) should yield the same final posterior distribution than merging both data sets first and then updating at once (bottom path).

identically distributed observations x_i ($i = 1, \ldots, n$) that follow the parametric sampling distribution with density $f(x_i|\theta)$, where $\theta \in \Theta$ is the parameter (representing the phenomenon of interest), such that the density of the complete data set x is $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$. The hypotheses H_0 and H_1 have prior probabilities $p(H_0)$ and $p(H_1) = 1 - p(H_0)$, and the corresponding densities of the hypothesis-based parameter distributions are $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$, respectively. Then the density of the overall prior distribution is (see e.g. Rouder, Haaf, & Vandekerckhove, 2018)

$$\pi(\theta) = p(H_0)\,\pi(\theta|H_0) + p(H_1)\,\pi(\theta|H_1)\,. \tag{1}$$

The Bayes factor

$$BF^{x} = \frac{\int f(x|\theta)\,\pi(\theta|H_1)\,d\theta}{\int f(x|\theta)\,\pi(\theta|H_0)\,d\theta} \tag{2}$$

is calculated using only the data x and the hypothesis-based parameter densities $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$, and allows to update the prior probabilities of the hypotheses to their posterior probabilities (see Figure 3a, left):

$$\frac{p(H_1|x)}{p(H_0|x)} = BF^{x} \cdot \frac{p(H_1)}{p(H_0)}\,. \tag{3}$$

In addition, revealed by simply applying Bayes rule consistently to the overall prior density $\pi(\theta)$ (depicted in detail by Schwaferts & Augustin, 2021), also the hypothesis-based parameter densities $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$ get updated by the data x to their posterior densities $\pi(\theta|H_0, x)$ and $\pi(\theta|H_1, x)$ (gray arrow in Figure 3a, left) (cp. also Kruschke & Liddell, 2018). In general (i.e. for non-degenerate prior distributions), these posteriors are different than the priors. These updated hypothesis-based posterior densities together with the posterior probabilities on the hypotheses describe the overall posterior distribution:

$$\pi(\theta|x) = p(H_0|x)\,\pi(\theta|H_0, x) + p(H_1|x)\,\pi(\theta|H_1, x)\,. \tag{4}$$

If a new data set y was observed (using the same experimental setup, i.e. following the same sampling distribution), this updated posterior distribution describes the starting point for a subsequent analysis with Bayes factors (Figure 3a, right). Consequently, the corresponding Bayes factor

$$BF^{y|x} = \frac{\int f(y|\theta)\,\pi(\theta|H_1, x)\,d\theta}{\int f(y|\theta)\,\pi(\theta|H_0, x)\,d\theta} \tag{5}$$

is also inherently influenced by the information within the previous data set x. In that, however, updating is consistent (a complete proof is provided by Schwaferts & Augustin, 2021).

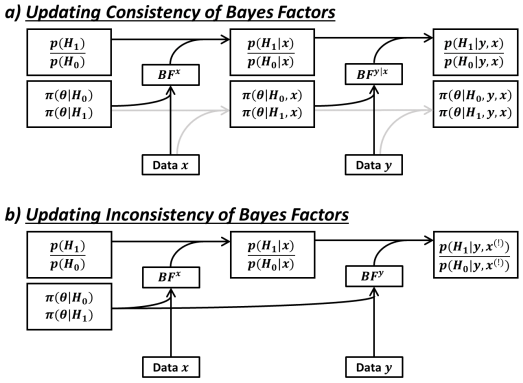

Figure 3. Consistent (a) and Inconsistent (b) Updating with Bayes Factors. Superscript (!) indicates that not all relevant information was extracted from the data set x.

Updating inconsistencies occur, if the second data set y is analyzed with the initial hypothesis-based prior distributions (i.e. $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$ as in equation (2)) although the first data set x was already observed (Figure 3b). This happens, if the initial hypothesis-based prior distributions do not get updated in the analysis of the first data set x, which is a violation of Bayes rule (cp. also Rouder & Morey, 2011, who noticed this issue and solved it properly by merging all data sets first). It is difficult to assess the prevalence of such updating inconsistencies in the current scientific literature, as the data is typically analyzed at once (not subsequently), and the calculation of the Bayes factor can be done without explicitly updating the hypothesis-based priors. Nevertheless, this update must not be neglected to be consistent (Figure 3a) and the extent to which this fact is overlooked might be indicated by Tendeiro and Kiers (2019): After a year of literature review about Bayes factors to understand them, the authors (and possibly their reviewers from the journal *Psychological Methods* as well) were convinced (see footnote 2 on p. 776 therein) that the hypothesis-based priors do not get updated to their posteriors. Interestingly, at the same place, the authors refer to Kruschke and Liddell (2018), who, in contrast, elaborated on the update of the hypothesis-based priors rather explicitly (see p. 157f and Fig. 4 and 6 therein).

This is a vivid sign of the confusion a researcher faces in the literature about Bayes factors.

Hypotheses as Sets of Parameters

At first, assume that hypotheses are represented by two disjoint subsets $\Theta_0, \Theta_1 \subset \Theta$ of the parameter space Θ:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \,. \tag{6}$$

These subsets need to be chosen to correspond to the theoretical positions that are contrasted within the research question. In addition to these theoretical positions, there is knowledge about the phenomenon of interest that is formalized by a parameter distribution with density $\pi(\theta)$. Without loss of generality, assume that this distribution has a positive density only for parameter values that are contained in one of the hypotheses. The prior probabilities of the hypotheses are obtained from this parameter distribution by

$$p(H_0) = \int_{\Theta_0} \pi(\theta)\, d\theta \quad \text{and} \quad p(H_1) = \int_{\Theta_1} \pi(\theta)\, d\theta \,, \tag{7}$$

and, once the data x were observed and the prior density $\pi(\theta)$ was updated to $\pi(\theta|x)$, the posterior probabilities of the hypotheses are

$$p(H_0|x) = \int_{\Theta_0} \pi(\theta|x)\, d\theta \quad \text{and} \quad p(H_1|x) = \int_{\Theta_1} \pi(\theta|x)\, d\theta \,. \tag{8}$$

How the data change the probabilities of the hypotheses is described by the Bayes factor

$$BF^x = \frac{p(H_1|x)}{p(H_0|x)} \,/\, \frac{p(H_1)}{p(H_0)} \,. \tag{9}$$

Accordingly, the probabilities of the hypotheses change but the hypotheses themselves, i.e. the sets Θ_0 and Θ_1, stay the same. In that, the Bayes factor can be interpreted appropriately as evidence quantification. Further, the overall prior distribution ($\pi(\theta)$) was updated completely to the overall posterior distribution ($\pi(\theta|x)$), providing updating consistency. Consequently, hypotheses can safely be represented by sets of parameters in the context of Bayes factors.

Before continuing, the current situation shall be characterized further (which will be needed below). Any given prior distribution with density $\pi(\theta)$ formalizes knowledge about the phenomenon of interest. One part of this knowledge relates to the one and another part to the other theoretical position (which are contrasted within the research question), formalized by the hypothesis-based prior densities $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$. Mathematically, these densities are obtained from the initial density $\pi(\theta)$ by

$$\pi(\theta|H_0) = \frac{1}{p(H_0)} \cdot \pi(\theta)|_{\Theta_0} \tag{10}$$

$$\pi(\theta|H_1) = \frac{1}{p(H_1)} \cdot \pi(\theta)|_{\Theta_1} \,, \tag{11}$$

where $\pi(\theta)|_{\Theta_0}$ and $\pi(\theta)|_{\Theta_1}$ are the densities $\pi(\theta)$ restricted to the sets Θ_0 and Θ_1, respectively. Now, consider the set $\mathcal{X}$ of all potentially observable data sets of any size $n \in \mathbb{N}_0$, with $n = 0$ referring to the empty data set (representing the prior situation). For a given prior distribution with density $\pi(\theta)$, denote the set of all potentially obtainable posterior densities as

$$\Pi := \{\pi(\theta|x) \mid x \in \mathcal{X}\} \,. \tag{12}$$

This set contains all possible posterior distributions, i.e. represents all possible bodies of knowledge about the phenomenon of interest, that might be available after some (yet unknown) data x were observed. Analogously, all different possible bodies of knowledge about the theoretical positions, respectively, are formally contained within the sets

$$\Pi_0 := \{\pi(\theta|H_0, x) \mid x \in \mathcal{X}\} \tag{13}$$

$$\Pi_1 := \{\pi(\theta|H_1, x) \mid x \in \mathcal{X}\} \,. \tag{14}$$

These sets contain only probability distributions with probability mass in the sets Θ_0 and Θ_1, respectively. In summary, a hypothesis and all potentially obtainable bodies of knowledge about this hypothesis (in the context of given prior knowledge and a certain experimental setup) can be described by the sets Θ_0 and Π_0 or Θ_1 and Π_1, respectively.

Hypotheses as Parameter Distributions

Now, prior distributions with densities $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$ shall represent the hypotheses H_0 and H_1, respectively. Ideally, the mapping between theoretical positions and hypotheses should be bijective, updating should be consistent, and theories should not change by seeing the data x, only their credibility should. In the following, two of each of these properties shall be assumed first and then evaluated w.r.t. the third.

Case 1: Bijective Mapping and Updating Consistency

Assume there is a bijective mapping between theoretical positions and hypotheses. As hypotheses are represented by prior distributions, not only by a set of parameters, two different parameter distributions represent two different hypotheses, i.e. two different theoretical positions. Consistent Bayesian updating dictates to update the prior distributions to the posterior distributions, which are (for non-degenerate cases) different than the respective prior distributions. As the parameter distributions change by observing the data, so do the hypotheses and theoretical positions. It that, observing data changes the theories that should be contrasted with each other, not only their credibility. This does not match with the fundamental characteristics of statistical inference by evidence quantification, leading to issues with the interpretation of Bayes factors.

Case 2: Bijective Mapping and Unchanged Theories

Again, assume there is a bijective mapping between theoretical positions and hypotheses, and that hypotheses are represented by the prior distributions, such that two different parameter distributions represent two different hypotheses, i.e. two different theoretical positions. If theories should not change by observing data, the posterior distribution needs to be the same as the prior distribution, which is outlined in Figure 3b and leads – for non-degenerate prior distributions – to updating inconsistency. In that, inference does not follow Bayes rule.

Case 3: Updating Consistency and Unchanged Theories

Now, also allow the mapping between theoretical positions and hypotheses to be non-bijective, in a sense that a single theoretical position might be formalized by a multitude of different hypotheses. If hypotheses are represented by prior distributions, then a set of different distributions corresponds to one theoretical position. How does this set look like, if updating should be consistent and a proper evidential interpretation of Bayes factors shall be kept?

To answer this question, assume that in the specifications of a statistical analysis each of both theoretical positions is instantiated by only one prior distribution with density $\pi(\theta|H_0)$ or $\pi(\theta|H_1)$, respectively, such that their supports do not overlap with each other (overlapping hypotheses will be discussed subsequently). Denote their supports (i.e. the sets of parameters in which the density has non-zero, positive mass) with Θ_0 and Θ_1, respectively. If updating shall be consistent, then parameter distributions are allowed to change by seeing the data. Considering all potentially observable data sets $x \in \mathcal{X}$ (of any size $n \in \mathbb{N}_0$), the initial prior densities $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$ might result in any posterior density within the sets Π_0 and Π_1 (equations (13) and (14)), respectively. To keep the proper evidential interpretation of Bayes factors, all these parameter densities within the sets Π_0 and Π_1 need to represent the same theoretical position, respectively. In that, the hypotheses H_0 and H_1 are represented by the sets Π_0 and Π_1, which, however, contain all potentially observable parameter distributions with positive probability mass restricted to the parameter sets Θ_0 and Θ_1, respectively. Two different parameter distributions with the same support (either Θ_0 or Θ_1) do not differentiate between two theoretical positions, only two different supports, i.e. sets of parameters, do. Accordingly, this situation is practically equivalent to representing hypotheses as sets of parameters.

Overlapping Hypotheses

Within these elaborations, it was assumed that the hypotheses are non-overlapping. Mathematically, they might also be overlapping. Consider the case, in which Θ_0 and Θ_1 are not (almost everywhere w.r.t. the prior density $\pi(\theta)$) disjoint. Then there are parameter values θ that are contained within both Θ_0 and Θ_1, such that, if these parameter values are true, the posterior distribution will be shifted – for an increasing sample size n – within this overlapping part. As a consequence, even an infinitely large data set cannot decisively distinguish between both hypotheses (i.e. answer the research question), and the Bayes factor has a finite limit. Formally, the behavior[1] of the Bayes factor is

$$BF \begin{cases} \to 0 & \text{if } \theta^* \in \Theta_0 \setminus \Theta_1 \\ \to \infty & \text{if } \theta^* \in \Theta_1 \setminus \Theta_0 \\ \to c(\theta^*) & \text{if } \theta^* \in \Theta_0 \cap \Theta_1 \end{cases} \quad \text{for } n \to \infty, \tag{15}$$

where θ^* is the true parameter and $c(\theta^*)$ is a fixed value that depends on θ^* (cp. Morey & Rouder, 2011, p. 411).

In that, if – for a given investigational setup – the theoretical positions (that are of interest in the context of the research question) are reasonably formalized by overlapping hypotheses, it might happen that the scientific investigation cannot answer the research question. This might be a waste of time and money, and cannot be argued to yield a "strong inference" (Platt, 1964) or constitute a "severe test" (Popper, 2002[1935]), claims frequently raised by promoters of the Bayes factor (see e.g. Dienes (2019, p. 365), Etz et al. (2018, p. 228), Schönbrodt and Wagenmakers (2018, p. 130)), which apparently require non-overlapping hypotheses. In this context it shall be noted that even Jeffreys (1961, e.g. p. 269) himself tried to avoid the possibility of obtaining a finite, non-zero Bayes factor limit at all costs: His derivation of the prior distributions for Bayes factors (now sometimes referred to as default Bayes factors (see e.g. Ly, Verhagen, & Wagenmakers, 2016)) aimed at being able to state decisive evidence (which corresponds to Bayes factors tending to ∞ or 0) even for a fixed number of observations (which is an even stronger requirement than within equation (15)). In this regard, it seems advisable to rethink the investigational setup such that overlapping hypotheses can be avoided.

Example

An example shall be provided that leads to a paradox with the representation of hypotheses via parameter distributions.

[1] This equation (15) has been formulated in a mathematically imprecise way in order to present the relevant points clearly. Actually, the condition $\theta^* \in \Theta_0 \setminus \Theta_1$ constitutes the case in which the true parameter θ^* is within the set $\Theta_0 \setminus \Theta_1$ such that for increasing n the posterior distribution will be shifted into this set $\Theta_0 \setminus \Theta_1$. The other conditions have to be read analogously. There might be cases in which – mathematically – θ^* is within one of the parameter sets that define the conditions, but the posterior distribution will not be shifted completely within the respective set for $n \to \infty$. These cases, however, lie exactly at the borders between the set-valued hypotheses and are expected to occur almost never w.r.t. the parameter distribution.

Assume the data x is characterized by the binomially distributed random quantity $X \sim Bin(n, \theta)$, with $n = 20$ and $\theta \in [0, 1]$ (probability of success) being the parameter of interest. Both hypotheses shall have an identical support $\Theta_0 = \Theta_1 = [0, 1]$, but different prior (beta) distributions (see Figure 4 left, rounded boxes):

$$H_0 : \theta \sim Beta(1, 1) \quad \text{vs.} \quad H_1 : \theta \sim Beta(15, 7)\,. \tag{16}$$

Further, both hypotheses shall have equal prior probabilities $p(H_0) = p(H_1) = 0.5$ (see Figure 4 left).

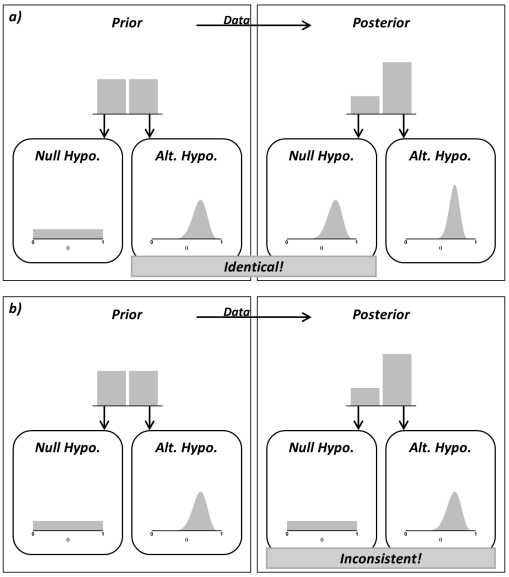

Figure 4. Example. Theoretical positions are represented by parameter distributions with consistent (a) and inconsistent (b) Bayesian updating. Consistent updating yields a situation in which the prior alternative hypothesis and the posterior null hypothesis represent the same theoretical position (This illustration of the hypothesis-based parameter distributions was inspired by Kruschke & Liddell, 2018).

Now, assume that $s = 14$ successes were observed. The resulting Bayes factor is $BF = 2.89$ leading to posterior probabilities $p(H_0|s) = 0.257$ and $p(H_1|x) = 0.743$, favoring the alternative hypothesis H_1 (see Figure 4 right). However, also the within-hypothesis beta distributions get updated by observing s to (see Figure 4a, right, rounded boxes; else updating is inconsistent (see Figure 4b))

$$H_{0|s} : \theta|s \sim Beta(15, 7) \quad \text{vs.} \quad H_{1|s} : \theta|s \sim Beta(29, 3)\,. \tag{17}$$

This example was constructed such that the posterior null hypothesis $H_{0|s}$ has the same distribution as the prior alternative hypothesis H_1. If a theoretical position is represented by a parameter distribution, then both of these hypotheses represent the same theoretical positions (Figure 4a). Although the prior alternative hypothesis H_1 *gains* credibility by observing s, the posterior null hypothesis $H_{0|s}$ has *less* credibility due to observing s. Paradoxically, both hypotheses represent the same theoretical position, so it is not clear whether the data agree or disagree with this theoretical position.

In order to solve this paradox, hypotheses need to be considered as sets of parameters. Both hypotheses hypothesize the same parameter set, representing the same theoretical position. Accordingly, the research question, that will be answered within this analysis, contrasts a theoretical position against itself. Naturally, no data can inform this pointless contrast.

Discussion

This paper elaborated that hypotheses should be represented by sets of parameters only, not by parameter distributions. If so, updating consistency and a proper evidential interpretation of Bayes factors are given, and if not, the foundational or evidential basis of Bayes factors is severely impaired. In that, a clear distinction between theoretical positions and knowledge about the phenomenon of interest is mandatory. It is important that the content of theoretical positions should only inform the specification of hypotheses (sets of parameters) and that the available knowledge should only inform the specification of the prior distribution.

Empirical Content

It is argued that by using parameter distributions to represent hypotheses, their *empirical content* can be increased (Vanpaemel & Lee, 2012, p. 1052), supplemented by references to Popper (2002[1935]). However, the elaborations within this paper might cast doubt. According to Popper (2002[1935], p. 103), the empirical content of a statement is "the class of its potential falsifiers", such that a higher empirical content is characterized by a larger class of potential falsifiers. Now, one needs a proper concept of "falsifiability" in the Bayesian framework, and it appears that defining such a concept is fundamentally difficult, as Popper's elaborations of induction are restricted to the *modus tollens* of deductive tests (Popper, 2002[1935], p. 19) while the Bayesian framework tries to formalize induction by a logic of partial beliefs (cp. e.g. Ly et al., 2016, p. 20). If, at all, one tries to find such a concept, one might say that a hypothesis is falsified if its probability is zero. Naturally, a zero probability of a hypothesis will not be obtained in a scientific investigation (which assumes non-zero prior probabilities), so – practically – one might stop if the probability of a hypothesis is sufficiently small and then decide to treat this hypothesis as falsified. In terms of the limit behavior of Bayes factors, this resembles the case in which the Bayes factor tends towards ∞ or 0 (cp. also Rouder, Haaf, & Vandekerckhove, 2018, p. 105). This refers to the first two cases in equation (15). In the third case, however, evidence will never be conclusive if $n \to \infty$, so it

will not be possible to "falsify" any of the hypotheses with the given experiment. In that, the class of potential falsifiers of H_0 is $\Theta_1 \setminus \Theta_0$ and the class of potential falsifiers of H_1 is $\Theta_0 \setminus \Theta_1$. Therefore, it is only the supports Θ_0 and Θ_1 that determine the empirical content of the hypotheses, not the exact shape of the prior distributions ($\pi(\theta|H_0)$ and $\pi(\theta|H_1)$). Consequently, beyond their mere supports, prior distributions do not increase the empirical content of hypotheses.

Nil-Hypotheses

Acknowledging that hypotheses are only the supports of the within-hypothesis prior distributions, it appears that many elaborations in the context of Bayes factors (e.g. Dienes, 2019; Gönen et al., 2005; Rouder, Haaf, & Aust, 2018; Rouder, Haaf, & Vandekerckhove, 2018; Rouder et al., 2009) do still use a sharp null hypothesis, which hypothesizes only one single parameter value (cp. also Tendeiro & Kiers, 2019, p. 787). In that, these hypotheses are identical to those employed in conventional null hypothesis significance testing (NHST), such that its heavy critique about the uselessness of these hypotheses (see e.g. Berkson, 1938; Cohen, 1994; Gigerenzer, 2004; Kirk, 1996) does apply to these Bayes factors as well. The inclusion of the parameter distribution into the statistical analysis does not tackle these issues (about the uselessness of the employed hypotheses). To do so, hypotheses need to be specified as sets of parameters that correspond to the theoretical positions that are of interest within the research question. Then, these hypotheses are typically not single-valued anymore. In this regard, the methodological development of Bayes factors with reasonably specified interval-valued hypotheses needs to be addressed more intensively. Although few elaborations exist (cp. Heck et al., 2020; Hoijtink, Mulder, van Lissa, & Gu, 2019; Morey & Rouder, 2011), this development is treated as rather ancillary within the Bayes factor literature. Alternative hypothesis-based methods (see e.g. Kruschke, 2015, 2018; Lakens, 2017; Lakens, Scheel, & Isager, 2018) already started to primarily address this necessity of allowing reasonably specified interval-valued hypotheses, and Bayes factors need to go along with them.

