

UvA-DARE (Digital Academic Repository)

Bayes factors for research workers

Ly, A.

[Link to publication](#)

License

Other

Citation for published version (APA):

Ly, A. (2018). *Bayes factors for research workers*.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

BAYES FACTORS FOR RESEARCH WORKERS

In this dissertation we advocate the use of Bayes factors in empirical research to replace or complement standard null hypothesis tests based on p -values. These Bayes factors were specifically designed to quantify the evidence for or against the existence of an effect. This was done by comparing two models with the same distributional assumptions, where the alternative model is an extension of the null model by incorporating one extra parameter. Instead of returning a decision to "reject" or "not reject", a Bayes factor $\text{BF}_{10}(d)$ returns a non-negative number that represents the evidence provided by the observed data d for the model that includes the effect. The returned number can be seen as a refinement of the binary decision with $\text{BF}_{10}(d) = \infty$ and $\text{BF}_{10}(d) = 0$ corresponding to definite rejection and acceptance of the null, respectively. Moreover, the Bayes factor allows its users to forgo the binary decision and acknowledge uncertainty, so that the evidence can be updated continually in light of new data, directly and easily. For empirical scientists to be able to use these Bayes factors, we implemented them in *Jeffreys's Amazing Statistics Program*, JASP, which is freely available and open-source.

(url: <https://jasp-stats.org>)

ALEXANDER-LY.COM

BAYES FACTORS FOR RESEARCH WORKERS

ALEXANDER LY

ALEXANDER LY BAYES FACTORS FOR RESEARCH WORKERS



Bayes Factors for Research Workers

Alexander Ly

ISBN 978-94-028-0897-1

This publication is typeset in L^AT_EX using the Memoir class

Printed by Ipkamp Printing B.V., Enschede

Cover by Viktor Beekman, viktorbeekman.nl

Copyright © 2017 by Alexander Ly

All rights reserved

The research of this doctoral thesis received financial assistance from the European Research Council (ERC; 283876)

Bayes Factors for Research Workers

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit

op 19 januari 2018, te 13.00 uur

door Alexander Ly

geboren te Heerlen

Promotiecommissie

Promotor:	Prof. dr. E. M. Wagenmakers	Universiteit van Amsterdam
Copromotor:	Dr. M. Marsman	Universiteit van Amsterdam
Overige leden:	Prof. dr. J. O. Berger	Duke University
	Prof. dr. M. D. Lee	UC Irvine
	Prof. dr. P. D. Grünwald	Universiteit Leiden
	Prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
	Dr. L. J. Waldorp	Universiteit van Amsterdam
	Dr. R. P. P. P. Grasman	Universiteit van Amsterdam
Faculteit:	Faculteit der Maatschappij- en Gedragswetenschappen	

*Voor opa Minh Chung Ly
mijn vader Han Giang Ly
mijn moeder Phuong Hoa Tran
mijn zus Julia-Selina Ly
mijn liefste Gracia Edwards
my family*

Contents

1	Introduction	1
1.1	Bayesian model learning	1
1.2	Chapter outline	2
I	Bayes Factor Rationale	7
2	Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology	9
2.1	Introduction	10
2.2	Historical and philosophical background of the Bayes factor	11
2.3	Jeffreys's procedure for constructing a default Bayes factor	16
2.4	Jeffreys's Bayes factor for the test of the nullity of a normal mean: The Bayesian t -test	19
2.5	Jeffreys's Bayes factor for the test of the nullity of a correlation	25
2.6	Conclusion	36
2.A	The default Bayes factor hypothesis tests proposed by Jeffreys in ToP	38
2.B	The hypergeometric function	38
2.C	The stretched beta density	38
2.D	Translation of Jeffreys's notation in ToP	39
3	An Evaluation of Alternative Methods for Testing Hypotheses, from the Perspective of Harold Jeffreys	41
3.1	Introduction	42
3.2	Rejoinder to Robert	42
3.3	Rejoinder to Chandramouli and Shiffrin	51
3.4	Conclusion	65
II	Bayes Factors for Common Designs	67
4	Bayesian Inference for Kendall's Rank Correlation Coefficient	69
4.1	Introduction	69
4.2	Methods	72
4.3	Results	75
4.4	Concluding comments	77