Knowledge vs. Theoretical Positions

The central message of this paper is that knowledge and theoretical positions about the phenomenon of interest need to be distinguished. Former inform the specification of the prior distribution, latter inform the specification of the hypotheses. In that sense, both of these mathematical objects (prior distribution, hypotheses) or real world concepts (knowledge, theoretical positions) are independent of each other. This can also be seen, as it is possible to specify a prior distribution without having hypotheses (as in a non-hypothesis-based Bayesian analysis) or as it is possible to specify hypotheses without having a prior distribution (as in non-Bayesian hypothesis-based analyses). Yet, it is possible to depict the prior distribution in dependence of the hypotheses via the within-hypothesis prior distributions (equations (10) and (11)) and the prior probabilities of the hypotheses (equation (7)). Strikingly, after combing these components to the overall prior distribution (equation (1)), its dependence on the hypotheses is gone! Naturally, what is known about the phenomenon of interest does primarily not depend on which hypothetical conjectures might be possible about it. This has serious implications about how to specify the essential quantities in a hypothesis-based Bayesian analysis: It is recommended to specify the overall prior distribution (as density $\pi(\theta)$) and the hypotheses (via Θ_0 and Θ_1) independently. If, in contrast, the within-hypothesis prior distributions shall be specified (via $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$), the applied scientist needs to make sure that by combining them with the prior probabilities of the hypotheses to the overall prior distribution (equation (1)) its dependence on the hypotheses is gone. This seems quite remarkable.

Outlook

Looking at the history of Bayesian statistics, it appears that prior distributions have always had a bad reputation. In this context, it seems that the idea of using prior distributions to formalize theoretical positions was motivated by the intention of correcting this bad reputation of prior distributions. For example, Vanpaemel and Lee (2012, both quotes on p. 1048) stated that they "do not agree that priors are an unwanted aspect of the Bayesian framework" and that they "believe that it is wrong to malign priors as a necessary evil". It can only be agreed on! Parameter distributions are a vital part of Bayesian statistics and must not be condemned! This elaboration clarified the distinction between knowledge (parameter distribution) and theoretical positions (hypotheses), and, therefore, tried to contribute to a correct employment of parameter distributions in the Bayesian framework.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (Second ed.). New York: Springer. doi: 10.1007/978-1-4757-4286-2

Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. *Lecture Notes-Monograph Series*, *6*, iii–160.2 (discussion: 160.3–199). Retrieved from `http://www.jstor.org/stable/4355509`

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*(203), 526–536. doi: 10.2307/2279690

Blume, J. D. (2011). Likelihood and its evidential framework. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 493–511). Elsevier. doi: 10.1016/B978-0-444-51862-0.50014-9

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

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. doi: 10.1177/2515245919876960

Ebner, L., Schwaferts, P., & Augustin, T. (2019). Robust Bayes factor for independent two-sample comparisons under imprecise prior information. In J. De Bock, C. P. de Campos, G. de Cooman, E. Quaeghebeur, & G. Wheeler (Eds.), *Proceedings of the eleventh international symposium on imprecise probability: Theories and applications* (Vol. 103, pp. 167–174). PMLR. Retrieved from `http://proceedings.mlr.press/v103/ebner19a.html`

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*, 219–234. doi: 10.3758/s13423-017-1317-5

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi: 10.1016/j.socec.2004.09.033

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, *59*, 252–257. doi: 10.1198/000313005X55233

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., … others (2020). A review of applications of the Bayes factor in psychological research. doi: 10.31234/osf.io/cu43g

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

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press. doi: 10.1017/CBO9780511790423

Jeffreys, H. (1961). *Theory of probability* (Third ed.). Oxford: Oxford University Press.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *26*(1), 1–9. doi: 10.1214/10-STS337

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. doi: 10.2307/2291091

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. doi: 10.1177/0013164496056005002

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York: Academic Press. doi: 10.1016/B978-0-12-405888-0.09999-2

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi: 10.1177/2515245918771304

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. doi: 10.3758/s13423-017-1272-1

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi: 10.1177/2515245918770963

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. doi: 10.1016/j.jmp.2008.03.002

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi: 10.1016/j.jmp.2015.06.004

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. doi: 10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. doi: 10.1037/a0024377

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353. doi: 10.1126/science.146.3642.347

Popper, K. (2002[1935]). *The logic of scientific discovery* (2nd ed.). London: Routledge Classics. doi: 10.4324/9780203994627

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*(1), 41–56. doi: 10.1080/03637751.2017.1394581

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. doi: 10.3758/s13423-017-1420-7

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689. doi: 10.3758/s13423-011-0088-7

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra: Psychology*, *2*(1). doi: 10.1525/collabra.28

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

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman and Hall. doi: 10.1201/9780203738665

Royall, R. (2004). The likelihood paradigm for statistical evidence. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 119–152). University of Chicago Press. doi: 10.7208/CHICAGO/9780226789583.003.0005

Rüger, B. (1998). *Test-und Schätztheorie: Band I: Grundlagen*. De Gruyter Oldenbourg. doi: 10.1524/9783486599701

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

Schwaferts, P., & Augustin, T. (2021). *Updating consistency in Bayes factors* (Tech. Rep. No. 236). Ludwig-Maximilians-University Munich, Department of Statistics. doi: 10.5282/ubm/epub.75073

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*(3), 196–201. doi: 10.1198/000313002137

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

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. doi: 10.1016/j.jmp.2010.07.003

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056. doi: 10.3758/s13423-012-0300-4

Contribution 4

Ebner, Schwaferts & Augustin (2019): Robust Bayes Factor for Independent Two-Sample Comparisons Under Imprecise Prior Information. (ISIPTA)

Proceedings of Machine Learning Research 103:167–174, 2019 ISIPTA 2019

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

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

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

Abstract

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

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

1. Introduction

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

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

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

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

2. Bayes Factor

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

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

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

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

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

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

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

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

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

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

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

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

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

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

as the marginal likelihood under H_0.

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

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

Finally, the marginal likelihood under H_1 ensues as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

is typically termed the effective sample size.

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

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

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

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

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

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

3. Robust Bayes Factor

3.1. Theory

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

3.2. Example

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

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

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

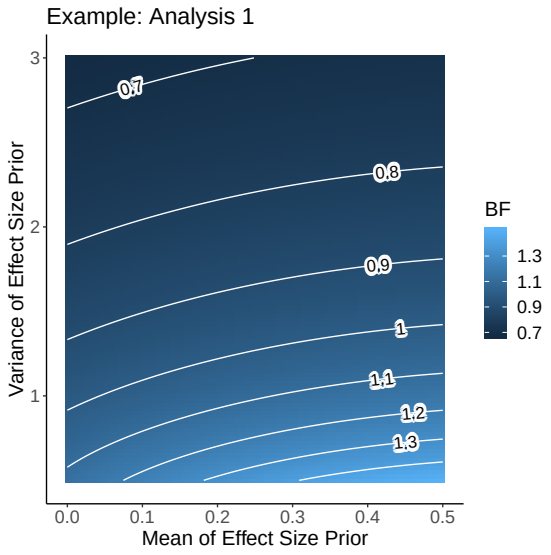

Figure 1: Dependence of the Bayes Factor value (color) on the mean μ_δ (x-axis) and variance σ_δ^2 (y-axis) of the normal effect size prior within the first exemplary analysis.

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

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

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

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

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

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

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

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

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

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

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

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

4. Discussion

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

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

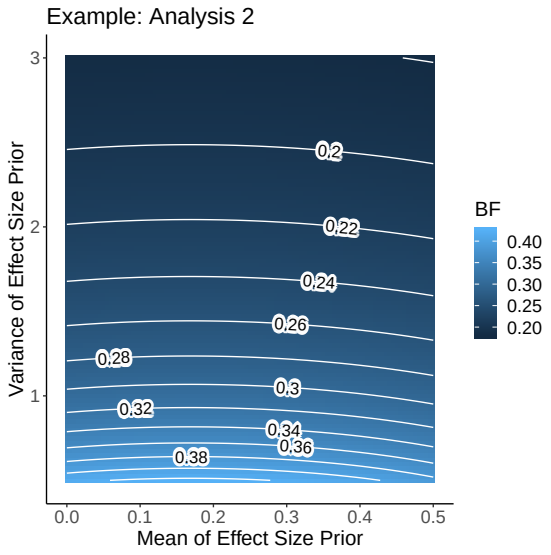

Figure 2: Dependence of the Bayes Factor value (color) on the mean μ_δ (x-axis) and variance σ^2_δ (y-axis) of the normal effect size prior within the second exemplary analysis.

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

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

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

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

5. Outlook

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

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

Appendix A. R Code

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contribution 5

Schwaferts & Augustin (2019): Imprecise Hypothesis-Based Bayesian Decision Making with Simple Hypotheses. (ISIPTA)

Proceedings of Machine Learning Research 103:338–345, 2019 ISIPTA 2019

Imprecise Hypothesis-Based Bayesian Decision Making With Simple Hypotheses

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

Abstract

Applied real-world decisions are frequently guided by the outcome of hypothesis-based statistical analyses. However, most often relevant information about the phenomenon of interest is available only imprecisely, and misleading results might be obtained, in particular, by either ignoring relevant information or pretending a level of knowledge that is not given. In order to be able to include (partial) information authentically in the imprecise form it is available, this paper tries to extend the framework of hypothesis-based Bayesian decision making with simple hypotheses to be able to deal with imprecise information about the three relevant quantities: hypotheses, prior beliefs, and loss function. Although straightforward at first glance, it appears that by specifying the hypotheses imprecisely, Bayesian updating of the prior beliefs might be inconsistent. In that, this paper provides the basic mathematical formulation to further extend imprecise hypothesis-based Bayesian decision theory to more elaborate contexts, such as those involving composite imprecise hypotheses, and in addition highlights the necessity of paying particular attention to the depicted updating issues.

Keywords: Hypotheses, Likelihood Ratio, Imprecise Probabilities, Bayesian Decision Theory, Sequential Updating, Inconsistency, Statistics in Psychological Research

1. Introduction

In the face of the currently discussed reproducibility crisis in psychological research (Ioannidis, 2005), Bayesian statistics is gaining popularity (e.g. Van De Schoot et al., 2017) also in this area. Classical hypotheses tests are argued to be replaced by the so called Bayes factor (e.g. Kass and Raftery, 1995; Gönen et al., 2005; Rouder et al., 2009), a Bayesian quantity for hypothesis comparisons, which might be seen as a generalization of the likelihood ratio to include prior information about the parameter of interest by employing prior distributions on it. If these distributions are degenerate, i.e. have all mass on a single parameter value, the Bayes factor equals the likelihood ratio.

In addition to the prior distributions on the parameter, a Bayesian analysis in the context of statistical hypotheses requires prior probabilities of these hypotheses, which might be interpreted as subjective *belief* in the respective hypotheses and get updated by the data. It is the Bayes factor, which quantifies the change in these subjective probabilities (e.g. Morey et al., 2016), and therefore the Bayes factor is interpreted as quantification of the *evidence* in the data w.r.t. the hypotheses. The posterior probabilities of the hypotheses might then be used to guide a *decision* together with an appropriately specified loss function in the context of Bayesian decision theory (see e.g. Berger, 1995; Huntley et al., 2014).

In that, the changing focus onto Bayesian statistics within psychological research might be seen as a step towards rising awareness of the distinction between evidence, belief and decision in the context of an analysis of statistical hypotheses (see e.g. Lavine and Schervish (1999) and especially Royall (2004)).

Naturally, statistical hypotheses depend on the real-world research question, which might not always be unambiguously formalized mathematically. Prior probabilities of the hypotheses are subjective in nature and only rarely accessible as precise numerical values. The loss function depends on a putative real-world decision problem such that a precise specification of the loss function might not be given by the researcher.

Yet, certain potentially incomplete information about hypotheses, prior beliefs and the loss function might be available, such that both ignoring these information or specifying the respective quantities in an overly precise way might yield misleading results or decisions. In order to avoid untrustworthy results, it is thus necessary to allow researchers to include information into a statistical analysis specifically in the imprecise form it is available. Therefore, this paper intends to formulate the simplest case (using simple hypotheses) of hypothesis-based Bayesian decision theory in a way to include partial information about hypotheses, prior beliefs and the loss function. This might be seen as a fundamental, but necessary step to extend the imprecise probability framework (Walley, 1991) to the Bayes factor analyses that are recently applied in psychological research, working at the interface between statistical developments and empirical sciences.

As mentioned above, there are two different types of prior distributions inherent to a Bayes factor analysis:

(hypothesis-based) priors on the parameters and a prior on the hypotheses that is used to further guide decisions. In that, the Bayes factor analysis might be generalized within the framework of imprecise probabilities at these two distinct parts. Allowing the priors on the parameter to be specified imprecisely is restricted to the Bayes factor analysis itself and given an account by Ebner et al. (2019). However, considerations about allowing imprecise priors on the hypotheses and imprecise quantities relevant for a corresponding decision might apply to more situations than typically addressed in a Bayes factor analysis. Therefore, it will be given a separate account within this paper and discussed by referring to the likelihood ratio, which might be seen as both special case (using degenerate priors on the parameters) and foundational basis (see e.g. Royall, 2004) of the Bayes factor.

The present paper is structured as follows. Section 2 collects the basic ingredients of the classical case of Bayesian decision making based on two precise simple hypotheses. This framework will in Section 3 be powerfully extended to the situation where single hypotheses are interval-valued and the loss functions and prior odds are imprecise. Section 4 warns that, however, in this context some inconsistency issues may arise under updating and assess them in greater detail. Section 5 provides a numerical example as illustration and Section 6 concludes with a brief outlook.

2. Precise Hypothesis-Based Bayesian Decision Making

Assume a parametric statistical model, such that observed data $x = (x_1, \ldots, x_n)$ are modeled as realizations of independent and identically distributed random variables X_i, $i = 1, \ldots, n$, with parametric probability density $f(x_i|\theta)$, $\theta \in \mathscr{D}_\theta$, which specifies the joint density as

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)\,. \tag{1}$$

All considered parameter vales θ are comprised within the parameter space $\mathscr{D}_\theta$ and, for the sake of simplicity (especially w.r.t. notation), the parameter is assumed to be a single real-valued scalar here. Generalizations to multidimensional parameters are possible, but are left to further research.

Further assume two precise simple hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1\,, \tag{2}$$

where θ_0 and θ_1 are precise hypothesized parameter values, which implies that one of these two values is considered to be true. In a Bayesian context there is a subjective prior distribution on the hypotheses ($p(H_0)$ and $p(H_1) = 1 - p(H_0)$), forming the prior odds

$$\pi := \frac{p(H_0)}{p(H_1)}\,. \tag{3}$$

The prior odds can be updated by the observed data x via Bayes rule to the posterior odds

$$\frac{p(H_0|x)}{p(H_1|x)} = \frac{\frac{f(x|\theta_0)\cdot p(H_0)}{f(x)}}{\frac{f(x|\theta_1)\cdot p(H_1)}{f(x)}} = LR^x(\theta_0, \theta_1)\cdot\pi\,, \tag{4}$$

where

$$LR^x(\theta_0, \theta_1) = \frac{f(x|\theta_0)}{f(x|\theta_1)} \tag{5}$$

is the likelihood ratio and frequently referred to as Bayes factor (see e.g. Liu and Aitkin, 2008), as both hypotheses in equation (2) might be formulated by degenerate probability distributions with all probability mass on θ_0 and θ_1, respectively.

In order to guide a decision between two actions a_0 and a_1, a loss function

$$\begin{aligned} L : \mathscr{H} \times \mathscr{A} &\to \mathbb{R}_0^+ \\ (H, a) &\mapsto L(H, a) \end{aligned} \tag{6}$$

with $\mathscr{H} = \{H_0, H_1\}$ and $\mathscr{A} = \{a_0, a_1\}$ need to be specified, quantifying the "badness" of choosing a if H is true. The expected posterior loss

$$\begin{aligned} \rho : \mathscr{A} &\to \mathbb{R}_0^+ \\ a &\mapsto p(H_0|x)L(H_0, a) + p(H_1|x)L(H_1, a) \end{aligned} \tag{7}$$

can be used to find the optimal action(s)

$$a^* = \operatorname*{argmin}_{a\in\mathscr{A}} \rho(a)\,. \tag{8}$$

Assume that, as is common practice in empirical research, the decision problem is formulated in regret form, where a_0 is associated with H_0 and a_1 with H_1 such that the correct decisions are evaluated to have zero loss, i.e. $L(H_0, a_0) = L(H_1, a_1) = 0$. Then it is only necessary to specify the ratio

$$k := \frac{L(H_0, a_1)}{L(H_1, a_0)} \tag{9}$$

in order to calculate the ratio of expected posterior losses

$$r := \frac{\rho(a_1)}{\rho(a_0)} = \pi \cdot LR^x(\theta_0, \theta_1)\cdot k \tag{10}$$

to determine

$$a^* = \begin{cases} a_0 & \text{if } r > 1 \\ a_1 & \text{if } r < 1 \end{cases}\,. \tag{11}$$

For $r = 1$ any action might be chosen.

3. Imprecise Hypothesis-Based Bayesian Decision Making

Within applied research, it is typically extremely difficult to specify the quantities θ_0, θ_1, π and k, which are necessary to determine a^*, as precise values. This is due to the fact that commonly some (potentially imprecise) information is available and several choices of precise values for these quantities are in accordance with it. Both ignoring the available relevant information and arbitrarily choosing among those plausible values, can hardly be an optimal strategy. Therefore, these quantities shall be specified imprecisely as an interval of values. Following Dubois' distinction (cp. Dubois, 1986, Section 1.4), these intervals have to be interpreted as conjunctive sets: they must be treated as a generalization of a single value and thus as an entity of its own. In that, as the interval of values replaces the respective precise value, the distribution is parametrically constructed by an interval (e.g. Augustin et al., 2014, Section 7.3.2). Also note that all four quantities θ_0, θ_1, π and k might be specifiable independently of each other, which allows subsequent calculations to be straightforward.

3.1. Imprecise Simple Hypotheses

Instead of a precise parameter value θ, the (imprecise) density of the data x is now dependent on an imprecise interval-valued parameter $\Theta = [\underline{\Theta}, \overline{\Theta}]$, i.e.

$$f(x|\Theta) = \{f(x|\theta)|\theta \in \Theta\} \tag{12}$$

with $\underline{\Theta}$ and $\overline{\Theta}$ being precise valued bounds on the parameter that are considered as defining the imprecise parameter Θ. Accordingly, the parameter space of Θ is now the set of all closed parameter intervals

$$\mathscr{D}_\Theta = \{[\underline{\Theta}, \overline{\Theta}]|\underline{\Theta} \in \mathscr{D}_\theta, \overline{\Theta} \in \mathscr{D}_\theta, \underline{\Theta} \leq \overline{\Theta}\}\,. \tag{13}$$

Consider imprecise, but simple hypotheses

$$H_0 : \Theta = \Theta_0 \quad \text{vs.} \quad H_1 : \Theta = \Theta_1\,, \tag{14}$$

where

$$\Theta_0 = [\underline{\Theta_0}, \overline{\Theta_0}]\,, \tag{15}$$

$$\Theta_1 = [\underline{\Theta_1}, \overline{\Theta_1}] \tag{16}$$

and $\underline{\Theta_0}$ (or $\underline{\Theta_1}$) is the lower bound as well as $\overline{\Theta_0}$ (or $\overline{\Theta_1}$) the upper bound for the simple hypothesized parameter value Θ under H_0 (or H_1). Although specified as intervals within this paper, simple imprecise hypotheses might also be generalized to hypothesize (convex) sets of parameters in general. Note that these hypotheses are not composite, as they consist only of one single, but imprecisely specified value. In contrast, composite hypotheses, for instance specified as

$$H_0 : \theta \in [\underline{\Theta_0}, \overline{\Theta_0}] \quad \text{vs.} \quad H_1 : \theta \in [\underline{\Theta_1}, \overline{\Theta_1}]\,, \tag{17}$$

would contain all precise parameter values within the respective intervals. That is exactly the crucial difference in interpreting composite and simple imprecise hypotheses. While the latter states that there is only one single parameter value which represents the hypothesis, yet there is not enough information available to precisely specify this single value, the former states that all the different parameter values, as a whole, represent the hypothesis. In that, composite hypotheses bound the unknown parameter value of a precise sampling model, while an imprecise parameter specifies an imprecise sampling model (e.g. Augustin et al., 2014, Section 7.2.5). As an outlook, composite imprecise hypotheses would be subsets of $\mathscr{D}_\Theta$ containing more than one parameter interval.

The Bayesian account to composite hypotheses is to employ a prior distribution on the hypothesized values and to calculate the respective marginal density of the observed data (as in a typical Bayes factor analysis (e.g. Morey et al., 2016)). While this prior is on the *parameter values* themselves within a precise composite hypothesis, it is on *parameter intervals* within an imprecise composite hypothesis. A simple imprecise hypothesis might therefore be described by a degenerate distribution with all mass on the respective parameter interval.

Accordingly, the fundamental technical difference between precise composite and simple imprecise hypotheses within the Bayesian framework is that only former requires the specification of a prior distribution on the hypothesized parameter values. In that, former might be incorporated within the Bayesian analysis by means of a marginal density and latter by means of the imprecise-valued density as in equation (12).

3.2. Imprecise Likelihood Ratio, Imprecise Prior Odds, and Imprecise Loss Function

Given data x, instead of a precise likelihood ratio, there is an interval-valued likelihood ratio

$$LR^x = [\underline{LR^x}, \overline{LR^x}]\,, \tag{18}$$

with

$$\underline{LR^x} = \min_{\substack{\theta_0 \in \Theta_0 \\ \theta_1 \in \Theta_1}} LR^x(\theta_0, \theta_1)\,, \tag{19}$$

$$\overline{LR^x} = \max_{\substack{\theta_0 \in \Theta_0 \\ \theta_1 \in \Theta_1}} LR^x(\theta_0, \theta_1)\,. \tag{20}$$

Note that within this paper a precise likelihood ratio value is denoted with its dependence on θ_0 and θ_1, whereas a interval-valued likelihood ratio is denoted without this dependence.

In addition, the prior odds

$$[\underline{\pi}, \overline{\pi}] \tag{21}$$

might be interval-valued with $\underline{\pi}$ being the lower bound and $\overline{\pi}$ being the upper bound of the subjectively specified prior odds, leading to the imprecisely defined posterior odds

$$[\underline{LR^x} \cdot \underline{\pi}\,, \overline{LR^x} \cdot \overline{\pi}]\,. \tag{22}$$

The loss function might also be specified imprecisely by

$$[\underline{k}, \overline{k}]\,, \tag{23}$$

where, analogously, $\underline{k}$ is the lower bound and $\overline{k}$ is the upper bound for stating, in generalization of (9), how much "worse" a_1 would be under H_0 than a_0 would be under H_1, if deciding correctly has 0 "badness" (for a more general account on robust loss functions see Dey and Michaes (2000)).

In contrast to the precise case, the ratio of expected posterior losses r, which was used to determine the optimal action, is not precise anymore:

$$[\underline{r}, \overline{r}]\,, \tag{24}$$

where

$$\underline{r} = \underline{\pi} \cdot \underline{LR^x} \cdot \underline{k} \tag{25}$$

$$\overline{r} = \overline{\pi} \cdot \overline{LR^x} \cdot \overline{k} \tag{26}$$

can be calculated from the respective lower and upper bounds of π, LR^x and k, as all these quantities are positive, and they vary independently. If one of these quantities is still precise, its lower and upper bounds are equal, for instance for a precise k it holds that $k = \underline{k} = \overline{k}$.