5 Informed Bayesian <i>t</i>-Tests	81
5.1 Introduction	81
5.2 Jeffreys's default Bayes factor	83
5.3 One-sample and paired samples <i>t</i> -test	84
5.4 Two-sample <i>t</i> -test	89
5.5 Example III: Reanalysis of 593 <i>t</i> -tests	95
5.6 Quantifying evidence for H_0	97
5.7 Concluding comments	99
6 A Limit-Consistent Bayes Factor for Testing the Equality of Two Poisson Rates	101
6.1 Introduction	101
6.2 Jeffreys's Bayes factor for the comparison of two Poisson rates	104
6.3 A limit-consistent Bayes factor for the comparison of two Poisson rates	107
6.4 Discussion	111
III Scientific Learning with Bayes Factors	113
7 Four Requirements for an Acceptable Research Programme	115
7.1 The power fallacy	116
7.2 The fallacy of the transposed conditional	116
7.3 Requirements of a research programme	117
7.4 Concluding comments	120
8 Bayesian Reanalyses from Summary Statistics: A Guide for Academic Consumers	123
8.1 Introduction	124
8.2 The Festinger & Carlsmith (1959) cognitive dissonance study	125
8.3 Bayesian reanalysis	126
8.4 Concluding comments	128
9 Replication Bayes Factors from Evidence Updating	129
9.1 Introduction	129
9.2 The Bayes factor	131
9.3 Bayesian updating in action	132
9.4 The replication Bayes factor reconceptualised	134
9.5 Example 1: A <i>t</i> -test to assess whether superstition improves performance	137
9.6 Example 2: A contingency table analysis to test whether more valuable stimuli are judged to be relatively rare	138
9.7 Concluding comments	140
9.A Deriving the <i>t</i> -value across all data sets	141
9.B Replication Bayes factors as conditional Bayes factors	143
9.C Replication paradox and solution	143

IV Analytic Results	147
10 Analytic Posteriors for Pearson’s Correlation Coefficient	149
10.1 Introduction	149
10.2 Notation and result	150
10.3 Analytic posteriors for the case $\beta = 0$	155
10.A A lemma distilled from the Bateman Project	157
11 Analytic Posteriors for the Binomial Rate Parameters, and the Odds Ratio	159
11.1 Introduction	159
11.2 Binomial distribution	159
11.3 Products of generalised beta prime distributions and the odds ratio	166
11.4 Concluding remarks	168
V Two Tutorials	169
12 A Tutorial on Bridge Sampling	171
12.1 Introduction	171
12.2 Four sampling methods to approximate the marginal likelihood . .	174
12.3 Case study: Bridge sampling for reinforcement learning models .	193
12.4 Discussion	205
12.A The bridge sampling estimator as a general case of methods 1 – 3 .	207
12.B Bridge sampling implementation: Avoiding numerical issues . .	207
12.C Correcting for the probit transformation	208
12.D Details on the application of bridge sampling to the individual-level EV model	210
12.E Details on the application of bridge sampling to the hierarchical EV model	211
13 A Tutorial on Fisher Information	213
13.1 Introduction	213
13.2 The role of Fisher information in frequentist statistics	217
13.3 The role of Fisher information in Bayesian statistics	222
13.4 The role of Fisher information in minimum description length . .	231
13.5 Concluding comments	242
13.A Generalisation to vector-valued parameters: The Fisher information matrix	244
13.B Frequentist statistics based on asymptotic normality	245
13.C Bayesian use of the Fisher-Rao metric: The Jeffreys’s prior . . .	249
13.D MDL: Coding theoretical background	256
13.E Regularity conditions	260
VI Conclusion	265
14 Discussion and Future Directions	267

CONTENTS

References	271
Nederlandse Samenvatting	299
Acknowledgements — Dankwoord	305
Publications	307

Chapter 1

Introduction

Abstract

The goal of this project was to develop and promote Bayesian hypothesis tests for social scientists. By and large, social scientists have ignored the Bayesian revolution in statistics, and, consequently, most social scientists still assess the veracity of experimental effects using the same methodology that was used by their advisors and the advisors before them. This state of affairs is undesirable: social scientists conduct groundbreaking, innovative research only to analyse their results using methods that are old-fashioned or even inappropriate. This imbalance between the science and the statistics has gradually increased the pressure on the field to change the way inferences are drawn from their data. However, three requirements need to be fulfilled before social scientists are ready to adopt Bayesian tests of hypotheses. First, the Bayesian tests need to be developed for problems that social scientists work with on a regular basis; second, the Bayesian tests need to be default or objective; and, third, the Bayesian tests need to be available in a user-friendly computer program.