The optimal action is

$$a^* = \begin{cases} a_0 & \text{if } \underline{r} \geq 1 \\ a_1 & \text{if } \overline{r} \leq 1 \end{cases}\,, \tag{27}$$

however, for $\underline{r} < 1 < \overline{r}$, the decision cannot be guided unambiguously and more information is required. This might be accomplished by collecting more data, such that the imprecise likelihood ratio interval will become smaller, or by obtaining more information about the decision problem, such that θ_0, θ_1, π or k might be specified more accurately, i.e. by smaller intervals. With this additional information, the resulting imprecise ratio of expected posterior losses $[\underline{r}, \overline{r}]$ might become smaller and with sufficient information might exclude 1, allowing the determination of the optimal action a^*. This will be illustrated by an example in Section 5.

Certainly, not being able to determine an optimal action in the context of a given data set might at first glance seem to be a disadvantage of the imprecise framework. However, this might only occur if some of the available information is imprecise, such that specifying precise values for the necessary quantities is arbitrary, can be characterized as overprecision and might yield potentially misleading, enforced decisions. Nevertheless, if necessary, enforcing a decision is still possible for $\underline{r} < 1 < \overline{r}$, yet the researcher is now aware of its spuriousness, which might have been masked due to the overprecision within the precise case.

4. Potential Bayesian Updating Issues with Imprecise Hypotheses

Although within the last section simple hypotheses were allowed to be imprecisely specified, this might be accompanied by Bayesian updating inconsistencies that appear while sequentially considering two separate data sets. On that note, (e.g. Seidenfeld, 1994; Huntley et al., 2014) already emphasized the importance of being cautions with sequential decision problems in the context of imprecise probabilities.

4.1. Precise Case

Consider the presence of a second data set $y = (y_1, \ldots, y_m)$ being modeled analogously to x, i.e.

$$f(y|\theta) = \prod_{i=1}^{m} f(y_i|\theta)\,, \tag{28}$$

and denote $z = (x, y)$ as the merged data set with

$$f(z|\theta) = \prod_{i=1}^{n+m} f(z_i|\theta) = f(y|\theta) \cdot f(x|\theta)\,. \tag{29}$$

Therefore, with precise simple hypotheses as in equation (2) it holds that

$$LR^z(\theta_0, \theta_1) = LR^y(\theta_0, \theta_1) \cdot LR^x(\theta_0, \theta_1) \tag{30}$$

and the posterior odds after seeing all the data z

$$LR^z(\theta_0, \theta_1) \cdot \pi = LR^y(\theta_0, \theta_1) \cdot LR^x(\theta_0, \theta_1) \cdot \pi \tag{31}$$

(as well as the ratio of expected posterior losses r) do not depend on whether the data was merged or not.

4.2. Imprecise Case

However, in the context of the imprecise hypotheses from equation (14), define

$$(\underline{\theta_0^x}, \underline{\theta_1^x}) := \underset{\substack{(\theta_0,\theta_1):\\ \theta_0 \in \Theta_0, \theta_1 \in \Theta_1}}{\operatorname{argmin}} \; LR^x(\theta_0, \theta_1)\,, \tag{32}$$

$$(\underline{\theta_0^y}, \underline{\theta_1^y}) := \underset{\substack{(\theta_0,\theta_1):\\ \theta_0 \in \Theta_0, \theta_1 \in \Theta_1}}{\operatorname{argmin}} \; LR^y(\theta_0, \theta_1)\,, \tag{33}$$

$$(\underline{\theta_0^z}, \underline{\theta_1^z}) := \underset{\substack{(\theta_0,\theta_1):\\ \theta_0 \in \Theta_0, \theta_1 \in \Theta_1}}{\operatorname{argmin}} \; LR^z(\theta_0, \theta_1) \tag{34}$$

as the respective tuples of hypothesized parameter values, which lead for each data set to the respective minimal likelihood ratio. As in general

$$\underline{\theta_0^x} \neq \underline{\theta_0^y} \neq \underline{\theta_0^z}\,, \tag{35}$$

$$\underline{\theta_1^x} \neq \underline{\theta_1^y} \neq \underline{\theta_1^z}\,, \tag{36}$$

it follows that

$$f(z|\underline{\theta_0^z}) \neq f(y|\underline{\theta_0^y}) \cdot f(x|\underline{\theta_0^x})\,, \tag{37}$$

$$f(z|\underline{\theta_1^z}) \neq f(y|\underline{\theta_1^y}) \cdot f(x|\underline{\theta_1^x}) \tag{38}$$

and accordingly

$$\underline{LR^z} \neq \underline{LR^y} \cdot \underline{LR^x}\,. \tag{39}$$

Analogue considerations lead to

$$\overline{LR^z} \neq \overline{LR^y} \cdot \overline{LR^x}\,, \tag{40}$$

and an example of this inequality is provided within Section 5.

Therefore, in general, the imprecise posterior odds after considering the merged data

$$[\underline{LR^z} \cdot \underline{\pi}\,, \overline{LR^z} \cdot \overline{\pi}] \neq [\underline{LR^y} \cdot \underline{LR^x} \cdot \underline{\pi}\,, \overline{LR^y} \cdot \overline{LR^x} \cdot \overline{\pi}] \tag{41}$$

differ from those after subsequently considering both data sets separately, which treats the posterior odds after the first data set x as prior odds for the second data set y.

Accordingly, it might seem that the imprecise ratio of expected posterior losses and the resulting decision might depend on whether the data was merged or not. In that, the Bayesian updating procedure for the odds on the hypotheses might be characterized as 'inconsistent' in terms of Rüger (1998, p. 190)'s work on the foundations of statistics.

4.3. Evaluation

Evaluating these updating inconsistencies in greater detail, two characteristics emerge.

First, although the interval-valued likelihood ratio LR^x of the data set x might be outlined by its bounds $\underline{LR^x}$ and $\overline{LR^x}$, consistent updating dictates to also consider the dependence of the likelihood ratio values within LR^x on the parameter values θ_0 and θ_1 as the result of the analysis.

This can be seen based on the following considerations. The interval-valued likelihood ratio LR^x of equation (18) consists of all likelihood ratio values obtained with parameters $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, i.e.

$$LR^x = \{LR^x(\theta_0, \theta_1) | \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\}\,. \tag{42}$$

In this regard, the values within the interval-valued likelihood ratio of the merged data z might be decomposed using equation (30) to

$$\begin{aligned} LR^z &= \{LR^z(\theta_0, \theta_1) | \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\} && (43)\\ &= \{LR^y(\theta_0, \theta_1) \cdot LR^x(\theta_0, \theta_1) | \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\}\,. && (44) \end{aligned}$$

It appears that for each value within LR^z the complete data set has to be evaluated using the same parameter values θ_0 and θ_1. However, for calculating e.g. $\underline{LR^y} \cdot \underline{LR^x}$, the first part of the data x was evaluated with different parameter values ($\underline{\theta_0^x}$ and $\underline{\theta_1^x}$) than the second part of the data y (evaluated with $\underline{\theta_0^y}$ and $\underline{\theta_1^y}$). Accordingly, the value $\underline{LR^y} \cdot \underline{LR^x}$ might not be contained within LR^z and updating might be inconsistent.

To enable consistent updating, from the first analysis of data set x, all values within the interval-valued likelihood ratio LR^x together with their dependence on θ_0 and θ_1, not only the bounds $\underline{LR^x}$ and $\overline{LR^x}$, are necessary to calculate the final interval-valued likelihood ratio LR^z in a subsequent analysis of both data sets x and y using equation (44).

Second, the values $\underline{LR^y} \cdot \underline{LR^x}$ and $\overline{LR^y} \cdot \overline{LR^x}$ might be considered as approximation of the interval LR^z by providing outer bounds, i.e.

$$LR^z = [\underline{LR^z}, \overline{LR^z}] \subseteq [\underline{LR^y} \cdot \underline{LR^x}, \overline{LR^y} \cdot \overline{LR^x}]\,. \tag{45}$$

This becomes apparent by considering the lower bound $\underline{LR^z}$, which is obtained with parameter values $\underline{\theta_0^z} \in \Theta_0$, $\underline{\theta_1^z} \in \Theta_1$. Applying equation (30) leads to

$$\underline{LR^z} = LR^z(\underline{\theta_0^z}, \underline{\theta_1^z}) = LR^y(\underline{\theta_0^z}, \underline{\theta_1^z}) \cdot LR^x(\underline{\theta_0^z}, \underline{\theta_1^z}) \tag{46}$$

and as $\underline{LR^x}$ and $\underline{LR^y}$ are minima, it also holds that

$$\underline{LR^x} \leq LR^x(\underline{\theta_0^z}, \underline{\theta_1^z}) \tag{47}$$

$$\underline{LR^y} \leq LR^y(\underline{\theta_0^z}, \underline{\theta_1^z})\,, \tag{48}$$

so that together (as all likelihood ratios are positive)

$$\underline{LR^y} \cdot \underline{LR^y} \leq LR^y(\underline{\theta_0^z}, \underline{\theta_1^z}) \cdot LR^x(\underline{\theta_0^z}, \underline{\theta_1^z}) = \underline{LR^z}\,. \tag{49}$$

Analogue considerations lead to

$$\overline{LR^y} \cdot \overline{LR^y} \geq \overline{LR^z}\,, \tag{50}$$

finally allowing the approximation in equation (45).

5. Example

A short fictitious example shall serve as illustration (replicable with the R code in the electronic appendix).

Person A provides a huge amount of allegedly fair coins and offers a bet to person B for 1€: Person A will randomly take one of the coins and flip it. If tails, then person B will get back 4€. Naturally, person B is suspicious about the coins being fair and eventually obtains the permission to examine some coins. Based on the outcome of that sample, person B will have to decide whether to accuse person A of cheating (action a_1) or not (action a_0).

Modelling the coin flips as independent Bernoulli experiments with parameter p for the probability of heads, person B considers the possibility of the coins being fair with the precise null hypothesis $H_0 : p = 0.5$. However, person B is unsure about the parameter p if person A is cheating. Due to the offer of person A, p might be at least 0.75, but on the other hand, if p might be too high, say $p > 0.9$, it might be too suspicious. Person B regards those parameter values $[0.75, 0.9]$ as plausible, but is not able to further describe the plausibility of each of these parameter value. Furthermore, person B considers the possibility that different coins

might have (slightly) different probabilities of heads and, therefore, chooses as alternative hypothesis the imprecise simple hypothesis $H_1 : p = [0.75, 0.9]$.

The loss $L(H_1, a_0)$ of not doing anything if the coins are truly biased is not too high, as the price of the bet is only 1€. Accusing person A of cheating if the coins are actually fair ($L(H_0, a_1)$), however, might result in a rather unpleasant situation. Naturally, both these losses are on a different scale, but need to be expressed in relation to each other. As this is rather difficult, Person B figures out that k might be somewhere between 8 and 20, being unable to further specify this value.

In a situation before checking the coins, person B is also not exactly sure what to belief about the coins. Certainly, with the offer of person A, the alternative hypothesis is at least as plausible as the null hypothesis. However, the coins look normal and so the null hypothesis is not absolutely implausible. After some consideration, person B determines that the prior odds are captured by $\pi = [1, 4]$.

Now, person B flips $n = 10$ coins, yielding heads $x = 9$ times. Based on this observation and the specifications given, person B calculates the interval-valued likelihood ratio

$$LR^x = [0.025, 0.052] \tag{51}$$

and the ratio of expected posterior losses

$$[0.202, 4.162], \tag{52}$$

which does not unambiguously favor one of the actions, as it contains the value 1.

Additional information is necessary to do so and person B flips another $m = 10$ coins, yielding heads $y = 5$ times. The corresponding interval-valued likelihood ratio is

$$LR^y = [4.214, 165.4]. \tag{53}$$

Combining those interval-valued likelihood ratios yields

$$[\underline{LR^y} \cdot \underline{LR^y}, \overline{LR^y} \cdot \overline{LR^y}] = [0.105, 8.601], \tag{54}$$

but knowing of the updating inconsistencies, person B treats this interval only as an approximation, resulting in an approximation of the ratio of expected posterior losses by

$$[0.843, 688.1]. \tag{55}$$

Still the value 1 is included within the interval and this approximation does not allow an unambiguous decision.

In order to account for the updating inconsistencies, person B merges both data sets $z = 9 + 5 = 14$ with $n + m = 20$, leading to the interval-valued likelihood ratio

$$LR^z = [0.219, 4.169], \tag{56}$$

which is truly different to and included by the interval in equation (54). The resulting ratio of expected posterior losses is

$$[1.754, 333.5], \tag{57}$$

which finally favors to not accuse person A of cheating (action a_0).

By providing the data (n, m, x and y), sufficient information is available for subsequent analyses to consider the dependency of respective likelihood ratio values on the parameter θ_1.

Person B specified the relevant quantities as best as possible to the partially available knowledge and the analysis of the first data set indicated a lack of information for guiding the decision. A precise account of the situation, on the other hand, might have pretended a precision, which is not available. For example, person B might have arbitrarily chosen $H_1 : p = 0.8$, $k = 8$ and $\pi = 1$ of those possible values that are in accordance with the available knowledge, leading to a precise likelihood ratio of $LR^x(0.5, 0.8) = 0.036$ and a ratio of expected posterior losses of $r = 0.29$ that favoured a_1.

Accordingly, person B would have accused person A of cheating, although the available information are rather ambiguous. Even worse, person B would not even be aware of the lack of information, as it was masked by the false precision of the arbitrarily chosen values.

6. Concluding Remarks

This paper elaborated on how to include partial information about simple hypotheses, prior beliefs and the loss function in the context of hypothesis-based Bayesian decision theory and depicted inconsistencies within the procedure of Bayesian updating that might arise from the use of imprecise simple hypotheses.

Typically, there is only one data set for the statistical analysis of an empirical study, so that the updating inconsistencies as depicted in Section 4 might not become visible. Furthermore, for guiding the decision based on a single data sets, within the context employed in this paper, only the bounds of the interval-valued likelihood ratio are necessary. Nevertheless, properly reporting the results of the analysis also requires to include the dependence of the likelihood ratio values within the interval-valued likelihood ratio on the parameter values. Naturally, as an alternative, the data can be made publicly accessible, so that all relevant information necessary for subsequent analyses might be extracted directly from the data.

Although two data sets were considered to outline the updating inconsistencies, this cannot be regarded as an unnatural approach, as one of the central characteristics of Bayesian learning is to employ a posterior distribution obtained from previous data as prior distribution for a subsequent analysis. Certainly, this reflects the natural way to accumulate information.

In addition, remark that within this paper, the (imprecise) prior odds are updated first to obtain the posterior odds before determining a potentially optimal decision. How-

ever, a different procedure might be possible as well. For each hypothesis a decision strategy might be calculated, which maps the potentially observed data to the optimal action. In that, a decision strategy might be chosen first based on the prior odds and then the optimal decision might be determined based on the observed data. While this equivalence of prior risk optimality and posterior loss optimality holds in the traditional case of precise probabilities and loss functions, it is no longer satisfied in more general settings (see explicitly Augustin (2003) and more generally the references in Section 4).

Sometimes, an applied researcher is not primarily interested in guiding a decision, but just in investigating a real-world phenomenon. In this case, a hypothesis-based statistical analysis might be superfluous and descriptive statistics seem to be sufficient (see also the literature about "new statistics", e.g. Cumming, 2014). Nevertheless, all information should be provided, such that other researchers are able to guide a decision.

Furthermore, the only quantities treated imprecisely within this paper were the hypotheses, the prior odds and the loss function, however, also the data themselves might be available imprecisely, representing ambiguity in the data values. Although, most commonly, data values in psychological research represent scores that are designed to be precise, extending this framework to allow imprecise data looks very promising, as the data are independent of the other imprecisely specified quantities.

In summary, this paper addressed the imprecise generalization of hypothesis-based Bayesian decision making using simple hypothesis and, therefore, employed the likelihood ratio. A Bayes factor analysis typically employs composite hypotheses as well and might therefore be considered as more complex than the context depicted here. Yet, even within this simple context updating inconsistencies might occur, emphasizing the importance of investigating them in greater detail particularly with regard to their presence in analyses using Bayes factors.

Appendix A. R Code

R code to replicate the example is provided electronically.

Acknowledgments

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

References

Thomas Augustin. On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view — a cautionary note on updating imprecise priors. In Jean-Marc Bernard, Teddy Seidenfeld, and Marco Zaffalon, editors, *ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, pages 31–45, Lugano, Waterloo, 2003. Carleton Scientific.

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

James O. Berger. *Statistical decision theory and Bayesian analysis. 2nd edition*. Springer, 1995.

Geoff Cumming. The new statistics: Why and how. *Psychological Science*, 25(1):7–29, 2014.

Dipak K. Dey and Athanasios C. Michaes. Ranges of posterior expected losses and ε-robust actions. In David Ríos Insua and Fabrizio Ruggeri, editors, *Robust Bayesian Analysis*, pages 145–159. Springer, 2000.

Didier Dubois. Belief structure, possibility theory and decomposable confidence measures on finite sets. *Computers and Artificial Intelligence*, 5:403–416, 1986.

Luisa Ebner, Patrick Schwaferts, and Thomas Augustin. Robust Bayes factor for independent two-sample comparisons under imprecise prior information. conditionally accepted subject to minor revision for: Jasper de Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler, editors, *Proceedings of the 11th International Symposium on Imprecise Probability: Theory and Applications (ISIPTA '19, Ghent), Proceedings in Machine Learning Research*, 2019.

Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall. The Bayesian two-sample t test. *The American Statistician*, 59(3):252–257, 2005.

Nathan Huntley, Robert Hable, and Matthias C. M. Troffaes. Decision making. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 190–206. John Wiley & Sons, 2014.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Michael Lavine and Mark J. Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119–122, 1999.

Charles C. Liu and Murray Aitkin. Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52(6):362–375, 2008.

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

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009.

Richard Royall. The likelihood paradigm for statistical evidence. In Mark L. Taper and Subhash R. Lele, editors, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pages 119–152. University of Chicago Press, 2004.

Bernhard Rüger. *Test-und Schätztheorie: Band I: Grundlagen*. De Gruyter Oldenbourg, 1998.

Teddy Seidenfeld. When normal form and extensive form solutions differ. In Dag Prawitz, Brian Skyrms, and Dag Westerstahl, editors, *Logic, Methodology and Philosophy of Science IX (Uppsala, 1991)*, pages 451–463. Elsevier, 1994.

Rens Van De Schoot, Sonja D. Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, and Sarah Depaoli. A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2):217–239, 2017.

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

Contribution 6

Schwaferts & Augustin (2021a): Imprecise Hypothesis-Based Bayesian Decision Making with Composite Hypotheses. (ISIPTA)

ISIPTA 2021

Imprecise Hypothesis-Based Bayesian Decision Making with Composite Hypotheses

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

Abstract

Statistical analyses with composite hypotheses are omnipresent in empirical sciences, and a decision-theoretic account is required in order to formally consider their practical relevance. A Bayesian hypothesis-based decision-theoretic analysis requires the specification of a prior distribution, the hypotheses, and a loss function, and determines the optimal decision by minimizing the expected posterior loss of each hypothesis. However, specifying such a decision problem unambiguously is rather difficult as, typically, the relevant information is available only partially. In order to include such incomplete information into the analysis and to facilitate the use of decision-theoretic approaches in applied sciences, this paper extends the framework of hypothesis-based Bayesian decision making with composite hypotheses into the framework of imprecise probabilities, such that imprecise specifications for the prior distribution, for the composite hypotheses, and for the loss function are allowed. Imprecisely specified composite hypotheses are sets of parameter sets that are able to incorporate blurring borders between hypotheses into the analysis. The imprecisely specified prior distribution gets updated via generalized Bayes rule, such that imprecise probabilities of the (imprecise) hypotheses can be calculated. These lead – together with the (imprecise) loss function – to a set-valued expected posterior loss for finding the optimal decision. Beneficially, the result will also indicate whether or not the available information is sufficient to guide the decision unambiguously, without pretending a level of precision that is not available.

Keywords: Decision Theory, Bayesian Statistics, Composite Hypotheses, Imprecise Probabilities

1. Introduction

There is an increased awareness within the applied sciences that it is important to consider the practical relevance of an effect in addition to its statistical significance (see e.g. Kirk, 1996). Implemented in hypothesis-based methodologies, this relates to specifying the hypotheses reasonably w.r.t. their practical implications (see e.g. methods using regions of practical equivalence (Kruschke, 2015, 2018) or equivalence tests (Lakens, 2017; Lakens et al., 2018)). Practical implications naturally depend on what results are used for. Aptly put by Berger and Wolpert (1988, p. 55): "But no matter what is meant by inference, if it is to be of any value, then somehow it must be used, or acted upon, and this does indeed lead back to the decision-theoretic framework." However, a decision-theoretic analysis (see e.g. Berger, 1985) is typically avoided in applied sciences (cp. e.g. the recommendation in Rouder et al., 2018, p. 110). This might be explained by the fact that many required quantities are very difficult to specify unambiguously, as relevant information is typically available only partially.

In order to facilitate such a decision-theoretic analysis, this paper intends to extend hypothesis-based analyses into a decision-theoretic framework that allows for impartial information to be included properly, building on the framework of imprecise probabilities (see e.g. Augustin et al., 2014a; Walley, 1991). Therefore, previous elaborations on imprecise hypothesis-based Bayesian decision making (Schwaferts and Augustin, 2019) shall be extended to composite hypotheses.

The paper starts by presenting the precise framework of hypothesis-based Bayesian decision making with composite hypotheses in Section 2, which serves as basis for its extension to the framework of imprecise probabilities in Section 3 by allowing imprecisely specified prior distributions (Section 3.1), composite hypotheses (Section 3.2), and loss functions (Section 3.3). A schematic example is provided in Section 4, followed by a discussion about scalability (Section 5.1) and the conditional perspective (Section 5.2), as well as by a brief outlook in Section 6.

2. Precise Hypothesis-Based Bayesian Decision Making

Within the context of decision making (for an extensive overview see Berger, 1985), the observed data x are commonly assumed to be parametrically distributed with density $f(x|\theta)$ and parameter $\theta \in \Theta$. Although generalizations to multidimensional parameters are possible, the parameter

θ is assumed to be a single real-valued scalar within this paper to keep the notation simple.

In the Bayesian setting, there is a prior distribution π_θ on the parameter θ with density $\pi(\theta)$. In the presence of the observed data x, this prior distribution gets updated via Bayes rule to the posterior distribution $\pi_{\theta|x}$ with density

$$\pi(\theta|x) = \frac{f(x|\theta)\cdot\pi(\theta)}{f(x)}, \tag{1}$$

where

$$f(x) = \int_\Theta f(x|\theta)\cdot\pi(\theta)\,d\theta \tag{2}$$

is the marginal density of the data x, assumed to be strictly positive for all x.

Within a Bayesian analysis, results are derived from this posterior distribution $\pi_{\theta|x}$ exclusively, e.g. by considering mean, median, or certain credibility intervals (see e.g. Kruschke, 2015). If a research question contrasts different theoretical positions, these need to be formalized as statistical hypotheses, which are then evaluated using the posterior distribution. Formally, composite hypotheses

$$h_0 : \theta \in \Theta_0 \quad \text{vs.} \quad h_1 : \theta \in \Theta_1 \tag{3}$$

are subsets $\Theta_0, \Theta_1 \subset \Theta$ of the parameter space.

For the given prior distribution on θ and the observed data x, the posterior probabilities of the hypotheses are

$$p(h_0|x) := p(\Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)\,d\theta, \tag{4}$$

$$p(h_1|x) := p(\Theta_1|x) = \int_{\Theta_1} \pi(\theta|x)\,d\theta, \tag{5}$$

where we assume non-degenerated cases with $p(h_0|x) > 0$ and $p(h_1|x) > 0$.

Frequently, contrasting statistical hypotheses (and the corresponding theoretical positions) is related to an applied research question or some practical implications, being formalized by a decision problem. Consider the case of a decision between two actions a_0 and a_1 (as only two hypotheses are considered within this paper). A loss function

$$L : \mathscr{H} \times \mathscr{A} \to \mathbb{R}_0^+ : (h,a) \mapsto L(h,a), \tag{6}$$

with $\mathscr{H} = \{h_0, h_1\}$ being the hypothesis space and $\mathscr{A} = \{a_0, a_1\}$ being the action space, quantifies the "badness" of deciding for $a \in \mathscr{A}$ if $h \in \mathscr{H}$ is true.

Typically, deciding for a_1 if h_1 is true and for a_0 if h_0 is true is considered to be a correct decision, such that – without loss of generality – the loss function can be stated in regret form, in which deciding correctly has zero loss, i.e. $L(h_1,a_1) = L(h_0,a_0) = 0$. The remaining values refer to the type-I-error ($L(h_0,a_1)$) and the type-II-error ($L(h_1,a_0)$) and are assumed to be non-zero. In that, it is possible here to specify the loss function L by one single quantity

$$k := \frac{L(h_0,a_1)}{L(h_1,a_0)}, \tag{7}$$

which specifies how bad the type-I-error is compared to type-II-error (if deciding correctly has zero loss).

As the hypotheses (equation (3)) represent sets of parameters, the loss function L – which was defined on the hypothesis space $\mathscr{H}$ and the action space $\mathscr{A}$ within this paper (equation (6)) – might also be depicted w.r.t. the parameter space Θ and the action space $\mathscr{A}$, formally

$$L_\theta : \Theta \times \mathscr{A} \to \mathbb{R}_0^+ : (\theta,a) \mapsto L_\theta(a). \tag{8}$$

An example of such a loss function in regret form w.r.t. the hypotheses is illustrated in the context of the example (Section 4) in Figure 1 (top).

With the expected posterior loss $\rho : \mathscr{A} \to \mathbb{R}_0^+$:

$$a \mapsto \rho(a) = L(h_1,a)\cdot p(h_1|x) + L(h_0,a)\cdot p(h_0|x), \tag{9}$$

the ratio of expected posterior losses

$$r := \frac{\rho(a_1)}{\rho(a_0)} = \frac{L(h_0,a_1)\cdot p(h_0|x)}{L(h_1,a_0)\cdot p(h_1|x)} \tag{10}$$

$$= k\cdot\frac{p(h_0|x)}{p(h_1|x)} \tag{11}$$

allows to determine the set $\mathscr{A}^*$ of optimal actions (in the context of the conditional Bayes decision principle Berger, 1985, p. 16)

$$\mathscr{A}^* = \begin{cases} \{a_0\} & \text{if } r < 1 \\ \{a_1\} & \text{if } r > 1 \\ \{a_0, a_1\} & \text{if } r = 1 \end{cases}. \tag{12}$$

Both actions are optimal, if $r = 1$. In this case, one might arbitrarily select one of the actions, as both actions have exactly the same expected posterior loss and can, therefore, be considered as practically equal.

3. Imprecise Hypothesis-Based Bayesian Decision Making

Within the framework of hypothesis-based Bayesian decision theory, imprecision shall be investigated for the prior distribution, the composite hypotheses, and the loss function, as an unambiguous precise specification of those quantities appears to bear the most difficulties for applied scientists (in contrast to precise data, likelihoods and parameters). Nevertheless, an extension towards imprecise data, imprecise likelihoods (see e.g. Walley, 1991, ch. 8) and imprecise parameters (for imprecise parameters in the context of simple hypotheses Schwaferts and Augustin, 2019) seems very powerful and shall be considered in future developments.

3.1. Imprecise Prior Distribution

A Bayesian analysis requires the specification of a prior distribution on the parameter. However, it is often impossible

to determine one single precise prior distribution describing adequately the extent and homogeneity of the knowledge at hand (e.g. Augustin et al., 2014b, Section 7.2). This impossibility is further illustrated by the fact that even the interpretation of a prior distribution is not unambiguously agreed on, with interpretations as knowledge (e.g. in Jaynes, 2003) or information (e.g. in Berger, 1985) or degrees of belief (e.g. in Jeffreys, 1961) or uncertainty (e.g. in Kruschke, 2015) about the parameter before observing the data. Instead, following the framework of imprecise probabilities (Walley, 1991), a set of prior distributions is considered to be more suitable. This set shall be denoted by Π_θ and referred to as imprecise prior distribution. It constitutes a quantity of its own and represents the prior situation. This imprecise prior gets updated after observing the data x to an imprecise posterior distribution

$$\Pi_{\theta|x} = \left\{ \pi_{\theta|x} \,\middle|\, \pi_\theta \in \Pi_\theta \right\}, \tag{13}$$

where each posterior distribution $\pi_{\theta|x}$ is obtained via Bayes rule (equation (1)) from one of the prior distributions $\pi_\theta \in \Pi_\theta$. Updating such a set of prior probabilities element by element is very natural in the context of Bayesian sensitivity analyses and so-called robust Bayesian approaches (e.g. Rios Insua and Ruggeri, 2000). Moreover, it can be justified by Walley's general coherence theory (Walley, 1991, Chapter 6ff.), where equation (13) is deduced from Walley's generalized Bayes rule.[1]

The resulting posterior distribution $\Pi_{\theta|x}$ represents the posterior situation (given the prior Π_θ and the data x) and underlies all further derivations in a Bayesian analysis.

3.2. Imprecise Composite Hypotheses

If two theoretical positions should be contrasted with each other, these need to be formalized as statistical hypotheses. However, determining which parameter values correspond to which theoretical position might not be unambiguous for all parameter values θ. To account for this, a hypothesis should comprise not only a single set of parameters (as in equation (3)) but a set of parameter sets. Denote these sets of parameter sets as

$$[\Theta]_0 := \{\Theta_0 \subset \Theta \,|\, \Theta_0 \text{ reasonable under } H_0\} \tag{14}$$
$$[\Theta]_1 := \{\Theta_1 \subset \Theta \,|\, \Theta_1 \text{ reasonable under } H_1\}, \tag{15}$$

where $[\Theta]_0$ contains all parameter sets Θ_0 that are reasonable for one hypothesis and $[\Theta]_1$ contains all parameter sets Θ_1 that are reasonable for the other hypothesis.

These sets are considered as entities on their own and formalize the theoretical positions that should be contrasted with each other, considering the available information as is.

1. See, in particular, the corresponding lower envelope theorem (Walley, 1991, Section 6.4.2) and (Walley, 1991, Section 7.8.1) on the coherence of envelopes of standard Bayesian inference.

Accordingly, the respective imprecisely specified hypotheses are

$$H_0 : \theta \in [\Theta]_0 \quad \text{vs.} \quad H_1 : \theta \in [\Theta]_1 . \tag{16}$$

This notation can be read as: The imprecise hypothesis H_0 states that the parameter θ is of a set Θ_0, which itself is only vaguely defined by $[\Theta]_0$ (H_1 analogously).

The crucial difference between precise hypotheses (equation (3)) and imprecise hypotheses (equation (16)) is that in the precise case, a certain parameter value θ might be assigned to *either* one hypothesis, the other hypothesis, both hypotheses (such that hypotheses are overlapping), or no hypothesis (such that this parameter value is not considered at all), while in the imprecise case, the assignment of a certain parameter value θ to the hypotheses might be *any* combination of these four options. Also note that the imprecise parameter sets $[\Theta]_0$ and $[\Theta]_1$ are different from

$$\bigcup_{\Theta_0 \in [\Theta]_0} \Theta_0 = \{\theta \in \Theta \,|\, \theta \in \Theta_0, \Theta_0 \in [\Theta]_0\} \tag{17}$$
$$\bigcup_{\Theta_1 \in [\Theta]_1} \Theta_1 = \{\theta \in \Theta \,|\, \theta \in \Theta_1, \Theta_1 \in [\Theta]_1\}, \tag{18}$$

which would represent two – most likely overlapping – precise hypotheses, and not two imprecisely specified hypotheses. Precise overlapping composite hypotheses imply that there is certainty that some parameter values $\theta \in \Theta$ are contained in both hypotheses, while the imprecise composite hypotheses state that there is uncertainty to which hypothesis some parameter values might be attributed. In that, latter hypotheses inherit far less requirements on the available information for their specification.

Although these formulations of imprecisely specified hypotheses might be employed in statistical analyses without being embedded into the decision theoretic context, this paper focuses on their use for guiding decisions. Then, imprecisely specified hypotheses might also be expressed by an imprecision in the parameter-based loss function L_θ (equation (8)), as illustrated in Figure 1 (center).

In order to obtain the (imprecise) posterior probabilities $P(H_0|x)$ and $P(H_1|x)$ of the imprecise hypotheses H_0 and H_1 using the imprecise posterior distribution $\Pi_{\theta|x}$, one might consider each combination of the distributions $\pi_{\theta|x} \in \Pi_{\theta|x}$ and the parameter sets $\Theta_0 \in [\Theta]_0$ or $\Theta_1 \in [\Theta]_1$, respectively, using equations (4) and (5):

$$P(H_0|x) = \left\{ p(h_0|x) \,\middle|\, \Theta_0 \in [\Theta]_0, \pi_{\theta|x} \in \Pi_{\theta|x} \right\} \tag{19}$$
$$P(H_1|x) = \left\{ p(h_1|x) \,\middle|\, \Theta_1 \in [\Theta]_1, \pi_{\theta|x} \in \Pi_{\theta|x} \right\}. \tag{20}$$

The (imprecise) ratio between these two imprecise quantities is

$$\left[\frac{P(H_0|x)}{P(H_1|x)}\right] := \left\{ \frac{p(h_0|x)}{p(h_1|x)} \,\middle|\, p(h_i|x) \in P(H_i|x), i = 0, 1 \right\} \tag{21}$$

with supremum

$$\overline{P} := \sup \left[\frac{P(H_0|x)}{P(H_1|x)} \right] \tag{22}$$

and infimum

$$\underline{P} := \inf \left[\frac{P(H_0|x)}{P(H_1|x)} \right] . \tag{23}$$

3.3. Imprecise Loss Values

Typically, the value k (see equation (7)), which completely specifies the loss function (as in Section 2), is difficult to specify unambiguously as a precise value due to insufficient information. An imprecise loss function, however, allows to consider a set K of reasonable values for k (illustrated in Figure 1, bottom).

The imprecise ratio of expected posterior losses is a set

$$R := \left\{ r = k \cdot \frac{p(h_0|x)}{p(h_1|x)} \,\middle|\, k \in K, \frac{p(h_0|x)}{p(h_1|x)} \in \left[\frac{P(H_0|x)}{P(H_1|x)} \right] \right\} \tag{24}$$

that considers all obtainable ratios r of expected posterior losses (equation (11)) that arise within the imprecisely specified setting.

With the supremum $\overline{K} := \sup K$ and infimum $\underline{K} := \inf K$ of K, the imprecise ratio of expected posterior losses R is bounded by

$$\overline{R} := \sup R = \overline{K} \cdot \overline{P} \tag{25}$$

$$\underline{R} := \inf R = \underline{K} \cdot \underline{P}, \tag{26}$$

as all these quantities are non-negative.

Now, the set $\mathscr{A}^*$ of optimal actions is

$$\mathscr{A}^* = \begin{cases} \{\,\} & \text{if } \underline{R} < 1 < \overline{R} \\ \{a_0\} & \text{if } 1 \leq \underline{R}, 1 < \overline{R} \\ \{a_1\} & \text{if } \underline{R} < 1, \overline{R} \leq 1 \\ \{a_0, a_1\} & \text{if } \underline{R} = \overline{R} = 1 \end{cases} . \tag{27}$$

The case with $\underline{R} = \overline{R}$ depicts the precise case (as in Section 2) generalized within the imprecise framework. For $\underline{R} < 1 < \overline{R}$, the available information is not sufficient to unambiguously declare one of the actions as optimal. Therefore, the set $\mathscr{A}^*$ of optimal actions is empty and the decision should be withheld. If so, further information about the imprecisely specified quantities might be obtained, such that they can be specified more precisely (i.e. by smaller sets), or additional data might be collected, allowing to obtain a less imprecise ratio of expected posterior losses R to obtain an optimal action unambiguously.

In a sense, one might say to be "indecisive" if $\mathscr{A}^* = \{\,\}$ or if $\mathscr{A}^* = \{a_0, a_1\}$, as both cases do not yield a single optimal action. However, these cases are fundamentally different. In the first case, there is not enough information to declare one action as superior, in the second case, there is enough information to state that both actions should be rated as practically equal. In that, we want to emphasize that an action is to be considered optimal within this paper, if there is enough information available to declare it as superior or practically equal to the other action (or actions; for the more general case see Section 5.1).

4. Example

Does drug Z help to treat the symptoms of disease D, measured on scale S? Respective actions are

a_0: do not administer drug Z to patients with disease D

a_1: administer drug Z to patients with disease D

To assess this question, a team of investigators plans to run an experiment, in which a number $n = 100$ of patients with disease D are treated with drug Z and the change s_j ($j = 1, \ldots, 100$) of their symptoms on a metric scale S is measured.

Previous investigations showed that treating patients with disease D with a placebo did not increase the symptoms and the standard deviation of the change in symptoms on scale S was 15. Therefore (and for the sake of simplicity within this example), the changes s_j ($j = 1, \ldots, 100$) are modelled as (independent and identically) normally distributed random quantities $S_j \overset{iid}{\sim} N(\delta, (15)^2)$, where the effect parameter $\delta \in \Delta = \mathbb{R}$ represents the difference in change of symptoms compared to the zero-change of the placebo.

A prior distribution for δ is difficult to determine. Although the investigators agree that a normal distribution $\delta \sim N(\mu, \sigma^2)$ might be reasonable (with $\mu \in M$ and $\sigma \in \Sigma$, such that the hyperparameter space is $M \times \Sigma$), they disagree slightly about the specification of the hyperparameters. After some discussions they determine that the set

$$\Pi_\delta = \left\{ N(\mu, \sigma^2) \,\middle|\, \mu \in [3, 12], \sigma \in [10, 20] \right\} \tag{28}$$

covers all different opinions about the hyperparameters, and decide to use it as imprecise prior distribution.

Further, determining whether (a_1) or not (a_0) to administer drug Z to treat disease D for which effect values δ is not easy. The investigators try to consider the consequences of different scenarios, but these cannot be outlined unambiguously as this is a new field of research and some essential information is still pending from other investigations. They can agree that it is safe to assume that effects δ smaller than $\underline{\delta} = 5$ do not justify administering the drug Z due to adverse effects, and that effects δ larger than $\overline{\delta} = 8$ do justify administering the drug Z as the benefits outweigh the adverse effects. For effect values within $[\underline{\delta}, \overline{\delta}] = [5, 8]$ the situation is less obvious. Therefore, they decide to use the imprecise composite hypotheses

$$H_0 : \delta \in [\Delta]_0 \quad \text{vs.} \quad H_1 : \delta \in [\Delta]_1 \tag{29}$$

with

$$[\Delta]_0 := \left\{ \Delta_0 = (-\infty, \tilde{\delta}] \,\middle|\, \tilde{\delta} \in [5,8] \right\} \tag{30}$$

$$[\Delta]_1 := \left\{ \Delta_1 = [\tilde{\delta}, \infty) \,\middle|\, \tilde{\delta} \in [5,8] \right\} . \tag{31}$$

Weighting potential adverse effects against possible benefits of the drug Z is challenging, due to impartial information about the adverse effects. After considering different cases, the investigators determine that a loss function with $k \in K = [\underline{K}, \overline{K}] = [3,15]$ might represent the current knowledge about the consequences of drug Z quite well.

This specification of K is sufficient for the remaining analyses, however, as illustration, consider that the loss function might also be expressed more extensively by relating it directly to the parameter δ (compare equation (8)) via the imprecise specifications (Figure 1, bottom)

$$L_\delta(a_0) = \begin{cases} \{0\} & \text{if } \delta < 5 \\ \{0,1\} & \text{if } 5 \leq \delta \leq 8 \\ \{1\} & \text{if } 8 < \delta \end{cases} \tag{32}$$

and

$$L_\delta(a_1) = \begin{cases} [3,15] & \text{if } \delta < 5 \\ \{0\} \cup [3,15] & \text{if } 5 \leq \delta \leq 8 \\ \{0\} & \text{if } 8 < \delta \end{cases} . \tag{33}$$

Also note, that by using this depiction, the loss function $L_\delta(a)$ might also be interpreted – for each action $a \in \mathscr{A}$ – as an imprecise gamble (illustrations in Figure 1), bridging to the mathematical foundation of the framework of imprecise probabilities in the spririt of Walley (1991) (see also e.g. Quaeghebeur, 2014; Miranda and de Cooman, 2014).

Now, the investigators perform the experiment, obtain the data $s = (s_1, \ldots, s_{100})$, and estimate the effect size (with the in-sample mean) to be $\hat{\delta} = m(s) = 10.03$.

Updating the imprecise prior Π_δ using the generalized Bayes rule (in this case element-wise using the Bayesian normal-normal model with known sample variance) results in the imprecise posterior distribution

$$\Pi_{\delta|s} = \left\{ N(\mu, \sigma^2) \,\middle|\, (\mu, \sigma) \in \mathscr{F} \right\} , \tag{34}$$

where $\mathscr{F}$ is the set (within the hyperparameter space $M \times \Sigma$) as displayed in Figure 2. It can be seen that, compared to the prior distribution Π_δ, the posterior distribution $\Pi_{\delta|s}$ is extremely narrowed down. This is because there is no prior-data-conflict (see e.g. Walter and Augustin, 2009) and the study is highly informative with $n = 100$ patients, such that the data s might easily overwhelm the initial prior uncertainty expressed by Π_δ.

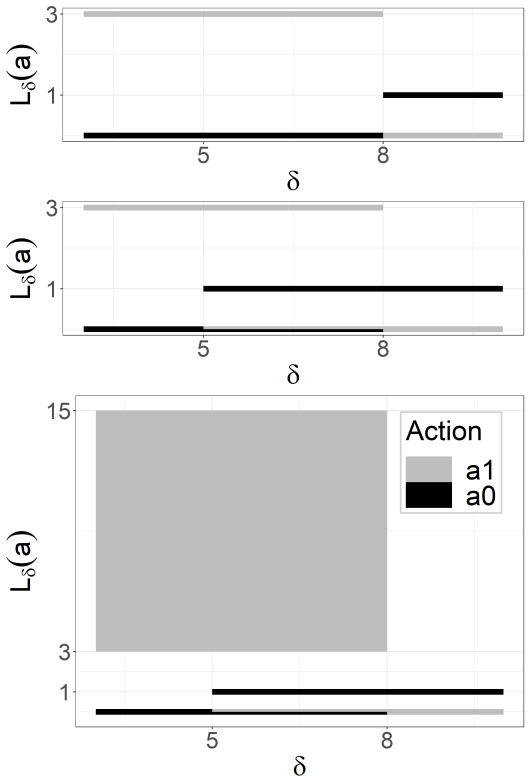

Figure 1: Loss Function. Loss values for both actions a_0 (black) and a_1 (gray) are depicted for varying effect values δ. The top plot illustrates a precise loss function in regret form with $\tilde{\delta} = 8$ (boundary between hypotheses) and $k = 3$. The center plot illustrates how using imprecisely specified hypotheses with $\tilde{\delta} \in [5,8]$ can also be expressed by an imprecisely specified loss function, although $k = 3$ is still precise (please note the overlapping lines at $L_\delta(a) = 0$). The bottom plot adds an imprecisely specified $k \in [3,15]$, leading to the loss function of the example (Section 4). Both the top and center plot are included to illustrate the extension of the precise case into the imprecise framework.

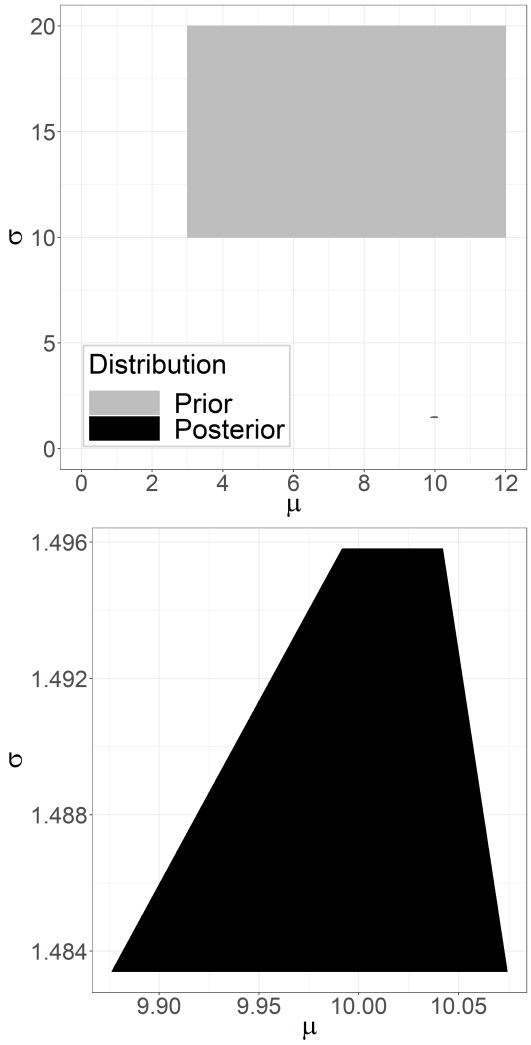

Figure 2: Prior and Posterior Distribution. The gray area (top plot) depicts the set of hyperparameters $(\mu,\sigma) \in M \times \Sigma$ for the normal distributions $N(\mu,\sigma^2)$ that define the imprecise prior distribution Π_δ. This set is bounded by the values $(3,10)$, $(3,20)$, $(12,10)$, and $(12,20)$. Updating the respective distributions leads to normal distributions with parameters $(9.87,1.48)$, $(9.99,1.49)$, $(10.07,1.48)$, and $(10.04,1.49)$, respectively, being extremal elements of the set $\mathscr{F}$ of hyperparamters that define the imprecise posterior distribution. This set is depicted as black area in the top plot bottom-right and enlarged in the bottom plot. Please note the substantial difference in scales between both plots and that this posterior set is not convex (despite its appearance).