1.1 Bayesian model learning

The Bayesian hypothesis tests developed here are designed to help empirical scientist (i) quantify the evidence in favour or against a hypothesis from the observed data, and, more importantly, (ii) extract information from the observed data to learn, construct and grow models and theories.

A *statistical model* is a simplification of reality and defines a functional relationship $f(d|\theta)$ between data d and so-called *parameters*. For instance, d can represent blood pressure measurements before and after treatment of a sample of patients, while θ represents the effect size of the treatment in the population of patients, and f is typically a normal distribution that accounts for the noise, due to only measuring a small sample of a larger population.

To test whether the treatment has an effect on the population of patients we compare the *null model* M_0 , the statistical model with the effect size restricted

at zero $\theta = 0$, against the *alternative model* \mathcal{M}_1 where the effect size θ is free to vary.

The *prior plausibility* of there being an effect before any datum is observed depends on the treatment. For instance, the prior probability of there being an effect is relatively high, say, $P(\mathcal{M}_1) = 0.9$ and $P(\mathcal{M}_0) = 0.1$, when the treatment involves the intake of a pill that includes an active component designed to lower blood pressure. Equivalently, we then say that the *prior model odds* of there being an effect is nine to one, that is, $\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} = 9$. The prior model odds can be updated in light of the observed data d using *Bayes' rule* which leads to the crucial equation

$$\underbrace{\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_0 | d)}}_{\text{Posterior model odds}} = \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{prior model odds}} \quad (1.1.1)$$

where $P(\mathcal{M}_i | d)$ is the *posterior model probability* of model \mathcal{M}_i updated by the data and $p(d | \mathcal{M}_i)$ is the marginal likelihood of \mathcal{M}_i . The term $\text{BF}_{10}(d)$ is known as the *Bayes factor* and equals the change from prior to posterior model odds brought about by the observed data d . The Bayes factor has an intuitive interpretation: $\text{BF}_{10}(d) = 7$ indicates that the observed data are 7 times more likely under \mathcal{M}_1 than under \mathcal{M}_0 , whereas $\text{BF}_{10}(d) = .2$ indicates that the observed data are 5 times more likely under \mathcal{M}_0 than under \mathcal{M}_1 . In general, the Bayes factor returns a non-negative number given the observed data d , and the higher (lower) the value of $\text{BF}_{10}(d)$, the more (less) evidence for \mathcal{M}_1 over \mathcal{M}_0 . In a similar fashion, if the patients' activity levels were also measured one can investigate whether the treatment makes people tired. Slowly and gradually one can then chart how the treatment influences the population of patients.

The Bayes factor is given by a ratio of marginal likelihood $p(d | \mathcal{M}_i)$ that represents how well model \mathcal{M}_i fits the observed data. This marginal likelihood can be thought of as the functional relationship $f_i(d | \theta)$ of model \mathcal{M}_i at the observed data d and weighted with respect to a so-called *prior distribution* $\pi_i(\theta)$ at each possible parameter value θ :

$$p(d | \mathcal{M}_i) = \int f_i(d | \theta) \pi_i(\theta) d\theta. \quad (1.1.2)$$

Hence, given two models, that is, the functional relationships $f_1(d | \theta)$ and $f_0(d | \theta)$, the statistician is required to choose two priors, namely, $\pi_0(\theta)$ and $\pi_1(\theta)$ to construct a Bayes factor. For the Bayes factor to be accessible to practitioners, they have to be computable for any data set d . This dissertation discuss both issues: The choice of priors for a Bayes factor, and its computations.

1.2 Chapter outline

1.2.1 Part I. Bayes factor rationale

The first part of the dissertation focusses on the philosophy, motivation and the construction of Bayes factors based on the work of Harold Jeffreys.

Chapter 2 elaborates on the principles upon which the Bayes factor is founded, how it is interpreted, and presents a general scheme with which Jeffreys selected prior distributions and constructed Bayes factors. The idea is to propose a Bayes factor that is *predictively matched* and *information consistent*. A predictively matched Bayes factor returns one for inconclusive data, whilst an information consistent Bayes factor returns infinite support for the alternative when the data are overwhelmingly in favour of there being an effect. This scheme is extracted from how Jeffreys treated the test of nullity of a normal mean, the Bayesian t -test and, subsequently, applied to construct a novel Jeffreys's Bayes factor for Pearson's correlation. This Bayes factor is analytic, thus, easily computed.