With the stated hypotheses (equation (29)), the imprecise posterior probabilities $P(H_0|s)$ and $P(H_1|s)$ are bounded by

$$0.0003 \leq P(H_0|s) \leq 0.103 \tag{35}$$
$$0.897 \leq P(H_1|s) \leq 0.9997\,, \tag{36}$$

leading to ratios

$$0.0003 = \underline{P} \leq \frac{P(H_0|s)}{P(H_1|s)} \leq \overline{P} = 0.115\,. \tag{37}$$

Together with K the imprecise ratio R of expected posterior losses is characterized by

$$\underline{R} = 0.0009 \quad \text{and} \quad \overline{R} = 1.725\,. \tag{38}$$

As $\underline{R} < 1 < \overline{R}$, the investigators cannot state an unambiguous conclusion with the obtained data s and the available vague information.

Some time later, other investigations about the adverse effects of drug Z were finalized showing that the adverse effects are rather mild. This allows the investigators to specify the loss function more precisely by $K = [3,5]$ and the imprecise ratio R of expected posterior losses narrows down to lie between

$$\underline{R} = 0.0009 \quad \text{and} \quad \overline{R} = 0.575\,, \tag{39}$$

now permitting to state a_1 as the optimal action, as $\overline{R} < 1$. Accordingly, the investigators recommend to administer drug Z to treat patients with disease D.

R code to replicate the example is provided electronically.

5. Discussion

5.1. Scalability

The elaborations within this paper were restricted towards a framework that uses only two hypotheses and, therefore, only two actions. While this is currently the most used framework for hypothesis-based analyses in applied sciences, the considerations within this paper might naturally be scaled towards using multiple hypotheses and actions.

In fact, the depicted case (with only two hypotheses and actions) that determines the optimal action(s) (equation (27)) by considering the ratio R of expected posterior losses (equation (24)) might be considered as a special case in the context of interval dominance in the imprecise decision theoretic framework. This shall be illustrated within the context of the example (Section 4).

As outlined in equations (32) and (33), the loss function might also be viewed w.r.t. the parameter δ, such that the conditions in these multi-case equations are determined by the hypothesis specifications and the values by the actual consequences of the decision. Naturally, additional

hypotheses and actions can easily be incorporated using this formulation (although the applied scientists might have more values to specify).

These conditions need to be evaluated w.r.t. the posterior distribution $\Pi_{\delta|s}$, determining the (imprecise) posterior probabilities to be bounded by

$$0.0003 \leq P(\delta < 5|s) \leq 0.0005 \quad (40)$$
$$0.0806 \leq P(5 \leq \delta \leq 8|s) \leq 0.1024 \quad (41)$$
$$0.8970 \leq P(8 < \delta|s) \leq 0.9189\,. \quad (42)$$

The expected posterior loss $\rho_\delta : \mathscr{A} \to \mathbb{R}_0^+$:

$$a \mapsto \rho_\delta(a) = \int_\Delta L_\delta(a) \cdot \pi(\delta|s)\, d\delta \quad (43)$$

is now based on the parameter-based loss function L_δ (in contrast to the hypothesis-based loss function L as in equation (6)), and needs to consider that L_δ is an imprecise quantity and that $\pi(\delta|s)$ denotes the probability densities of the imprecise posterior $\Pi_{\delta|s}$. Respective values are bounded by

$$0.897 \leq \rho_\delta(a_0) \leq 1.021 \quad (44)$$
$$0.0009 \leq \rho_\delta(a_1) \leq 1.544\,. \quad (45)$$

Interval dominance (see e.g. Huntley et al., 2014) compares interval-valued expected posterior losses. Then, an action is declared dominated if there exists another action that has an interval-valued expected posterior loss being strictly less. Such a dominated action can be ruled out. In line with the considerations about optimal actions in Section 3.3, we consider an action as optimal (in the context of interval dominance) if it dominates or practically equals[2] every other action. If there are other actions that cannot be dominated by an action, information is lacking to treat this action as superior and it should not be considered as optimal. In the case considered here, $\rho_\delta(a_0)$ lies within the range of $\rho_\delta(a_1)$, so neither expected posterior loss interval dominates the other one. Thus, no action proves itself as superior.

If, as in the example, additional information were gathered to narrow k down to be within $[3,5]$, the expected posterior loss $\rho_\delta(a_1)$ of action a_1 is then bounded by

$$0.0009 \leq \rho_\delta(a_1) \leq 0.514\,, \quad (46)$$

now being completely below the expected posterior loss $\rho_\delta(a_0)$ of action a_0. Action a_1 dominates action a_0 and is thus considered to be optimal. Apparently, these are the same results as in Section 4 and R code to replicate these numbers is provided electronically.

This illustrates that the framework depicted within this paper can be considered as a special case of the imprecise Bayesian decision theoretic framework using the concept of interval dominance and the conditional Bayes principle, which might be easily extended to additional hypotheses and actions. By being restricted to this special case of only two hypotheses and actions, a simple regret form of the loss function can be used, allowing to determine the optimal action easily via the ratio of expected posterior losses. This allows to extend hypothesis-based analysis into the imprecise decision theoretic framework without exceedingly complicated mathematical formulas, a fact that might be welcomed by applied scientists.

5.2. Conditional Perspective

A rigorous conditional perspective was taken within this Bayesian decision-theoretic approach: The prior gets updated first to the posterior before considering the decision problem and finding the optimal action based on this posterior distribution. However, there is also a different, i.e. unconditional, decision-theoretic approach, which starts by finding an optimal decision function (mapping all possible data sets to optimal actions) by minimizing the prior risk. This approach takes all potentially observable data sets into account, and focuses on the actually observed data only as a second step, evaluating the decision function at the concretely observed sample. Within the precise case, both approaches yield eventually the same optimal action (cp. e.g. Berger, 1985, p. 159). This is, however, not necessarily true within the imprecise case (Augustin, 2003),[3] see also (Seidenfeld, 1994) in a related game theoretic context, – a fact that breathes new life into an old debate between the conditional and the frequentist point of view. The existence of this decision-theoretic dynamic inconsistency in the context of point-wise updating set-valued distributions requires the applied scientists to reason about whether to use the conditional or the unconditional perspective in their analyses. However, the conditional perspective is generally considered to be the preferred point of view within Bayesian statistics, as it does not consider other potential data sets that were not observed (Jeffreys, 1961). The argumentation by Berger (1985, p. 160, notation adapted, italics preserved)

> Note that, from the conditional perspective together with the utility development of the loss, the *correct* way to view the situation is that of minimizing $[\rho(a)]$. One should condition on what is known, namely $[x]$ (...), and average the utility over what is unknown, namely θ. The desire to minimize [the prior risk] would be deemed rather bizarre from this perspective.

2. As depicted in Section 2, an action practically equals another action if their expected posterior losses are precise and with identical value.

3. The proof of the equivalence of prior risk and posterior loss essentially relies on the interchangeability of integrals over the parameter space and the sample space, which is no longer valid if also maxima/minima of distributions have to be considered in between. This suggests that similar dynamic inconsistencies may occur as soon as imprecise losses are considered.

carries over to the rigorous generalization developed here, at least as long as single experiments with a single decision are considered.

6. Outlook

Within this paper, the hypothesis-based Bayesian decision-theoretic framework with composite hypotheses was extended to include imprecise specifications of the prior distribution, the hypotheses, and the loss function. These three quantities are expected to be the most difficult to specify in current applied sciences, if a hypothesis-based decision-theoretic Bayesian analysis is intended. Therefore, their imprecise extension might provide a useful framework for applied scientists.

This approach might also be seen as an extension to the Bayes factor (Gönen et al., 2005; Kass and Raftery, 1995; Rouder et al., 2018), a quantity involved in updating prior probabilities of hypotheses to their posterior probabilities (as in Section 2, equations (4) and (5)) and interpreted as quantification of the evidence within the data w.r.t. the hypotheses. This extension covers the ability to include actions and a (hypothesis-based) loss function into the statistical analysis (such that the practical implication of the study can be considered on a formal level) and the opportunity to treat impartial information about the prior (see also Ebner et al., 2019), the hypotheses, and the loss function as it is, without requiring a level of precision that is not available.

Appendix A. R Code

R code to replicate the example is provided electronically.

Acknowledgments

The authors thank the reviewers for their valuable comments and open sharing of ideas that allowed to significantly improve the quality of this work.

Competing Interests

The authors declare to have no competing interests.

References

Thomas Augustin. On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view — a cautionary note on updating imprecise priors. In Jean-Marc Bernard, Teddy Seidenfeld, and Marco Zaffalon, editors, *ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, pages 31–45, Lugano, Waterloo, 2003. Carleton Scientific. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.9683.

Thomas Augustin, Frank P. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley, Chichester, 2014a. doi: 10.1002/9781118763117.

Thomas Augustin, Gero Walter, and P. Coolen, Frank. Statistical inference. In Thomas Augustin, Frank P. Coolen, Gert de Cooman, and Matthias Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014b. doi: 10.1002/9781118763117.ch7.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1985. doi: 10.1007/978-1-4757-4286-2.

James O. Berger and Robert L. Wolpert. The likelihood principle. *Lecture Notes-Monograph Series*, 6:iii–160.2 (discussion: 160.3–199), 1988. URL http://www.jstor.org/stable/4355509.

Luisa Ebner, Patrick Schwaferts, and Thomas Augustin. Robust Bayes factor for independent two-sample comparisons under imprecise prior information. In Jasper De Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 167–174. PMLR, 2019. URL http://proceedings.mlr.press/v103/ebner19a.html.

Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall. The Bayesian two-sample t test. *The American Statistician*, 59:252–257, 2005. doi: 10.1198/000313005X55233.

Nathan Huntley, Robert Hable, and Matthias C. M. Troffaes. Decision making. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 190–206. John Wiley & Sons, 2014. doi: 10.1002/9781118763117.ch8.

Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. doi: 10.1017/CBO9780511790423.

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

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995. doi: 10.2307/2291091.

Roger E. Kirk. Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5):746–759, 1996. doi: 10.1177/0013164496056005002.

John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan*. Academic Press, New York, 2015. doi: 10.1016/B978-0-12-405888-0.09999-2.

John K. Kruschke. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018. doi: 10.1177/2515245918771304.

Daniël Lakens. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. doi: 10.1177/2515245918770963.

Enrique Miranda and Gert de Cooman. Lower previsions. In Thomas Augustin, Frank P. Coolen, Gert de Cooman, and Matthias Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 28–55. Wiley, 2014. doi: 10.1002/9781118763117.ch2.

Erik Quaeghebeur. Desirability. In Thomas Augustin, Frank P. Coolen, Gert de Cooman, and Matthias Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 1–27. Wiley, 2014. doi: 10.1002/9781118763117.ch1.

David Rios Insua and Fabbrizzio Ruggeri, editors. *Robust Bayesian Analysis. Lecture Notes in Statistics 152*. Springer, New York, 2000. doi: 10.1007/978-1-4612-1306-2.

Jeffrey N. Rouder, Julia M. Haaf, and Joachim Vandekerckhove. Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1):102–113, 2018. doi: 10.3758/s13423-017-1420-7.

Patrick Schwaferts and Thomas Augustin. Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In Jasper De Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 338–345. PMLR, 2019. URL http://proceedings.mlr.press/v103/schwaferts19a.html.

Teddy Seidenfeld. When normal form and extensive form solutions differ. In Dag Prawitz, Brian Skyrms, and Dag Westerstahl, editors, *Logic, Methodology and Philosophy of Science IX (Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science, Uppsala, 1991)*, pages 451–463. Elsevier, 1994. doi: 10.1016/S0049-237X(06)80056-X.

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

G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009. doi: 10.1080/15598608.2009.10411924.

Contribution 7

Schwaferts & Augustin (2021c): How to Guide Decisions with Bayes Factors. (Preprint)

How to Guide Decisions with Bayes Factors

Patrick Schwaferts, Thomas Augustin

patrick.schwaferts@stat.uni-muenchen.de
thomas.augustin@stat.uni-muenchen.de

Ludwig-Maximilians-Universität Munich
Department of Statistics
Methodological Foundations of Statistics and its Applications
Ludwigsstraße 33, 80539 Munich, Germany

Abstract

Some scientific research questions ask to guide decisions and others do not. By their nature frequentist hypothesis-tests yield a dichotomous test decision as result, rendering them rather inappropriate for latter types of research questions. Bayes factors, however, are argued to be both able to refrain from making decisions and to be employed in guiding decisions. This paper elaborates on how to use a Bayes factor for guiding a decision. In this regard, its embedding within the framework of Bayesian decision theory is delineated, in which a (hypothesis-based) loss function needs to be specified. Typically, such a specification is difficult for an applied scientist as relevant information might be scarce, vague, partial, and ambiguous. To tackle this issue, a robust, interval-valued specification of this loss function shall be allowed, such that the essential but partial information can be included into the analysis as is. Further, the restriction of the prior distributions to be proper distributions (which is necessary to calculate Bayes factors) can be alleviated if a decision is of interest. Both the resulting framework of hypothesis-based Bayesian decision theory with robust loss function and how to derive optimal decisions from already existing Bayes factors are depicted by user-friendly and straightforward step-by-step guides.

Keywords: Bayesian Statistics, Bayes Factor, Decision Theory, Robustness, Imprecise Probabilities

1 Introduction

The result of a classic frequentist hypothesis test is a dichotomous test decision. However, scientific research questions are very versatile and there is not always the demand to guide a decision. By the their nature, frequentist hypothesis tests prohibit a statistical hypothesis-based analysis without making decisions. In that sense, a statistical framework that provides results without requiring an underlying (potentially artificially constructed) decision problem seems to be advantageous. The Bayes factor – a Bayesian quantity that is used for hypothesis comparisons (Jeffreys, 1961; Kass and Raftery, 1995; Gönen et al., 2005; Rouder et al., 2009) – is argued to do so, as it is typically interpreted as evidence quantification w.r.t. the contrasted hypotheses (see e.g. Morey et al., 2016) without requiring a decision to be made. In this regard, Rouder et al. (2018) state that "[r]efraining from

making decisions strikes [them] as advantageous in most contexts."

Naturally, the evidence (as quantified by the Bayes factor) might then be used to update beliefs in the considered hypotheses and subsequently to guide a respective decision. The essential point, however, is that the researcher might stop the analysis after calculating a Bayes factor without guiding a decision, e.g. if merely the evidence quantification is of interest. Then the result of the Bayesian analysis is the Bayes factor itself and not a decision. Yet, for those research situations that do indeed aim at guiding a decision, the Bayes factor might naturally be used to do so. The aim of this elaboration is to outline the decision theoretic framework in which Bayes factors are involved.

Further, it shall be acknowledged that the specification of the relevant quantities within such a decision theoretic framework as precise values might not always be possible for an applied scientist, as the available relevant information might be scarce, vague, partial, and ambiguous. To tackle this issue, also a robust version of the framework shall be outlined in which the applied researcher is allowed to specify the essential quantities less precisely as sets of values, such that the partial nature of the relevant information might be captured more accurately. Although such robust specifications might be possible for all essential quantities (see e.g. Schwaferts and Augustin, 2019, 2021), the present elaboration is restricted to a robustly specified inverval-valued loss function, as it is this quantity which characterizes the difference between a decision-theoretic and a non-decision-theoretic analysis, yet its precise specification is expected to bear serious difficulties for applied scientists.

The elaborations within this paper are structured as follows: After delineating the general (Section 2) and the hypothesis-based (Section 3) framework of Bayesian decision theory, its relation with Bayes factors is depicted (Section 4). To facilitate a more user-friendly employment of the hypothesis-based Bayesian decision theoretic framework, a robust interval-valued specification of the loss function was allowed (Section 5) and the restriction of the prior distributions to be proper can be alleviated (Section 6). Both the resulting framework (Secion 7.1) and how to derive optimal actions from existing Bayes factor values (Secion 7.2) are presented in respective step-by-step guides.

2 Bayesian Decision Theory

Within the framework of Bayesian decision theory (e.g. Berger, 1985; Robert, 2007), the objective is to decide between different actions. In accordance with the context of Bayes factors, only two actions shall be considered, namely a_0 and a_1, being comprised within the action space $\mathcal{A} = \{a_0, a_1\}$.

The researcher plans to conduct an investigation that yields data $\boldsymbol{x}$, which is characterized by a parametric sampling distribution with parameter $\theta \in \Theta$, where Θ is the parameter space. Accordingly, the density of the data is $f(\boldsymbol{x}|\theta)$.

In a Bayesian setting, a prior distribution on the parameter θ with density $\pi(\theta)$ needs to be specified. This prior reflects the information (or belief or knowledge or uncertainty) about the parameter before the investigation is conducted.

In addition, also a loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}_0^+ : (\theta, a) \mapsto L(\theta, a)$ needs to be specified, which quantifies the "badness" of the consequences of deciding for the action $a \in \mathcal{A}$ if the

parameter value $\theta \in \Theta$ is true. Usually, the exact shape of this loss function is inaccessible and hypothesis-based analyses are able to tackle this issue. These are depicted within the next section, but first – to delineate the ideal Bayesian solution – assume L is fully known.

Now, after specifying the parametric sampling distribution, the prior, as well as the loss function, the investigation can be conducted and the data $\boldsymbol{x}$ are observed. This allows to update the prior distribution via Bayes rule to the posterior distribution with density

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\,\pi(\theta)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\,\pi(\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\,\pi(\theta)\,d\theta}\,. \tag{1}$$

There are plenty of resources available about how to obtain this posterior (e.g. Gelman et al., 2013; Kruschke, 2015), which reflects the information (or belief or knowledge or uncertainty) about the parameter after the investigation was conducted.

Based on this posterior distribution, it is possible to calculate the expected posterior loss $\rho : \mathcal{A} \to \mathbb{R}_0^+$ for each action by integrating the loss function L over the posterior density:

$$\rho(a) = \int_\Theta L(\theta, a)\,\pi(\theta|\boldsymbol{x})\,d\theta\,. \tag{2}$$

The optimal action a^* has minimal expected posterior loss:

$$a^* = \arg\min_{a\in\mathcal{A}} \rho(a)\,. \tag{3}$$

3 Hypothesis-Based Bayesian Decision Theory

As mentioned, typically, the loss function L is not fully accessible as the essential information about it might be scarce, vague, partial, and ambiguous. A commonly employed solution is a hypothesis-based analysis: The researcher considers each possible parameter value θ and assesses which action should be preferred if this parameter value would be true. These considerations lead to two sets of parameters Θ_0 and Θ_1 for which the actions a_0 and a_1 should be preferred, respectively. These sets define the hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \tag{4}$$

employed in conventional analyses, such as hypothesis tests or Bayes factors.

From the posterior density $\pi(\theta|\boldsymbol{x})$ it is possible to determine the posterior probabilities of the parameters sets Θ_0 and Θ_1, i.e. of the hypotheses H_0 and H_1, by

$$P(H_0|\boldsymbol{x}) = \int_{\Theta_0} \pi(\theta|\boldsymbol{x})\,d\theta \qquad \text{and} \qquad P(H_1|\boldsymbol{x}) = \int_{\Theta_1} \pi(\theta|\boldsymbol{x})\,d\theta\,, \tag{5}$$

respectively. The ratio of these beliefs $P(H_0|\boldsymbol{x})/P(H_1|\boldsymbol{x})$ is referred to as posterior odds.

The underlying assumption of hypothesis-based analyses is that the loss values within these sets Θ_0 and Θ_1 are constant, respectively (see Figure 1). This assumption shall be referred to as *simplification assumption* and is inherent to a statistical analysis which considers statistical hypotheses and derives applied conclusions based on respective (hypothesis-based) results. In addition (without loss of generality), the loss values for deciding correctly (i.e. for a_0 if $\theta \in \Theta_0$ or for a_1 if $\theta \in \Theta_1$) can be set to 0. The resulting loss function is

Table 1: Simplified Hypothesis-Based Loss Function.

$L(\theta, a)$	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$a = a_0$	0	k_0
$a = a_1$	k_1	0

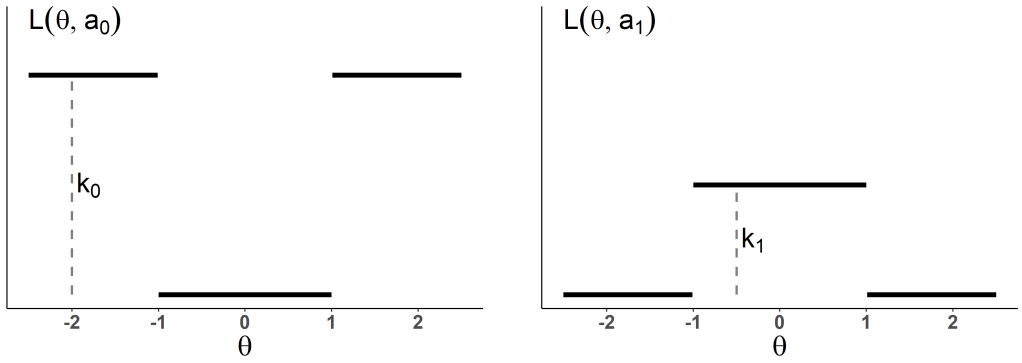

Figure 1: Hypothesis-Based Loss Function. Assume $\Theta = \mathbb{R}$, $\Theta_0 = [-1, 1]$, and $\Theta_1 = (-\infty, -1) \cup (1, \infty)$. The hypothesis-based loss function L (y-axis) in regret form (see Table 1) in dependence of the parameter θ (x-axis) and the actions a_0 (left) and a_1 (right) is assumed to be constant within the sets Θ_0 and Θ_1, respectively. This is an assumption (*simplification assumption*) inherent to a hypothesis-based statistical analysis which – at least implicitly – considers an underlying applied decision problem.

in regret form (depicted in Table 1) and has only two values to specify: $k_0 := L(a_0, \theta)$ if $\theta \in \Theta_1$ and $k_1 := L(a_1, \theta)$ if $\theta \in \Theta_0$.

With this simplified loss function (Table 1), the expected posterior loss of each action can be calculated as

$$\rho(a_0) = \int_\Theta L(\theta, a_0)\, \pi(\theta|\boldsymbol{x})\, d\theta = k_0 \cdot P(H_1|\boldsymbol{x}) \tag{6}$$

$$\rho(a_1) = \int_\Theta L(\theta, a_1)\, \pi(\theta|\boldsymbol{x})\, d\theta = k_1 \cdot P(H_0|\boldsymbol{x}) \tag{7}$$

and the action with minimal expected posterior loss shall be selected.