Chapter 3 gives additional insights on Bayes factors as a response to two comments from renowned researchers. In this rejoinder we took the opportunity to further elaborate on the Jeffreys-Lindley-Bartlett paradox, the distinction between inference and decision making as well on the difference between a testing and an estimation problem.

1.2.2 Part II. Bayes factors for common designs

The second part of the dissertation focusses on the Bayes factors that were developed for other scenarios that empirical scientists commonly encounter.

Chapter 4 outlines a Bayesian methodology to estimate and test the Kendall rank correlation coefficient τ . The key idea is to model the test statistic rather than the data, and exploit the analytic result derived for the Bayes factor for Pearson's correlation.

Chapter 5 also exploits the result derived for Pearson's correlation, but this time to define, if one wishes, an informed Bayes factor to test the nullity of a normal mean. An extension of Jeffreys's default t -test is presented that allows researchers to incorporate expert knowledge into the prior specification of the effect size parameter δ . Specifically, two families of prior distributions for δ are considered: the family of shifted and scaled t distributions (which includes Jeffreys's Cauchy prior as a special case) and the family of shifted and scaled normal distributions. For both families we derive the marginal posterior distribution of δ and the Bayes factor. For the normal family the solutions are completely analytic; for the t family the solutions contain a one-dimensional integral that can easily be evaluated numerically. The impact of incorporation of prior knowledge is illustrated with three examples.

Chapter 6 introduces the desideratum of *limit-consistency* as a means to facilitate the selection of prior distribution with good properties. This desideratum is relevant for tests of equality between two processes, and it concerns the hypothetical scenario where data acquisition for one process is terminated early whereas data acquisition of the second process continues indefinitely. In such cases, the Bayes factor ought to approach a finite limit. The Bayes factor Jeffreys proposed for the two-sample Poisson problem, unfortunately, violates limit-consistency and we propose a generalisation of Jeffreys's test that is limit-consistent.

1.2.3 Part III. Scientific learning with Bayes factors

The third part of the dissertation focusses on the use of Bayes factors in the empirical sciences as a tool for scientific learning. It also touches upon the “crisis of confidence” (e.g., Baker, 2016, Levent et al., 2012, Pashler and Wagenmakers, 2012).

Chapter 7 highlights how psychologists have been at the forefront of efforts to assess and improve reproducibility in science by way of large-scale replication initiatives, such as the Reproducibility Project: Psychology (Open Science Collaboration, 2015), the *Social Psychology* special issue on replication (Nosek and Lakens, 2014), and the various ManyLabs efforts (Ebersole et al., 2016; Klein et al., 2014). This chapter is a comment on Witte and Zenker (2016) who believe that a “different” use of p -values can resolve the crisis of confidence. We disagree, as statistics alone cannot avoid another crisis. Instead, we argue that confirmatory research should be preregistered. By preregistering an experiment one avoids hindsight bias and controls the problem of multiple testing. Moreover, we also believe that science should be open and transparent, and that researchers should acknowledge uncertainty, as this gives a more honest and better reflection of the scientific process.

Chapter 8 shows how easy it is to do a Bayesian reanalysis even without access to the full data set. This is interesting for researchers who want to complement their p -values with a Bayes factor. A Bayesian reanalysis is also useful for editors, reviewers, readers, and reporters, as it allows for the quantification of the evidence on a continuous scale. In addition, we also provide tools that allow for an assessment of the robustness of the evidence within the data to changes to the prior distribution. Furthermore, by expanding a summary statistic into a posterior one can gauge which posterior parameter ranges are more credible than others. Moreover, this posterior can be used as an informed prior for a subsequent study.

Chapter 9 describes a general method that allows experimenters to quantify the evidence from the data of a direct replication attempt given data already acquired from an original study. This general method was designed to help researchers build a body of knowledge based on the data from the increased number of replication studies in response to psychology’s crisis of confidence.

1.2.4 Part IV. Analytic results

The fourth part presents various analytic results that have been used in the construction of the Bayesian tests presented in this dissertation.

Chapter 10 provides the analytic posterior for Pearson’s correlation coefficient for a large class of priors, and Bernardo’s reference prior in particular. This result is used to construct the analytic Bayes factor given in Chapter 2 and forms the basis of Chapters 4 and 5.