Only the ratio $k := k_1/k_0$ is required to determine this optimal action. This ratio k states how much worse it would be to decide for a_1 if $\theta \in \Theta_0$ is true (type-I-error) than to decide for a_0 if $\theta \in \Theta_1$ is true (type-II-error), if deciding correctly has loss 0. With the ratio of expected posterior losses

$$\varrho(k) := \frac{\rho(a_1)}{\rho(a_0)} = k \cdot \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})} \tag{8}$$

the optimal action is

$$a^* = \begin{cases} a_0 & \text{if} \quad \varrho(k) > 1 \\ a_1 & \text{if} \quad \varrho(k) < 1 \end{cases} . \tag{9}$$

For $\varrho(k) = 1$ any action might be chosen.

4 Bayes Factors

By assessing even the prior distribution in the light of the hypotheses, it is possible to obtain the prior probabilities in the hypotheses (illustrated in Figure 2, left):

$$P(H_0) = \int_{\Theta_0} \pi(\theta)\, d\theta \qquad \text{and} \qquad P(H_1) = \int_{\Theta_1} \pi(\theta)\, d\theta\,, \tag{10}$$

Analogously, the ratio of these beliefs $P(H_0)/P(H_1)$ is referred to as prior odds.

In addition, the prior distribution can be restricted to each of the hypotheses, referred to as within-hypothesis priors (illustrated in Figure 2, middle and right), and the corresponding densities are

$$\pi(\theta|H_0) = \frac{1}{P(H_0)}\, \pi(\theta) \cdot \mathbf{1}(\theta \in \Theta_0) \tag{11}$$

$$\pi(\theta|H_1) = \frac{1}{P(H_1)}\, \pi(\theta) \cdot \mathbf{1}(\theta \in \Theta_1)\,, \tag{12}$$

where $\mathbf{1}(s) = 1$ if the statement s is true and $\mathbf{1}(s) = 0$ if s is false.

The overall prior distribution can be decomposed (cp. Rouder et al., 2018) into the prior probabilities of the hypotheses and the within-hypothesis priors (Figure 2):

$$\pi(\theta) = P(H_0)\, \pi(\theta|H_0) + P(H_1)\, \pi(\theta|H_1)\,. \tag{13}$$

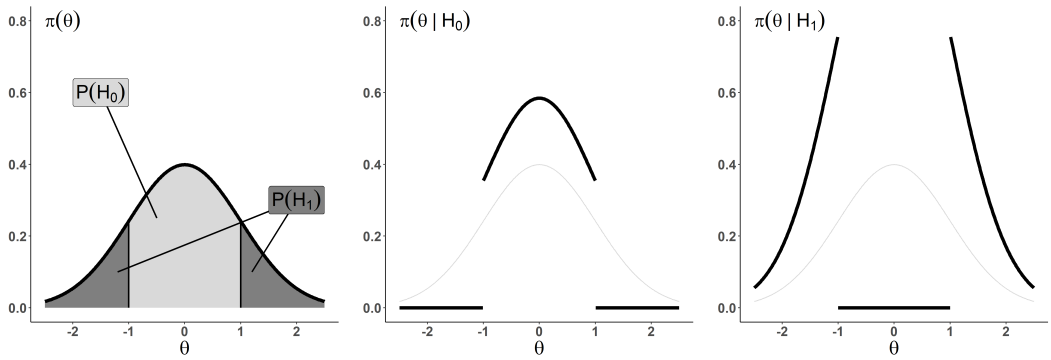

Figure 2: Prior Decomposition. Assume $\Theta = \mathbb{R}$, $\Theta_0 = [-1, 1]$, $\Theta_1 = (-\infty, -1) \cup (1, \infty)$ and a standard normal distribution for $\theta \sim N(0, 1)$. Left: The prior density $\pi(\theta)$ is depicted as solid line. $P(H_0)$ and $P(H_1)$ can be calculated as respective areas under this density, depicted as light gray and dark gray, respectively. Middle: The within-hypothesis density $\pi(\theta|H_0)$ as in equation (11) is depicted as solid line. Right: The within-hypothesis density $\pi(\theta|H_1)$ as in equation (12) is depicted as solid line.

Instead of considering the overall prior distribution together with the hypotheses (which leads to the posterior odds, as in Section 3), the Bayes factor is obtained by considering only the within-hypothesis priors together with the hypotheses:

$$BF := \frac{\int_{\Theta_0} f(\boldsymbol{x}|\theta)\, \pi(\theta|H_0)\, d\theta}{\int_{\Theta_1} f(\boldsymbol{x}|\theta)\, \pi(\theta|H_1)\, d\theta}\,. \tag{14}$$

The posterior odds can then be calculated from the Bayes factor and the prior odds:

$$\frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})} = BF \cdot \frac{P(H_0)}{P(H_1)}\,. \tag{15}$$

The optimal decision can now be obtained as in the previous section (Section 3) by considering the loss function.

5 Robust Loss Function

However, a precise specification of the value k is typically not accessible, as essential information about the "badness" of the consequences of the decision are scarce, vague, partial, and ambiguous. Yet, this partial information needs to be included into the analysis, as ignoring it facilitates suboptimal decisions. A decision cannot be guided properly without considering its consequences.

This partial information about the loss can be captured less arbitrarily and more robustly by an interval $[\underline{K}, \overline{K}]$ than by a precise value k (cp. e.g. Walley, 1991; Augustin et al., 2014). To do so, the researcher has to determine a lower bound $\underline{K}$ and an upper bound $\overline{K}$ for reasonable k values (i.e. for the ratio of how much worse the type-I-error is compared to the type-II-error, if deciding correctly has a loss of 0).

To perform a robust analysis (cp. also Ríos Insua and Ruggeri, 2012) with this interval-valued specification, it is possible to obtain and consider the optimal action for each value within this interval $[\underline{K}, \overline{K}]$.

If the optimal action is the same for each k within $[\underline{K}, \overline{K}]$, then this action should be chosen. If not, the decision should be withheld, because the data or the information about the decision problem are not sufficient to unambiguously guide the decision.

Formally (see also Schwaferts and Augustin, 2019, 2020, 2021), the ratios of expected posterior losses need to be calculated for both the lower and upper bound, respectively:

$$\varrho(\underline{K}) = \underline{K} \cdot \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})} \qquad \text{and} \qquad \varrho(\overline{K}) = \overline{K} \cdot \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}\,. \tag{16}$$

Then, the optimal action is

$$a^* = \begin{cases} a_0 & \text{if} \quad \varrho(\underline{K}) \geq 1 \\ a_1 & \text{if} \quad \varrho(\overline{K}) \leq 1 \end{cases}\,. \tag{17}$$

For $\varrho(\underline{K}) < 1 < \varrho(\overline{K})$, the decision should be withheld.

6 Improper Priors

Furthermore, the calculation of Bayes factors comes along with a restriction (Jeffreys, 1961) on the prior distribution: It must be a proper distribution, i.e. the density has to integrate to 1:

$$\int_\Theta \pi(\theta)\, d\theta = 1\,. \tag{18}$$

In contrast, an improper prior distribution is characterized by a non-integrable function (e.g. $\pi(\theta) \propto c$ with $c > 0$ being a constant, see Figure 3, dotted line) and, technically, this prior distribution is no proper probability distribution. However, these improper priors are frequently allowed within Bayesian prior specifications, because they might lead to proper posterior distributions (see Figure 3, solid line). In this case, the posterior odds $P(H_0|\boldsymbol{x})/P(H_1|\boldsymbol{x})$ can be calculated reasonably and a decision can be guided consistently.

The prior odds, however, might not be reasonable (e.g. with $P(H_0)/P(H_1) = 0$ as in Figure 3). Accordingly, the Bayes factor (calculated via equation (15))

$$BF = \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})} \Big/ \frac{P(H_0)}{P(H_1)} \tag{19}$$

cannot be calculated reasonably due to its dependence on the prior odds. Therefore, Bayes factors require – in contrast to a Bayesian analysis in general – proper prior distributions. This is truly a limitation, as improper priors are frequently employed for representing non-knowledge or for letting the data speak for themselves (e.g. Gelman et al., 2013).

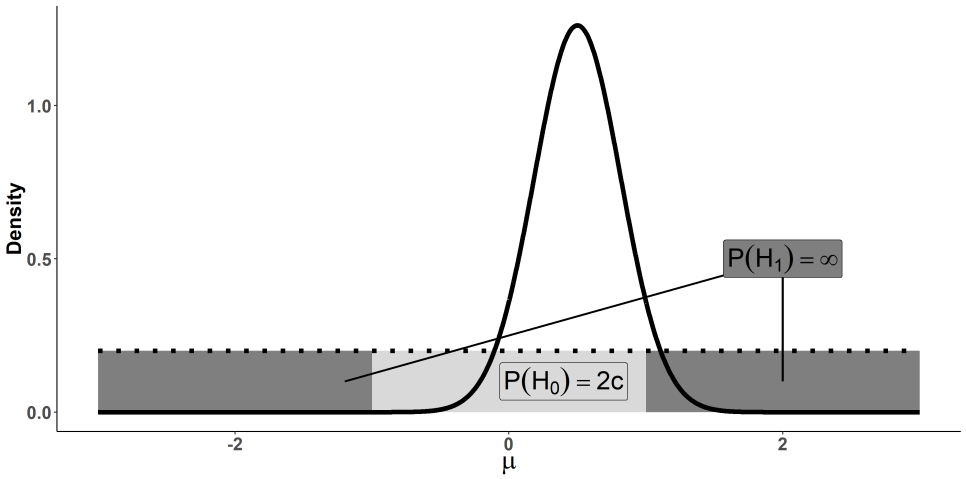

Figure 3: Improper Prior. Assume the model $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, with known variance $\sigma^2 = 1$ and unknown parameter $\mu \in \mathbb{R}$, the hypotheses $\Theta_0 = [-1, 1]$, $\Theta_1 = (-\infty, -1) \cup (1, \infty)$. The function $\pi(\mu) = c$, with $c = 0.2$ being an arbitrary constant, characterizes the improper prior distribution for μ (dotted line). For a sample of size $n = 10$ with in-sample mean $\bar{x} = 0.5$, the posterior distribution is proper (solid line), such that its density integrates to 1. The prior "beliefs" into the hypotheses are with $P(H_0) = 2c$ and $P(H_1) = \infty$ not reasonably interpretable (light gray and dark gray, respectively).

This issue is alleviated in hypothesis-based Bayesian decision theoretic accounts, as improper priors typically yield proper posterior odds. Accordingly, a researcher who is interested in guiding a decision might employ the decision theoretic framework directly without explicitly calculating the Bayes factor. Then, also improper priors might be employed.

Please note that it is still an ongoing debate whether non-knowledge can be formalized by a precise improper prior distribution and if so, which improper prior distribution shall be employed. Although the authors of this paper doubt that non-knowledge can be formalized by a precise prior distribution, even if it is improper (cp. e.g. Augustin et al., 2014),

this issue shall not be addressed here. In general, it is important that the employed prior distribution matches with the available information (or non-information) about the phenomenon of interest, and this applies to every point of view within this debate. In this regard, the present elaboration emphasizes only that it is mathematically possible to employ improper priors if decisions are of interest, which is an advantage of the (more general) hypothesis-based Bayesian decision theoretic account over Bayes factors.

7 Step-By-Step Guides

7.1 Hypothesis-Based Bayesian Decision Theory

In order to apply this hypothesis-based Bayesian decision theoretic framework with robust loss function, a researcher might follow the following steps.

Step 1: Actions. First of all, the researcher needs to specify the actions. It is recommended to explicitly state and report these actions, e.g. by (this example is taken from Bartolucci et al., 2011)

a_0: do not administer aspirin to prevent myocardial infarction

a_1: administer aspirin to prevent myocardial infarction

If the researcher has difficulties stating the actions, maybe there is no decision to guide and a descriptive analysis might suffice (cp. also Cumming, 2014; Kruschke and Liddell, 2018).

Step 2: Sampling Distribution. Next, the researcher should provide a detailed description of the investigation and how it is characterized (i.e. the sampling distribution). It is recommended to also explicitly state the employed parameter θ and its interpretation. This is the basis for specifying the hypotheses.

Step 3: Prior Distribution. In the Bayesian setting, it is possible to include prior information (or belief or knowledge or uncertainty) into the analysis. In that, the researcher has to specify a prior distribution on the parameter. It is recommended to fully report the available prior information about the parameter θ and why this leads to the prior density $\pi(\theta)$.

Of course, this specification is far from being unambiguous. However, this is a fundamental issue inherent to every Bayesian analysis (not only Bayesian decision theoretic accounts) and solving this issue is not the intention of this elaboration. Nevertheless, solutions, such as sensitivity analyses (found in almost every Bayesian textbook, e.g. Gelman et al., 2013), exist. It is recommended at this step of the analysis to also state all other possible prior densities that are in accordance with the available prior information, as these serve as basis for a subsequent sensitivity analysis.

Naturally, also non-informative priors might be specified and they might also be improper (as long as they lead to proper posterior distributions).

Step 4: Assumption. If the researcher is unable to specify the loss function L, then a hypothesis-based simplification as in Section 3 might be a solution. This simplification is an assumption on the loss function, namely that the loss function is constant within each of two parameter sets. If this assumption is not appropriate, it might lead to errors (which

are inherent to every hypothesis-based analysis) and the researcher needs to be aware of this consequence. It is recommended to explicitly report that this assumption was made. Transparency is one of the basic principles in science (cp. Gelman and Hennig, 2017).

Step 5: Hypotheses. Now, the researcher has to consider each possible parameter value θ and assess which action should be preferred if this parameter value would be true. All parameters for which a_0 or a_1 should be preferred are comprised within the sets Θ_0 or Θ_1, respectively. Certainly, there are parameter values that define the border between both sets Θ_0 and Θ_1. It is recommended to explicitly state what these values mean in real-life and why they define reasonable borders between Θ_0 and Θ_1.

Step 6: Errors. Deciding for a_1 if $\theta \in \Theta_0$ is the type-I-error and deciding for a_0 if $\theta \in \Theta_1$ is the type-II-error. Both errors should be delineated, as they serve as basis for specifying the ratio k. It is recommended to explicitly state these errors and their consequences, e.g. by

> Type-I-error: administer aspirin to prevent myocardial infarction, but the effect is negligible. Consequence: patients unnecessarily suffer side effects of aspirin.
>
> Type-II-error: do not administer aspirin to prevent myocardial infarction, although it would have an effect. Consequence: some patients suffer a myocardial infarction, which could have been prevented.

Of course, this is only a schematic illustration and in real empirical studies these elaborations will be more comprehensive.

Step 7: Loss Magnitude. The researcher has to imagine that the "badness" of deciding correctly is 0. In this context, the researcher has to determine how much worse the type-I-error is compared to the type-II-error. This is the value k. As a precise value for k is difficult to determine, it might be easier to specify a range $[\underline{K}, \overline{K}]$ of plausible values for k. It is recommended to report all considerations that lead to this specification.

Step 8: Investigation. Now, the investigation can be conducted and it is recommended to preregister[1] the previous specifications, the design of the experiment, and the planned (decision theoretic) analysis of the data (cp. Nosek et al., 2018; Klein et al., 2018). Registered reports[2] even allow to obtain a peer-review prior to collecting the data.

Step 9: Posterior Distribution. The observed data are used to obtain the posterior distribution as well as the posterior beliefs in the hypotheses $P(H_0|\boldsymbol{x})$ and $P(H_1|\boldsymbol{x})$. There are countless references on how to do this (e.g. Gelman et al., 2013; Kruschke, 2015).

Step 10: Optimal Action. The researcher has to calculate $\varrho(\underline{K})$ and $\varrho(\overline{K})$ as in equation (16) to find the optimal action as in equation 17).

For $\varrho(\underline{K}) < 1 < \varrho(\overline{K})$, the decision should be withheld, because the data or the information about the decision problem are not sufficient to unambiguously guide the decision. In this case, a reasonable strategy might be to collect more data or to gather more information about the decision problem, especially about the consequences of the errors, to narrow down $[\underline{K}, \overline{K}]$. However, it is recommended to transparently report that a decision was withheld at first and which subsequent steps were taken to obtain more information.

[1] Study designs can be preregistered e.g. at `www.cos.io/initiatives/prereg`.
[2] Information about registered reports can be found e.g. at `www.cos.io/rr`.

Step 11: Publish Data. Of course, other researchers might need the data to guide their decisions. It is to expect that they have different prior knowledge and that their decisions employ different hypotheses. Without having access to the data set (but only to the reported analysis), it might be difficult, or even impossible, for them to guide their decisions properly, emphasizing the importance of open science[3].

7.2 From Bayes Factors to Decisions

Sometimes, a researcher wants to use the results of a previous study to guide a decision. Assume a Bayes factor *BF* was already calculated and shall now be used to guide this decision.

Step A: Applicability of the Sampling Distribution. Confirm that the interpretation of the parameter θ is actually relevant for the decision of interest. If this is not the case, the available data (or Bayes factor) can hardly be used to guide the decision of interest.

Step B: Applicability of the Hypotheses. Certain specific hypotheses were assumed in order to calculate the Bayes factor. These need to match with the decision problem of interest. To assess this, the potential actions of the decision problem of interest need to be delineated as in *Step 1* and the parameter sets Θ_0 and Θ_1 need to be obtained as in *Step 5*. These sets have to be equivalent to the hypotheses that were employed in the calculation of the Bayes factor. If this is not the case, it is recommended to not use this Bayes factor value and restart the decision theoretic account within the previous section (Section 7.1). In this regard, it is helpful if the data set, that was used to calculate the original Bayes factor, is fully accessible.

Step C: Applicability of the Prior Distribution. Confirm that the employed within-hypothesis prior distributions for calculating the Bayes factor match with the available information about the phenomenon of interest. If this is not the case, it is recommended to not use this Bayes factor value and restart the decision theoretic account within the previous section (Section 7.1). Again, to do so it is helpful if the data set, that was used to calculate the original Bayes factor, is fully accessible.

Step D: Prior Odds. As the calculation of the Bayes factor does not require the prior odds, only the within-hypothesis prior distributions, former need to be specified to guide a decision. In this regard, the researcher has to specify the prior probabilities of the hypotheses. Analogue to *Step 3*, as this is part of the Bayesian prior specification, it is recommended to fully report the available information about the parameter and why it leads to the prior probabilities of the hypotheses.

Step E: Loss Function. Specify the (interval-valued) loss function by following *Steps 4, 6, and 7*.

Step F: Posterior Odds. Use the available Bayes factor *BF* to calculate the posterior odds via equation (15).

Step G: Optimal Action. The optimal action can be derived as in *Step 10*.

[3]Comprehensive information about open science are provided e.g. by the LMU Open Science Center: `www.osc.uni-muenchen.de`.

8 Concluding Remarks

Statisticians and methodologists do – in general – not know all the different fields of applications and research contexts a statistical method will eventually be employed in. The scientific endeavor is extremely versatile and research problems might arise that have not been thought of before. In that, versatility of research methods is of paramount importance. While it might be considered as disadvantageous that frequentist hypothesis tests are restricted to a dichotomous decision context, it might similarly be considered as disadvantageous if Bayes factors are restricted to only an evidential, non-decision context. Fortunately, the mathematics underlying Bayes factors suggest their involvement in guiding decisions. In this regard, Bayes factors might be seen as evidential quantification or as a quantity in the context of guiding decisions, depending on the goals of the scientific investigation.

In order to use Bayes factors correctly when guiding decision, their embedding within the framework of Bayesian decision theory has to be considered. It is important that the research context as well as the decision problem are formalized appropriately. If misspecified, results inform past the research question. Naturally, the specification of essential quantities (such as the prior distribution, the hypotheses, or the loss function) is an applied problem and might be rather difficult for the applied scientist. In order to alleviate these issues, these quantities might be specified robustly as interval-valued or set-valued quantities. Then the researcher might consider a range or a set of plausible values, avoiding the necessity to (arbitrarily) commit to one single precise value. Within this elaboration only an interval-valued loss value was considered, as it keeps the calculations simple (compare Section 5) yet allows to include essential loss information (about the consequences of the decision) into the analysis. Naturally, also the prior distribution and the hypotheses might be included into the analysis as set-valued quantities (see e.g. Ebner et al., 2019). How to deal with set-valued quantities on a formal level is extensively elaborated on within the field of imprecise probabilities (see e.g. Walley, 1991; Augustin et al., 2014; Huntley et al., 2014).

In summary, this elaboration delineates the decision theoretic embedding of Bayes factors by outlining the framework of hypothesis-based Bayesian decision theory, supplemented by considerations about robust loss specifications and straightforward step-by-step guides. These guides try to help those applied scientists who want to guide decisions with Bayes factors.

References

Augustin T., Coolen F.P., de Cooman G., and Troffaes M.C.M., editors (2014). *Introduction to Imprecise Probabilities*. John Wiley, Chichester. URL `http://dx.doi.org/10.1002/9781118763117`.

Bartolucci A.A., Tendera M., and Howard G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *The American Journal of Cardiology*, 107(12):1796–1801. URL `http://dx.doi.org/10.1016/j.amjcard.2011.02.325`.

Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition. URL `http://dx.doi.org/10.1007/978-1-4757-4286-2`.

Cumming G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1):7–29. URL http://dx.doi.org/10.1177/0956797613504966.

Ebner L., Schwaferts P., and Augustin T. (2019). Robust Bayes factor for independent two-sample comparisons under imprecise prior information. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 167–174. PMLR. URL http://proceedings.mlr.press/v103/ebner19a.html.

Gelman A. and Hennig C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180:967–1033. URL http://dx.doi.org/10.1111/rssa.12276.

Gelman A., Stern H.S., Carlin J.B., Dunson D.B., Vehtari A., and Rubin D.B. (2013). *Bayesian Data Analysis*. Chapman & Hall. URL http://dx.doi.org/10.1201/9780429258411.

Gönen M., Johnson W.O., Lu Y., and Westfall P.H. (2005). The Bayesian two-sample t test. *The American Statistician*, 59:252–257. URL http://dx.doi.org/10.1198/000313005X55233.

Huntley N., Hable R., and Troffaes M.C.M. (2014). Decision making. In T. Augustin, F.P.A. Coolen, G. de Cooman, and M.C.M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 190–206. John Wiley & Sons. URL http://dx.doi.org/10.1002/9781118763117.ch8.

Jeffreys H. (1961). *Theory of Probability*. Oxford University Press, Oxford, third edition.

Kass R.E. and Raftery A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795. URL http://dx.doi.org/10.2307/2291091.

Klein O., Hardwicke T.E., Aust F., Breuer J., Danielsson H., Mohr A.H., IJzerman H., Nilsonne G., Vanpaemel W., Frank M.C., and Frank M.C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, 4(1):20(1–15). URL http://dx.doi.org/10.1525/collabra.158.

Kruschke J.K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan*. Academic Press, New York. URL http://dx.doi.org/10.1016/B978-0-12-405888-0.09999-2.

Kruschke J.K. and Liddell T.M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206. URL http://dx.doi.org/10.3758/s13423-016-1221-4.

Morey R.D., Romeijn J.W., and Rouder J.N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18. URL http://dx.doi.org/10.1016/j.jmp.2015.11.001.

Nosek B.A., Ebersole C.R., DeHaven A.C., and Mellor D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606. URL http://dx.doi.org/10.1073/pnas.1708274114.

Ríos Insua D. and Ruggeri F., editors (2012). *Robust Bayesian Analysis*. Springer Science & Business Media. URL http://dx.doi.org/10.1007/978-1-4612-1306-2.

Robert C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, second edition. URL http://dx.doi.org/10.1007/0-387-71599-1.

Rouder J.N., Haaf J.M., and Vandekerckhove J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1):102–113. URL http://dx.doi.org/10.3758/s13423-017-1420-7.

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

Schwaferts P. and Augustin T. (2019). Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In J. De Bock, C.P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probability: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 338–345. PMLR. URL http://proceedings.mlr.press/v103/schwaferts19a.html.

Schwaferts P. and Augustin T. (2020). Bayesian decisions using regions of practical equivalence (ROPE): Foundations. Technical Report 235, Ludwig-Maximilians-University Munich, Department of Statistics. URL http://dx.doi.org/10.5282/ubm/epub.74222.

Schwaferts P. and Augustin T. (2021). Imprecise hypothesis-based Bayesian decision making with composite hypotheses. In A. Cano, J. De Bock, E. Miranda, and S. Moral, editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, page 280–288. PMLR. URL https://proceedings.mlr.press/v147/schwaferts21a.html.

Walley P. (1991). *Statistical Reasoning With Imprecise Probabilities*. Chapman & Hall, London.

Contribution 8

Schwaferts & Augustin (2020): Bayesian Decisions Using Regions of Practical Equivalence (ROPE): Foundations. (Technical Report)

Patrick Schwaferts
Thomas Augustin

Bayesian Decisions using Regions of Practical Equivalence (ROPE): Foundations

Technical Report Number 235, 2020
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Bayesian Decisions using Regions of Practical Equivalence (ROPE): Foundations

Patrick Schwaferts Thomas Augustin

Ludwig-Maximilians-Universität Munich
Department of Statistics
Methodological Foundations of Statistics and its Applications
Ludwigsstraße 33, 80539 Munich, Germany

Abstract

Kruschke [2018] proposes the so called HDI+ROPE decision rule about accepting or rejecting a parameter null value for practical purposes using a region of practical equivalence (ROPE) around the null value and the posterior highest density interval (HDI) in the context of Bayesian statistics. Further, he mentions the so called ROPE-only decision rule within his supplementary material, which is based on ROPE, but uses the full posterior information instead of the HDI.

Of course, if it is about formalizing and guiding decisions then statistical decision theory is the framework to rely on, and this technical report elaborates the decision theoretic foundations of both decision rules.

It appears that the foundation of the HDI+ROPE decision rule is rather artificial compared to the foundation of the ROPE-only decision rule, such that latter might be characterized as being closer to the underlying practical purpose than former. Still, the ROPE-only decision rule employs a truly arbitrary, and thus debatable, choice of loss values.

Keywords: Bayesian Decision Theory, Region of Practical Equivalence, ROPE, HDI+ROPE, ROPE-only, Imprecise Probabilities

1 Introduction

When it comes to applying statistics, there is an increased awareness that black-and-white thinking might lead to severe issues within the process of science, and thus binary decisions should be treated with caution [see e.g. Kruschke, 2018]. Reporting estimates together with the uncertainty about them might be seen as a fruitful alternative [see e.g. Cumming, 2014]. However, sometimes a decision is necessary and the use of statistical decision theory [see e.g. Berger, 1995, Robert, 2007] suggests itself. In that regard, every proposed or employed decision rule might be assessed on the basis of its decision theoretic foundation.

Kruschke [e.g. 2015, 2018] proposes a decision rule based on posterior highest density intervals (HDI) and regions of practical equivalence (ROPE).

A $(1-\alpha)$ highest density interval for a certain distribution of a parameter (prior or poste-

rior) is an interval[1] that contains all parameter values with the highest probability densities and integrates to a probability of $1 - \alpha$. Kruschke [2018] employs $(1 - \alpha) = 0.95$ and uses the posterior distribution when referring to a highest density interval (HDI; also referred to as highest posterior density (HPD) interval), which will be adopted within this technical report.

A region of practical equivalence (ROPE) refers to a certain parameter value of interest, which might also be called "null value" and frequently (but not necessarily) the parameter value of interest is zero. A ROPE for a null value is a "range of parameter values that are equivalent to the null value for practical purposes" [Kruschke, 2018, p. 272]. Accordingly, "the limits of the ROPE depend on the practical purpose of the ROPE. If the purpose is to assess the equivalence of drug-treatment outcomes, then the ROPE limits depend on the real-world costs and benefits of the treatment and the ability to measure the outcome" [Kruschke, 2015, p. 338].

Once the ROPE is specified (before observing the data) and the HDI is calculated (after observing the data), the decision rule by Kruschke [2018, p. 272], referred to as HDI+ROPE decision rule, is as follows:

- If the HDI falls completely inside the ROPE, then accept the null value for practical purposes.
- If the HDI falls completely outside the ROPE, then reject the null value for practical purposes.
- Else, withhold a decision.

In addition to the HDI+ROPE decision rule, Kruschke [2018, supp. p. 5] mentions another exemplary decision rule within his supplementary material[2] that is based on the ROPE alone and considers the posterior distribution instead of the HDI. Referred to as ROPE-only decision rule, it states:

- If more than 95% of the posterior distribution fall within the ROPE, then accept the null value for practical purposes.
- If less than 5% of the posterior distribution fall within the ROPE, then reject the null value for practical purposes.
- Else, withhold a decision.

Within his supplementary material, Kruschke [2018, supp. p. 3–5] delineates preliminary ideas about the decision theoretic foundation of the HDI+ROPE decision rule. In addition, a foundation of the ROPE-only decision rule is also pending. In that, the purpose of this

[1] Certainly, it might be possible that a HDI is a set of parameters, which is not an interval, however, in accordance with Kruschke [2018], these cases are not considered within this technical report.

[2] This technical report is based on the supplementary material Version 1 of February 25, 2018, available at the Open Science Framework with the url `https://osf.io/jwd3t/` and downloaded at August 18, 2020.

technical report is to elaborate the decision theoretic foundations of both decision rules more profoundly.

Therefore, Bayesian decision theory is briefly recalled in Section 2, before outlining the foundations of the ROPE-only decision rule (Section 3) and of the HDI+ROPE decision rule (Section 4). A concluding discussion in Section 5 compares the foundations of both decision rules w.r.t. their interpretation and connection to the underlying real-world decision.

2 Recall of Bayesian Decision Theory

The observed data $\boldsymbol{x} \in \mathcal{X}$, where the sample space $\mathcal{X}$ comprises all potential data sets, are modeled parametrically as realization of the random quantity X with density $f(\boldsymbol{x}|\theta)$, where $\theta \in \Theta$ is a real-valued parameter and Θ the parameter space.

Within a Bayesian context, there is a prior distribution with density $\pi(\theta)$ on the parameter θ, which gets updated via Bayes formula to the posterior distribution with density $\pi(\theta|\boldsymbol{x})$ once the data $\boldsymbol{x}$ are observed.

In the context of an applied decision, one of different potential actions $\mathsf{a} \in \mathcal{A}$ should be selected, where $\mathcal{A}$ denotes the action space.

Deciding for a certain action $\mathsf{a} \in \mathcal{A}$ if a certain parameter value $\theta \in \Theta$ is true has consequences and the "badness" of these consequences is formally captured by a loss[3] function

$$L : \Theta \times \mathcal{A} \to \mathbb{R}_0^+ \,. \tag{1}$$

Naturally, the meaning of this "badness" is comparative and can only be judged w.r.t. the loss values of other actions and parameter values, yet their comparative meaning should reflect the characteristics within the applied real-world decision. However, it might be rather difficult[4] to specify an exact loss function that matches those characteristics, which are usually accessible only vaguely. As a solution, the loss function might be simplified using ROPE (see Section 3.1) and specified in an imprecise manner (see Section 3.2).

As the posterior density $\pi(\theta|\boldsymbol{x})$ is available within a Bayesian analysis, it is possible to calculate the expected posterior loss of each action $\mathsf{a} \in \mathcal{A}$

$$\rho : \mathcal{A} \to \mathbb{R}_0^+ \tag{2}$$

by

$$\rho(\mathsf{a}) = \int_\Theta L(\theta, \mathsf{a}) \pi(\theta|\boldsymbol{x}) d\theta \,. \tag{3}$$

[3] Sometimes decision theory is depicted with a utility function instead of a loss function, which quantifies the "utility" instead of the "badness" of the respective consequences.

[4] Of course, there are situations in which a loss function might be specified exactly, as e.g. some special cases in economy in which the loss might be related to monetary outcomes or obtained from preferences [see e.g. Berger, 1995, ch. 2.2]. In many research situations, however, the necessary information to do so might not be available.

Intuitively and following the conditional Bayes principle [see e.g. Berger, 1995], the action a with minimal expected posterior loss $\rho(\mathsf{a})$ should be chosen and is called (posterior loss[5]) Bayes action.

Taken together, all three quantities $\pi(\theta)$, $\boldsymbol{x}$, and L are required to find the (optimal) Bayes action.

Before observing the data, only the prior density $\pi(\theta)$ and the loss function L are available. Therefore, it is possible to consider each potentially observable data set $\boldsymbol{x} \in \mathcal{X}$ and evaluate which action $\mathsf{a} \in \mathcal{A}$ would be the corresponding Bayes action. This is formally captured by a decision rule

$$\delta : \mathcal{X} \to \mathcal{A}\,. \tag{4}$$

In the context of the conditional Bayes principle, the optimal decision rule has the following shape

$$\delta(\boldsymbol{x}) = \underset{\mathsf{a}\in\mathcal{A}}{\arg\min}\, \rho(\mathsf{a}) = \underset{\mathsf{a}\in\mathcal{A}}{\arg\min} \int_{\Theta} L(\theta, \mathsf{a})\pi(\theta|\boldsymbol{x})d\theta \tag{5}$$

and states the Bayes action for each potential data set.

Of course, it is possible to formulate other decision rules, but these might not find the Bayes action for each data set. In that, refer to the decision rule in equation (5), which matches every data set with the corresponding Bayes action, as Bayes rule[6] (not to be confused with Bayes formula for calculating $\pi(\theta|\boldsymbol{x})$ from $f(\boldsymbol{x}|\theta)$ and $\pi(\theta)$).

Note that a Bayes action and a Bayes rule always refer to a certain loss function. With a different loss function a different decision rule might be a Bayes rule and a different action might be a Bayes action for a given data set.

[5] Within this technical report, the term "Bayes action" always refers to a posterior loss Bayes action.

[6] This depiction of a Bayes rule as minimizing the expected posterior loss is based on one of the fundamental theorems of Bayesian decision theory [c.p. e.g. Berger, 1995, p. 159 Result 1]. In general, the definition of a Bayes rule might involve the minimization of the prior risk (which considers all potentially observable data sets) and this theorem states equivalence with minimizing the expected posterior loss. In anticipation of Section 3, this theorem, however, might not necessarily hold within the framework of imprecise probabilities in general and counterexamples involve imprecisely specified probabilities [Augustin, 2003]. Yet the involvement of the framework of imprecise probabilities within this technical report comprises only an imprecisely stated loss function (and no imprecisely specified probabilities, see Section 3.2), so that an equivalence analogue to this fundamental theorem should hold within the context depicted here. This should be addressed in further research. In any case, a Bayesian analysis typically sticks to a conditional point of view that conditions on the actually observed data and does not consider other potential data sets, which were not observed. In that, Berger [1995, p. 160, notation adapted, italics preserved] reasons:

> Note that, from the conditional perspective together with the utility development of the loss, the *correct* way to view the situation is that of minimizing [$\rho(\mathbf{a})$]. One should condition on what is known, namely [$\boldsymbol{x}$] (...), and average the utility over what is unknown, namely θ. The desire to minimize [the prior risk] would be deemed rather bizarre from this perspective.

In summary, even if this fundamental theorem might not hold within the context employed here to depict the foundation of the ROPE-only rule (see Section 3), the depicted approach still appears to be reasonable.

3 Foundations of the ROPE-only decision rule

3.1 ROPE as Simplification

As made obvious by the quotes about the ROPE in Section 1, the ROPE cannot be separated from the underlying practical purposes. Implied by both the HDI+ROPE decision rule and the ROPE-only decision rule, the practical purpose is to decide between two actions a_0 and a_1. The first action a_0 is in accordance with the null value $\theta_0 \in \Theta$ and the second is in discordance with the null value θ_0. Accordingly, indicated by subscript P for "practical purpose", the action space $\mathcal{A}_P = \{\mathsf{a}_0, \mathsf{a}_1\}$ comprises these two actions of the practical purpose.

Kruschke [2018, p. 272] refers to these actions as "accept the null value for practical purposes" (a_0) and "reject the null value for practical purposes" (a_1). However, we want to refrain from using this terminology, because it tempts to ignore the actual real-world decision and to derive conclusions about actions that might not even be specified. Instead, we highly recommend to explicitly state the actions of interest, such that the real-world decision of interest might be formalized properly.

The corresponding loss function $L_P : \Theta \times \mathcal{A}_P \to \mathbb{R}_0^+$ quantifies the "badness" of each of those two practical actions under each parameter. With this loss function it would be possible to determine the Bayes action for the observed data set, however, the exact shape of this loss function L_P is hardly accessible in real life. Therefore, a way to deal with this issue is necessary and a first approach might be to simplify this loss function. Considerations in the context of ROPE lead to such a simplification.

By construction, under the null value θ_0 the loss of a_0 is smaller than the loss of a_1, i.e. $L_P(\theta_0, \mathsf{a}_0) < L_P(\theta_0, \mathsf{a}_1)$, as former action is in accordance and latter action in discordance with the null value.

If not specifying the exact values of the loss function L_P, the researcher might (or should) still be able to determine the appropriate action for each parameter value $\theta \in \Theta$. In that, there is a set Θ_0 of parameter values for which a_0 is appropriate (containing the null value θ_0) and a set $\Theta_1 = \Theta \setminus \Theta_0$ of the remaining parameter values for which a_1 is appropriate. The first set Θ_0 is the ROPE and usually an interval.

However, different parameter values within these sets, respectively, might still have different loss values. As these exact values are still hardly accessible in real life, a possible simplification is to treat each parameter value within the ROPE Θ_0 as "equivalent to the null value for practical purposes" [Kruschke, 2018, p. 272], i.e. assuming identical loss values for parameters within Θ_0:

$$\forall \theta_i, \theta_j \in \Theta_0 \, \forall \mathsf{a} \in \mathcal{A}_P : L_P(\theta_i, \mathsf{a}) = L_P(\theta_j, \mathsf{a}) \,. \tag{6}$$

In addition to the paramete values within the ROPE Θ_0, also the parameter values outside the ROPE, i.e. within Θ_1, might be treated as equivalent for practical purposes by

Table 1: Simplified loss function for the actions of the practical purpose using a regret form.

$L_P(\theta, \mathsf{a})$	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\mathsf{a} = \mathsf{a_0}$	0	k_0
$\mathsf{a} = \mathsf{a_1}$	k_1	0

employing identical loss values:

$$\forall \theta_i, \theta_j \in \Theta_1 \, \forall \mathsf{a} \in \mathcal{A}_P : L_P(\theta_i, \mathsf{a}) = L_P(\theta_j, \mathsf{a}) \, . \tag{7}$$

In that, this simplified loss function needs only four values to be specified and, without loss of generality, a regret form might be employed, in which $\mathsf{a_0}$ and $\mathsf{a_1}$ have zero loss if $\theta \in \Theta_0$ and $\theta \in \Theta_1$, respectively. The remaining two loss values shall be denoted by (see Table 1)

$$\begin{aligned} k_0 &:= L_P(\theta, \mathsf{a_0}) \quad \forall \theta \in \Theta_1 \\ k_1 &:= L_P(\theta, \mathsf{a_1}) \quad \forall \theta \in \Theta_0 \, . \end{aligned}$$

Using this simplification, the expected posterior loss of each action $\mathsf{a} \in \mathcal{A}_P$ is

$$\begin{aligned} \rho(\mathsf{a_0}) &= \int_\Theta L_P(\theta, \mathsf{a_0}) \pi(\theta|\boldsymbol{x}) d\theta \\ &\overset{\substack{eq. \\ (6) \\ (7)}}{=} 0 \cdot \int_{\theta \in \Theta_0} \pi(\theta|\boldsymbol{x}) d\theta + k_0 \cdot \int_{\theta \in \Theta_1} \pi(\theta|\boldsymbol{x}) d\theta \\ &= k_0 \cdot p(\theta \in \Theta_1 | \boldsymbol{x}) \end{aligned}$$

and analogously

$$\rho(\mathsf{a_1}) = k_1 \cdot p(\theta \in \Theta_0 | \boldsymbol{x}) \, . \tag{8}$$

With $k := k_1/k_0$, the ratio of expected posterior losses is

$$\varrho(k) := \frac{\rho(\mathsf{a_1})}{\rho(\mathsf{a_0})} = \frac{k_1 \cdot p(\theta \in \Theta_0 | \boldsymbol{x})}{k_0 \cdot p(\theta \in \Theta_1 | \boldsymbol{x})} = k \cdot \frac{p(\theta \in \Theta_0 | \boldsymbol{x})}{p(\theta \in \Theta_1 | \boldsymbol{x})} \tag{9}$$

and the corresponding Bayes action is

$$\mathsf{a}_{\mathsf{Bayes}}(k) = \begin{cases} \mathsf{a_0} & \text{if } \varrho(k) > 1 \\ \mathsf{a_1} & \text{if } \varrho(k) < 1 \end{cases} \, . \tag{10}$$

If $\varrho(k) = 1$, then either action might be chosen.

The term $p(\theta \in \Theta_0 | \boldsymbol{x}) / p(\theta \in \Theta_1 | \boldsymbol{x})$ can be calculated simply from the posterior density $\pi(\theta|\boldsymbol{x})$, however, k need to be specified w.r.t. to the practical purpose.

3.2 Framework of Imprecise Probabilities

Specifying a precise value for k, which defines the simplified loss function, might still be difficult for applied scientists and ideas from the framework of imprecise probabilities [Walley, 1991] come in handy. In addition, the foundations of the ROPE-only decision rule can be depicted elegantly within this framework.

In general, this framework is based on the fact that there is more to uncertainty than can be captured within precise probability values [e.g. Ellsberg, 1961, Levi, 1980, Walley, 1991, Etner et al., 2012]. As a solution, sets or intervals of probability values, so called imprecise probabilities, are employed instead of single precise probability values. These intervals are treated as an entity of its own [c.p. Walley, 1991] and numerous sources on how to calculate with imprecise probabilities are available [see e.g. Augustin et al., 2014, for a depiction of the state of the art within different fields of application at that time]. Naturally, this framework is appropriate whenever some relevant but potentially vague information about probabilities is available, yet it is not enough to unambiguously specify exact probability values. For example, within the Bayesian context, a researcher might be unable to specify the exact shape of a prior distribution and several different distributions are in accordance with the available prior knowledge. By comprising all these potential distributions within a set of distributions, the researcher obtains an imprecise prior distribution, which reflects the available knowledge and uncertainty as is, without pretending a level of precision that is not available [see also the framework of robust Bayesian statistics, e.g. Ríos Insua and Ruggeri, 2012].

Similarly, in the context of a real-world decision, some potentially vague information about potential consequences is supposed to be available. Yet, an applied scientist is usually unable to unambiguously specify a precise loss function as several different loss functions might be in accordance with the available (vague) information. An arbitrary specification of a loss function will result in an arbitrary decision. Not employing a loss function at all, on the other hand, leads to a decision that lacks a relation to the underlying real-world situation and is therefore arbitrary as well. In that, it seems obvious that partially available information about the loss function has to be included into the analysis in the form it is available.

Thus, analogue to imprecisely specified probabilities, the loss function might be specified imprecisely. In the context of the simplified loss function as depicted in Section 3.1, instead of a precise value k, an open[7] interval of values $K = \left(\underline{K}, \overline{K}\right)$ might be employed, where $\underline{K}$ and $\overline{K}$ denote the lower and upper bound, respectively, for stating how much "worse" $\mathsf{a_1}$ would be if $\theta \in \Theta_0$ than $\mathsf{a_0}$ would be if $\theta \in \Theta_1$ (if deciding correctly has zero "badness"). As every value $k \in K$ defines a different (simplified) loss function, the interval K defines a set of loss functions. For each of those loss functions, i.e. for every $k \in K$, it is possible

[7] Of course, K might also be specified by a closed interval $\left[\underline{K}, \overline{K}\right]$ and in many situations this might be more reasonable. However, in order to derive the ROPE-only decision rule, as stated by Kruschke [2018] within his supplementary material, K needs to be an open interval.

to determine the Bayes action $\mathsf{a}_{\mathsf{Bayes}}(k)$, once the data are available and the posterior distribution of θ is calculated. If the Bayes action $\mathsf{a}_{\mathsf{Bayes}}(k)$ is the same for all $k \in K$, then this action should be selected, else information is lacking to unambiguously guide a decision and a decision should be withheld.

Formally, an interval-valued ratio of expected posterior losses

$$\big(\varrho(\underline{K}), \varrho(\overline{K})\big) \tag{11}$$

is obtained by considering the interval-valued K, leading to the Bayes action

$$\mathsf{a}_{\mathsf{Bayes}}(K) = \begin{cases} \mathsf{a}_0 & \text{if } \varrho(\underline{K}) > 1 \\ \mathsf{a}_1 & \text{if } \varrho(\overline{K}) < 1 \end{cases} . \tag{12}$$

For $\varrho(\underline{K}) \leq 1 \leq \varrho(\overline{K})$, the decision should be withheld.

3.3 An Arbitrary Choice

By setting K arbitrarily to $K = (1/19, 19)$, the ROPE-only decision rule is obtained, because – according to the imprecise decision theoretic framework, especially considering equation (12) – action a_0 ("accept the null value for practical purposes") is optimal if

$$\begin{aligned} & \varrho(1/19) > 1 \\ \Leftrightarrow \quad & \frac{1}{19} \cdot \frac{p(\theta \in \Theta_0|\boldsymbol{x})}{p(\theta \in \Theta_1|\boldsymbol{x})} > 1 \\ \Leftrightarrow \quad & p(\theta \in \Theta_0|\boldsymbol{x}) > 19 \cdot p(\theta \in \Theta_1|\boldsymbol{x}) \\ \Leftrightarrow \quad & p(\theta \in \Theta_0|\boldsymbol{x}) > 19 \cdot (1 - p(\theta \in \Theta_0|\boldsymbol{x})) \\ \Leftrightarrow \quad & 20 \cdot p(\theta \in \Theta_0|\boldsymbol{x}) > 19 \\ \Leftrightarrow \quad & p(\theta \in \Theta_0|\boldsymbol{x}) > 0.95 \end{aligned}$$

and, analogously, action a_1 ("reject the null value for practical purposes") is optimal if

$$\begin{aligned} & \varrho(19) < 1 \\ \Leftrightarrow \quad & p(\theta \in \Theta_0|\boldsymbol{x}) < 0.05\,, \end{aligned}$$

which reflect exactly those conditions defining the ROPE-only decision rule.