Chapter 11 provides various analytic posteriors for two scenarios involving discrete data. One of these results is used in Chapter 6 and can also be used to define a robustness analysis in a binomial test. In addition, analytic expression are given from which one can construct a one-sided binomial Bayes factor. The last

result is an analytic expression for the odds ratio in a 2-by-2 contingency table, which is a topic for future research.

1.2.5 Part V. Two tutorials

The fifth part of the dissertation focusses on tools to construct Bayes factors and statistical modelling in general.

Chapter 12 elaborates on how bridge sampling (Meng and Wong, 1996) can be used to transform MCMC output into an estimate of the marginal likelihood. The bridge sampler is particularly useful for complicated models with hierarchical structures and when the marginal likelihood is intractable.

Chapter 13 gives general background on mathematical statistics and the role of Fisher information in particular. In this tutorial we clarify the concept of Fisher information as it manifests itself across three different statistical paradigms. Firstly, in the frequentist paradigm, Fisher information is used to construct hypothesis tests and confidence intervals using maximum likelihood estimators; secondly, in the Bayesian paradigm, Fisher information is used to define a default prior; finally, in the minimum description length paradigm, Fisher information is used to measure model complexity.

The dissertation is concluded with a discussion on future directions.

Part I

Bayes Factor Rationale

Chapter 2

Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology

Abstract

Harold Jeffreys pioneered the development of default Bayes factor hypothesis tests for standard statistical problems. Using Jeffreys's Bayes factor hypothesis tests, researchers can grade the decisiveness of the evidence that the data provide for a point null hypothesis \mathcal{H}_0 versus a composite alternative hypothesis \mathcal{H}_1 . Consequently, Jeffreys's tests are of considerable theoretical and practical relevance for empirical researchers in general and for experimental psychologists in particular. To highlight this relevance and to facilitate the interpretation and use of Jeffreys's Bayes factor tests we focus on two common inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian equivalent of the t -test) and testing the nullity of a correlation. For both Bayes factor tests, we explain their development, we extend them to one-sided problems, and we apply them to concrete examples from experimental psychology.

Keywords: Correlation test, hypothesis testing, model selection, statistical evidence, t -test.

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

2.1 Introduction

Consider the common scenario where a researcher entertains two competing hypotheses. One, the null hypothesis \mathcal{H}_0 , is implemented as a statistical model that stipulates the nullity of a parameter of interest (i.e., $\mu = 0$); the other, the alternative hypothesis \mathcal{H}_1 , is implemented as a statistical model that allows the parameter of interest to differ from zero. How should one quantify the relative support that the observed data provide for \mathcal{H}_0 versus \mathcal{H}_1 ? Harold Jeffreys argued that this is done by assigning prior mass to the point null hypothesis (or “general law”) \mathcal{H}_0 , and then calculate the degree to which the data shift one’s prior beliefs about the relative plausibility of \mathcal{H}_0 versus \mathcal{H}_1 . The factor by which the data shift one’s prior beliefs about the relative plausibility of two competing models is now widely known as the Bayes factor, and it is arguably the gold standard for Bayesian model comparison and hypothesis testing (e.g., Berger, 2006; Lee and Wagenmakers, 2013; Lewis and Raftery, 1997; Myung and Pitt, 1997; O’Hagan and Forster, 2004).

In his brilliant monograph “Theory of Probability”, Jeffreys introduced a series of default Bayes factor tests for common statistical scenarios. Despite their considerable theoretical and practical appeal, however, these tests are hardly ever used in experimental psychology and other empirical disciplines. A notable exception concerns Jeffreys’s equivalent of the t -test, which has recently been promoted by Jeffrey Rouder, Richard Morey, and colleagues (e.g., Rouder et al., 2009). One of the reasons for the relative obscurity of Jeffreys’s default tests may be that a thorough understanding of “Theory of Probability” requires not only an affinity with mathematics but also a willingness to decipher Jeffreys’s non-standard notation.

In an attempt to make Jeffreys’s default Bayes factor tests accessible to a wider audience we explain the basic principles that drove their development and then focus on two popular inferential scenarios: Testing the nullity of a normal mean (i.e., the Bayesian t -test), and testing the nullity of a correlation. We illustrate Jeffreys’s methodology using data sets from psychological studies. The chapter is organised as follows: The first section provides some historical background and outlines four of Jeffreys’s convictions regarding scientific learning. The second section shows how the Bayes factor is a natural consequence of these four convictions. We decided to include Jeffreys’s own words where appropriate, so as to give the reader an accurate impression of Jeffreys’s ideas as well as his compelling style of writing. The third section presents the procedure from which so-called default Bayes factors can be constructed. This procedure is illustrated with the redevelopment of the Bayesian counterpart for the t -test and the Bayesian correlation test. For both the t -test and the correlation test, we also derive one-sided versions of Jeffreys’s original tests. We apply the resulting Bayes factors to data sets from psychological studies. The last section concludes with a summary and a discussion.

2.2 Historical and philosophical background of the Bayes factor

2.2.1 Life and work

Sir Harold Jeffreys was born in 1891 in County Durham, United Kingdom, and died in 1989 in Cambridge. Jeffreys first earned broad academic recognition in geophysics when he discovered the earth's internal structure (Bolt, 1982; Jeffreys, 1924). In 1946, Jeffreys was awarded the Plumian Chair of Astronomy, a position he held until 1958. After his "retirement" Jeffreys continued his research to complete a record-breaking 75 years of continuous academic service at any Oxbridge college, during which he was awarded medals by the geological, astronomical, meteorological, and statistical communities (Cook, 1990; Huzurbazar, 1991; Lindley, 1991; Swirles, 1991). His mathematical ability is on display in the book "Methods of Mathematical Physics", which he wrote together with his wife (Jeffreys and Jeffreys, 1946).

Our first focus is on the general philosophical framework for induction and statistical inference put forward by Jeffreys in his monographs "Scientific Inference" (Jeffreys, 1931, second edition 1955, third edition 1973) and "Theory of Probability" (henceforth ToP; first edition 1939, second edition 1948, third edition 1961). An extended modern summary of ToP is provided by (Robert et al., 2009). Jeffreys's ToP rests on a principled philosophy of scientific learning (ToP, Chapter I). In ToP, Jeffreys distinguishes sharply between problems of parameter estimation and problems of hypothesis testing. For estimation problems, Jeffreys outlines his famous parameterisation-invariant "Jeffreys's priors" (ToP, Chapter III); for testing problems, Jeffreys proposes a series of default Bayes factor tests to grade the support that observed data provide for a point null hypothesis \mathcal{H}_0 versus a composite \mathcal{H}_1 (ToP, Chapter V). A detailed summary of Jeffreys's contributions to statistics is available online at www.economics.soton.ac.uk/staff/aldrich/jeffreysweb.htm.

For several decades, Jeffreys was one of only few scientists who actively developed, used, and promoted Bayesian methods. In recognition of Jeffreys's persistence in the face of relative isolation, E. T. Jaynes's dedication of his own book, "Probability theory: The logic of science", reads: "Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it" (Jaynes, 2003). In 1980, the seminal work of Jeffreys was celebrated in the 29-chapter book "Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys" (e.g, Geisser, 1980; Good, 1980; Lindley, 1980; Zellner, 1980). In one of its chapters, Dennis Lindley discusses ToP and argues that "The *Theory* is a wonderfully rich book. Open it at almost any page, read carefully, and you will discover some pearl." (Lindley, 1980, p. 37).

Despite discovering the internal structure of the earth and proposing a famous rule for developing parameterisation-invariant prior distributions, Jeffreys himself considered his greatest scientific achievement to be the development of the Bayesian hypothesis test by means of default Bayes factors (Senn, 2009). In what follows, we explain the rationale behind Jeffreys's Bayes factors and demonstrate their use for two concrete tests.

2.2.2 Jeffreys's view of scientific learning

Jeffreys developed his Bayes factor hypothesis tests as a natural consequence of his perspective on statistical inference, a philosophy guided by principles and convictions inspired by Karl Pearson's classic book *The Grammar of Science* and by the work of W. E. Johnson and Dorothy Wrinch. Without any claim to completeness or objectivity, here we outline four of Jeffreys's principles and convictions that we find particularly informative and relevant.