In any other case, i.e. for $0.05 \leq p(\theta \in \Theta_0|\boldsymbol{x}) \leq 0.95$, both the imprecise decision theoretic framework using $K = (1/19, 19)$ and the ROPE-only decision rule recommend to withhold a decision.

4 Foundations of the HDI+ROPE decision rule

4.1 Action Space and Decision Rule

The general idea of the decision theoretic foundation of the HDI+ROPE decision rule was described within the supplementary material by Kruschke [2018, supp. p. 3–5]. However, some aspects depicted there are merely preliminary[8], so this technical report intends to outline this foundation more profoundly. In line with this idea and the considerations depicted by Rice et al. [2008] (which are also referred to by Kruschke [2018]), the corresponding action in the context of the HDI+ROPE decision rule comprises two aspects:

- the determination of the HDI and
- the assessment of the relation between the HDI and the ROPE (inside, outside, or overlap).

The action space w.r.t. the first aspect – indicated with subscript I for "interval" – contains all possible closed parameter intervals

$$\mathcal{A}_I = \{[a, b] | a, b \in \Theta, a < b\} \tag{13}$$

and the objective is to decide for the element within $\mathcal{A}_I$ that is the HDI.

The action space w.r.t. the second aspect – indicated with subscript R for "relation" – contains all three possible relations between a parameter interval and a predefined ROPE:

$$\mathcal{A}_R = \{\mathsf{r}_0, \mathsf{r}_1, \mathsf{r}_2\} \tag{14}$$

with

r_0: The parameter interval falls completely within the ROPE.

r_1: The parameter interval falls completely outside the ROPE.

r_2: The parameter interval and the ROPE overlap.

In conjunction, the overall action space is $\mathcal{A}_I \times \mathcal{A}_R$ and the corresponding decision rule maps the sample space $\mathcal{X}$ to this action space:

$$\delta_{HDI+ROPE} : \mathcal{X} \to \mathcal{A}_I \times \mathcal{A}_R \,. \tag{15}$$

[8] Within equation (1) on page 4 within Kruschke [2018]'s supplementary material, the argument s of the function $\mathbf{1}(s)$ is sometimes a set, yet it should be a statement. The explanation of one of the terms states "cost of reject if HDI overlaps ROPE" [Kruschke, 2018, supp. p. 4 eq. (1)], yet the term might rather refer to a cost of rejection if the HDI is within the ROPE. As outlined within this technical report (see esp. equation (30)), the cost of deciding correctly should be identical for each relation between HDI and ROPE, which is not necessarily the case in Kruschke's formula (1).

In that regard, Kruschke [2018] states that his ideas are "merely suggestive" [supp. p. 4] and his "goal is [only] to point out that formal expressions are possible for the loss implicit to the intuitive HDI+ROPE rule" [supp. p. 5]. In that, the elaborations within this technical report are based on this initial work by Kruschke [2018].

The exact shape of this decision rule

$$\delta_{HDI+ROPE}(\boldsymbol{x}) = \begin{pmatrix} \delta_I(\boldsymbol{x}) \\ \delta_R(\delta_I(\boldsymbol{x})) \end{pmatrix} \tag{16}$$

can be depicted using the functions

$$\delta_I : \mathcal{X} \to \mathcal{A}_I \,, \tag{17}$$

which maps the data $\boldsymbol{x}$ to the corresponding HDI, and

$$\delta_R : \mathcal{A}_I \to \mathcal{A}_R \,, \tag{18}$$

which maps an interval in parameter space $[a, b] \in \mathcal{A}_I$ to its correct relation with a predefined ROPE Θ_0 by

$$\delta_R([a,b]) = \begin{cases} \mathsf{r_0} & \text{if } [a,b] \cap \Theta_0 = [a,b] \\ \mathsf{r_1} & \text{if } [a,b] \cap \Theta_0 = \emptyset \\ \mathsf{r_2} & \text{if } [a,b] \cap \Theta_0 \neq [a,b] \wedge [a,b] \cap \Theta_0 \neq \emptyset \end{cases} . \tag{19}$$

4.2 Loss Function

4.2.1 Determination of HDI

It is possible to state a loss function for which the determination of the HDI δ_I is a Bayes rule, namely [see e.g. Schervish, 1995, Rice et al., 2008]

$$L_I : \Theta \times \mathcal{A}_I \to \mathbb{R}_0^+ : L_I(\theta, [a,b]) = (b-a) + c \cdot \mathbf{1}(\theta \notin [a,b]) \,, \tag{20}$$

where $\mathbf{1}(s) = 1$ if the statement s is true and $\mathbf{1}(s) = 0$ if s is false. The value c denotes a constant which determines the minimum density of a parameter to be included within the HDI (see below).

The expected posterior loss w.r.t. this loss function is

$$\begin{aligned} \rho_I([a,b]) &= \int_\Theta L_I(\theta, [a,b]) \pi(\theta|\boldsymbol{x}) d\theta \\ &= \int_\Theta \left[(b-a) + c \cdot \mathbf{1}(\theta \notin [a,b])\right] \pi(\theta|\boldsymbol{x}) d\theta \\ &= (b-a) + c \int_\Theta \mathbf{1}(\theta \notin [a,b]) \pi(\theta|\boldsymbol{x}) d\theta \\ &= (b-a) + c \int_{\Theta \setminus [a,b]} \pi(\theta|\boldsymbol{x}) d\theta \end{aligned}$$

and minimizing this expected posterior loss over the action space $\mathcal{A}_I$ yields as Bayes action the interval $[a, b]$ that contains all parameters with posterior density larger than c^{-1} [see e.g. Schervish, 1995, Rice et al., 2008]. By setting c appropriately, the 95%-HDI is obtained as Bayes action for a given data set $\boldsymbol{x}$ and the decision rule δ_I is a Bayes rule w.r.t. the loss function L_I.

4.2.2 Relation between HDI and ROPE

It is also possible to state a loss function L_R for which the assessment of the relation between a parameter interval $[a, b]$ and a predefined ROPE Θ_0 is a Bayes rule.

As outlined in Section 2, a loss function is defined on the parameter space Θ and on the action space, which is $\mathcal{A}_R$ (as defined in equation (14)) within this context. However, the employed loss function will depend on the ROPE Θ_0 and the parameter interval $[a, b]$ as well. Although the ROPE might be treated as given, this is not the case for the parameter interval $[a, b]$, especially when considering the overall decision rule $\delta_{HDI+ROPE}$ (as in the following Section 4.2.3). Accordingly, this dependence of L_R on $[a, b] \in \mathcal{A}_I$ needs to be taken into account, so that

$$L_R : \Theta \times \mathcal{A}_R \times \mathcal{A}_I \to \mathbb{R}_0^+ : (\theta, \mathsf{r}, [a, b]) \mapsto L_R^{[a,b]}(\theta, \mathsf{r})\,. \tag{21}$$

Considering δ_R in isolation, as within this subsection, also $[a, b]$ might be treated as given.

Although this loss function is technically defined using the parameter space Θ, this dependence is not necessary:

$$\forall \mathsf{r} \in \mathcal{A}_R : \forall \theta_i, \theta_j \in \Theta : L_R^{[a,b]}(\theta_i, \mathsf{r}) = L_R^{[a,b]}(\theta_j, \mathsf{r}) =: L_R^{[a,b]}(\mathsf{r})\,. \tag{22}$$

A candidate of this loss function is depicted in Table 2, formally stated as

$$\begin{aligned}
L_R^{[a,b]}(\mathsf{r}) = {} & c_1 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_0) \cdot \mathbf{1}([a, b] \cap \Theta_0 = [a, b]) + \\
& c_1 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_1) \cdot \mathbf{1}([a, b] \cap \Theta_0 = \emptyset) + \\
& c_1 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_2) \cdot \mathbf{1}([a, b] \cap \Theta_0 \neq [a, b] \wedge [a, b] \cap \Theta_0 \neq \emptyset) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_0) \cdot \mathbf{1}([a, b] \cap \Theta_0 = \emptyset) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_0) \cdot \mathbf{1}([a, b] \cap \Theta_0 \neq [a, b] \wedge [a, b] \cap \Theta_0 \neq \emptyset) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_1) \cdot \mathbf{1}([a, b] \cap \Theta_0 = [a, b]) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_1) \cdot \mathbf{1}([a, b] \cap \Theta_0 \neq [a, b] \wedge [a, b] \cap \Theta_0 \neq \emptyset) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_2) \cdot \mathbf{1}([a, b] \cap \Theta_0 = [a, b]) + \\
& c_2 \cdot \mathbf{1}(\mathsf{r} = \mathsf{r}_2) \cdot \mathbf{1}([a, b] \cap \Theta_0 = \emptyset)\,,
\end{aligned} \tag{23}$$

where c_1, c_2 are arbitrary positive constants with $c_1 < c_2$ and, again, $\mathbf{1}(s) = 1$ if the statement s is true and $\mathbf{1}(s) = 0$ if s is false.

As L_R does not depend on the parameter θ, the expected posterior loss ρ_R of each action

Table 2: Loss function for finding the relation r between a parameter interval and a ROPE (see equation 23).

$L_R^{[a,b]}(\mathsf{r})$		Interval to ROPE		
		within	outside	overlap
Decision	$\mathsf{r} = \mathsf{r}_0$	c_1	c_2	c_2
	$\mathsf{r} = \mathsf{r}_1$	c_2	c_1	c_2
	$\mathsf{r} = \mathsf{r}_2$	c_2	c_2	c_1

$\mathsf{r} \in \mathcal{A}_R$ w.r.t. this loss function is the loss value itself:

$$\begin{aligned}\rho_R(\mathsf{r}) &= \int_\Theta L_R^{[a,b]}(\theta, \mathsf{r})\pi(\theta|\boldsymbol{x})d\theta \\ &= \int_\Theta L_R^{[a,b]}(\mathsf{r})\pi(\theta|\boldsymbol{x})d\theta \\ &= L_R^{[a,b]}(\mathsf{r}) \int_\Theta \pi(\theta|\boldsymbol{x})d\theta \\ &= L_R^{[a,b]}(\mathsf{r}) \end{aligned} \tag{24}$$

Minimizing ρ_R over the action space $\mathcal{A}_R$ results in the relation r that is obtained by the decision rule δ_R in equation (19). In that, this decision rule δ_R is a Bayes rule w.r.t. the loss function L_R for all parameter intervals $[a, b] \in \mathcal{A}_I$.

The loss function L_R was defined using only two different values c_1 and c_2. In that, two restrictions are imposed on the loss function L_R:

(I) Deciding correctly has the same loss c_1 independent of which relation is true.

(II) Deciding falsely has the same loss c_2 independent of which relation is true and which incorrect relation was chosen.

Of course, it would be possible to employ a loss function without these restrictions that uses e.g. nine different loss values instead of only two. However, the first restriction (I) will be necessary for combining both loss functions L_I and L_R, because then, independent of the parameter value $\theta \in \Theta$, for every potential interval $[a, b] \in \mathcal{A}_I$ the decision rule δ_R yields a relation $\mathsf{r} \in \mathcal{A}_R$ with the identical loss c_1, i.e.

$$\forall\theta \in \Theta \;\; \forall[a, b] \in \mathcal{A}_I : L_R^{[a,b]}(\theta, \delta_R([a, b])) = c_1 \,, \tag{25}$$

a fact that will be referred to later.

The second restriction (II) is employed both out of convenience and to emphasize an important characteristic: Assume that up to six different values larger than c_1 would be employed instead of c_2 within the loss function L_R in equation (23). Still, the expected posterior loss $\rho_R(\mathsf{r})$ of each action r is the loss value $L_R^{[a,b]}(\mathsf{r})$ itself (equation (24)) and, again, the minimization leads to the action obtained by δ_R. In that, the decision rule

δ_R is a Bayes rule w.r.t. this loss function independent of the exact values that are used instead of c_2. The exact specification of these values does not contribute to guiding the decision about the relation $\mathsf{r} \in \mathcal{A}_R$, only the fact that they are larger than c_1. Therefore, if important information is incorporated within these values this information will not be used for guiding the decision, so a single value c_2 might be employed out of convenience.

4.2.3 Overall

Adding both loss functions lead to

$$L_{HDI+ROPE} : \Theta \times (\mathcal{A}_I \times \mathcal{A}_R) \to \mathbb{R}_0^+ \tag{26}$$

with

$$L_{HDI+ROPE}(\theta, ([a,b], \mathsf{r})) = L_I(\theta, [a,b]) + L_R^{[a,b]}(\theta, \mathsf{r}) \tag{27}$$

for which the overall decision rule $\delta_{HDI+ROPE}$ is a Bayes rule.

This can be seen by considering the expected posterior loss

$$\begin{aligned}
\rho_{HDI+ROPE}([a,b], \mathsf{r}) &= \int_\Theta L_{HDI+ROPE}(\theta, ([a,b], \mathsf{r}))\pi(\theta|\boldsymbol{x})d\theta \\
&= \int_\Theta \left[L_I(\theta, [a,b]) + L_R^{[a,b]}(\mathsf{r}) \right] \pi(\theta|\boldsymbol{x})d\theta \\
&= \int_\Theta L_I(\theta, [a,b])\pi(\theta|\boldsymbol{x})d\theta + \int_\Theta L_R^{[a,b]}(\mathsf{r})\pi(\theta|\boldsymbol{x})d\theta \\
&= \rho_I([a,b]) + L_R^{[a,b]}(\mathsf{r})\,.
\end{aligned} \tag{28}$$

The corresponding Bayes action

$$\underset{([a,b],\mathsf{r}) \in \mathcal{A}_I \times \mathcal{A}_R}{\arg\min} \; \rho_I([a,b]) + L_R^{[a,b]}(\mathsf{r})\,, \tag{29}$$

is obtained by minimizing this expected posterior loss.

The first part $\rho_I([a,b])$ does not depend on $\mathsf{r} \in \mathcal{A}_R$ and as outlined in in the previous Section 4.2.2, for all possible parameter intervals $[a,b]$, the second part $L_R^{[a,b]}(\mathsf{r})$ can be minimized by choosing its correct relation $\mathsf{r} \in \mathcal{A}_R$ with the predefined ROPE, which is obtained by the decision rule δ_R. Therefore, for any parameter interval $[a,b]$, the optimal relation is $\mathsf{r} = \delta_R([a,b])$.

The optimal parameter interval $[a,b] \in \mathcal{A}_I$ can now be obtained as

$$\begin{aligned}
&\underset{[a,b] \in \mathcal{A}_I}{\arg\min} \; \rho_I([a,b]) + L_R^{[a,b]}(\delta_R([a,b])) \\
\overset{\substack{eq. \\ (25)}}{=} &\underset{[a,b] \in \mathcal{A}_I}{\arg\min} \; \rho_I([a,b]) + c_1 \\
= &\underset{[a,b] \in \mathcal{A}_I}{\arg\min} \; \rho_I([a,b])\,,
\end{aligned} \tag{30}$$

which is the 95%-HDI – for an appropriate choice of c (see Section 4.2.1).

Taken together, as stated at the beginning of this subsection, the Bayes action $([a, b], \mathsf{r})$ w.r.t. $L_{HDI+ROPE}$ is obtained by $\delta_{HDI+ROPE}$.

The first restriction (I) mentioned in the previous Section 4.2.2 (equation (25)) is employed for finding the optimal parameter interval in the overall case, i.e. within equation (30). Without this restriction (I), $L_R^{[a,b]}(\delta_R([a, b]))$ would not be a constant c_1, but a value that depends on the interval $[a, b]$. In that, the minimization in equation (30) could yield an interval, which is not the HDI, a fact that is referred to as "paradoxical behavior" by Kruschke [2018, supp. p. 4] (who also refers to Casella et al. [1993] in this context).

4.3 Final Decision

In contrast to the foundation of the ROPE-only decision rule, the loss function $L_{HDI+ROPE}$ does not allow a reasonable employment of the framework of imprecise probabilities. This is because L_I need to be as it is in order to obtain the HDI and L_R uses only the fact that c_1 is smaller than c_2. As depicted in Section 4.2.2, any additional information within these constants is not being used. Therefore, no potentially vague information can be captured within $L_{HDI+ROPE}$. As a consequence, the framework of imprecise probabilities cannot be employed within this context to elegantly formalize withholding a decision between a_0 and a_1. Therefore, the action space of the final decision comprises all a_0, a_1 and the action to withhold the decision.

Of course, in the context of the HDI+ROPE decision rule, there is a bijective mapping between $\mathcal{A}_R$ and the action space for this final decision:

$$\mathsf{r}_0 \mapsto \mathsf{a}_0 \qquad \mathsf{r}_1 \mapsto \mathsf{a}_1 \qquad \mathsf{r}_2 \mapsto \text{withhold decision}\,. \tag{31}$$

Accordingly, this last step does not need a separate decision theoretic account, as the final actions might be employed instead of the three relations $\mathsf{r} \in \mathcal{A}_R$.

Nevertheless, from a content point of view, this final step should be treated separately from the determination of the relation between the HDI and the ROPE. As outlined within this Section 4, the HDI+ROPE decision rule is primarily focusing on technical aspects of how to obtain the HDI and determine its relation with a predefined ROPE, and it is this final step that tries to build the connection to the underlying real-world decision of interest.

5 Discussion

The decision theoretic foundations of both Kruschke's HDI+ROPE decision rule [Kruschke, 2015, 2018] and the ROPE-only decision rule [Kruschke, 2018, supp. p. 5] are outlined within this technical report. Both decision rules are depicted as Bayes rules w.r.t. certain loss functions. In that, different loss functions are considered: First, although inaccessible, there is an underlying "true" loss function characterizing the real-world decision of interest.

Second, in the context of considerations about the ROPE (see Section 3.1), this "true" loss function is simplified, such that it might be specified by only a single number. Still, this simplified loss function characterizes the real-world decision of interest. Third, there is a loss function w.r.t. to finding the HDI for a given data set and, fourth, a rather artificial loss function might be employed in the context of determining the relation between a parameter interval and a pre-defined ROPE. Fifth, the loss function in the context of the HDI+ROPE decision rule is a combination of the previous two.

Naturally, by considering these different loss functions, different decision rules are characterized as Bayes rules. Put aptly by Rice et al. [2008, p. 3], "for the precise 'question' asked by loss function L and the stated modeling assumptions, one can think of the Bayes rules as providing the 'best' answer". In that, these five loss functions are asking:

- How should I decide in the real-world decision problem?
- Given the simplification, how should I decide in the real-world decision problem?
- Which interval is the HDI of the posterior distribution?
- How is the relation of the HDI and the ROPE?
- Which interval is the HDI of the posterior distribution and how is the relation of it with the ROPE?

The first question is of interest but cannot be answered, because the loss function is inaccessible. The second question does relate to the real-world decision of interest and might be used as a proxy for the first question (given the employed simplification is reasonable), as the corresponding simplified loss function still contains information w.r.t. to the real-world decision of interest. By allowing this loss function to be specified imprecisely, relevant information might be incorporated into the analysis as it is available. In this context, the ROPE-only decision rule is optimal when resorting to an arbitrary choice of interval-valued loss functions.

The third question does not address the real-world decision of interest at all. Although the fourth question contains the ROPE, the corresponding loss function considers only the bounds of the ROPE and not respective loss values that are in accordance with the real-world decision of interest (as within the second (simplified) loss function). In that, the fourth question relates to the real-world decision of interest only marginally and primarily addresses a rather technical interval comparison. Therefore, as a combination of the previous two, the fifth question does not primarily ask about the real-world decision problem, yet is implicitly used as a proxy for it when employing the HDI+ROPE decision rule.

In summary, the ROPE-only decision rule might be characterized as being closer to the real-world decision of interest than the HDI+ROPE decision rule. This might also be seen by the fact, that both the "true" underlying loss function and the posterior distribution are essential to derive the optimal decision in a Bayesian framework, yet the HDI+ROPE decision rule uses less of these information than the ROPE-only decision rule: First, former

simplifies the posterior distribution by sorting back to the less-informative HDI. Second, former employs only the bounds of the ROPE and latter also information about the loss-magnitude.

Of course, the arbitrary choice about the loss value interval within the simplified loss function in the context of the ROPE-only decision rule (see Section 3.3) has to be criticized. The corresponding interval should be chosen based on the real-world decision of interest. As it is to expect that at least some information about this loss value in the simplified loss function is available[9], the framework of imprecise probabilities offers an elegant way to include this essential but vague information.

References

T. Augustin. On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view — a cautionary note on updating imprecise priors. In J.-M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA '03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, pages 31–45, Lugano, Waterloo, 2003. Carleton Scientific.

T. Augustin, F. P. A. Coolen, G. De Cooman, and e. Troffaes, Matthias C. M. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

J. O. Berger. *Statistical decision theory and Bayesian analysis. 2nd edition*. Springer, 1995.

G. Casella, J. T. G. Hwang, and C. Robert. A paradox in decision-theoretic interval estimation. *Statistica Sinica*, 3(1):141–155, 1993.

G. Cumming. The new statistics: why and how. *Psychological Science*, 25(1):7–29, 2014.

D. Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75:643–669, 1961.

J. Etner, M. Jeleva, and J.-M. Tallon. Decision theory under ambiguity. *Journal of Economic Surveys*, 26(2):234–270, 2012.

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

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

I. Levi. *The enterprise of knowledge: an essay on knowledge, credal probability, and chance*. MIT Press, Cambridge, 1980.

[9]Guiding a decision without any information at all about the consequences is truly arbitrary. If so, it is indispensable to put effort into obtaining this information first.

K. M. Rice, T. Lumley, and A. A. Szpiro. Trading bias for precision: decision theory for intervals and sets. Technical report, 2008. Retrieved from `https://biostats.bepress.com/uwbiostat/paper336/`.

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

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007.

M. J. Schervish. *Theory of statistics*. Springer, 1995.

P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman & Hall, 1991.

Eidesstattliche Versicherung

Hiermit erkläre ich, Patrick M. Schwaferts, an Eides statt, dass die Dissertation von mir selbstständig und ohne unerlaubte Hilfe oder Hilfsmittel angefertigt worden ist.

Hereby I, Patrick M. Schwaferts, declare in lieu of an oath that the present dissertation was composed autonomously without any illicit aid.

Tutzing, 21. Oktober 2021

(Patrick M. Schwaferts)