2.2.2.1 Conviction i: Inference is inductive

Jeffreys's first conviction was that scientific progress depends primarily on induction (i.e., learning from experience). For instance, he states "There is a solid mass of belief reached inductively, ranging from common experience and the meanings of words, to some of the most advanced laws of physics, on which there is general agreement among people that have studied the data." (Jeffreys, 1955, p. 276) and, similarly: "When I taste the contents of a jar labelled 'raspberry jam' I expect a definite sensation, inferred from previous instances. When a musical composer scores a bar he expects a definite set of sounds to follow when an orchestra plays it. Such inferences are not deductive, nor indeed are they made with certainty at all, though they are still widely supposed to be." (Jeffreys, 1973, p. 1). The same sentiment is stated more forcefully in ToP: "... the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements" (Jeffreys, 1961, p. 5). Hence, inference is inductive and should be guided by the data we observe.

2.2.2.2 Conviction ii: Induction requires a logic of partial belief

Jeffreys's second conviction is that in order to formalise induction one requires a logic of partial belief: "The idea of a reasonable degree of belief intermediate between proof and disproof is fundamental. It is an extension of ordinary logic, which deals only with the extreme cases." (Jeffreys, 1955, p. 275). This logic of partial belief, Jeffreys showed, needs to obey the rules of probability calculus in order to fulfil general desiderata of consistent reasoning –thus, degrees of belief can be thought of as probabilities (cf. Ramsey, 1926). Hence, all the unknowns should be instantiated as random variables by specifying so-called prior distributions before any datum is collected. Using Bayes' theorem, these priors can then be updated to posteriors conditioned on the data that were actually observed.

2.2.2.3 Conviction iii: The test of a general law requires it be given prior probability

Jeffreys's third conviction stems from his rejection of treating a testing issue as one of estimation. This is explained clearly and concisely by Jeffreys himself:

"My chief interest is in significance tests. This goes back to a remark in Pearson's *Grammar of Science* and to a paper of 1918 by C. D.

Broad. Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property ϕ , and we do not know r , the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find them all with ϕ , the posterior probability that all n are ϕ s is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been inspected. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919–1923. Our point was that giving prior probability $1/(n+1)$ to a general law is that for n large we are already expressing strong confidence that no general law is true. The way out is obvious. To make it possible to get a high probability for a general law from a finite sample the prior probability must have at least some positive value independent of n ." (Jeffreys, 1980, p. 452)

The allocation of probability to the null hypothesis is known as the simplicity postulate (Wrinch and Jeffreys, 1921), that is, the notion that scientific hypotheses can be assigned prior plausibility in accordance with their complexity, such that "the simpler laws have the greater prior probabilities" (e.g., Jeffreys, 1961, p. 47; see also Jeffreys, 1973, p. 38). In the case of testing a point null hypothesis, the simplicity postulate expresses itself through the recognition that the point null hypothesis represents a general law and, hence, requires a separate, non-zero prior probability.

Jeffreys's view of the null hypothesis as a general law is influenced by his background in (geo)physics. For instance, Newton's law of gravity postulates the existence of a fixed universal gravitational constant G . Clearly, this law is more than just a statement about a constant; it provides a model of motion that relates data to parameters. In this context, the null hypothesis should be identified with its own separate null model M_0 rather than be perceived as a simplified version of an encompassing model M_1 .

Hence, Jeffreys's third conviction holds that in order to test the adequacy of a null hypothesis, the model that instantiates that hypothesis needs to be assigned a separate prior probability, which can be updated by the data to a posterior probability.

2.2.2.4 Conviction iv: Classical tests are inadequate

Jeffreys's fourth conviction was that classical "Fisherian" p -values are inadequate for the purpose of hypothesis testing. In the preface to the first edition of ToP, Jeffreys outlines the core problem: "Modern statisticians have developed extensive mathematical techniques, but for the most part have rejected the notion of the probability of a hypothesis, and thereby deprived themselves of any way of saying precisely what they mean when they decide between hypotheses" (Jeffreys, 1961, p. ix). Specifically, Jeffreys pointed out that the p -value significance test "(...) does not give the probability of the hypothesis; what it does give is a convenient,

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

though rough, criterion of whether closer investigation is needed.” (Jeffreys, 1973, p. 49). Thus, by selectively focusing on the adequacy of predictions under the null hypothesis—and by neglecting the adequacy of predictions under the alternative hypotheses—researchers may reach conclusions that are premature (see also the Gosset-Berkson critique, Berkson, 1938; Wagenmakers et al., 2017c):

“Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts.” (Jeffreys, 1961, p. 390).

Jeffreys also argued against the logical validity of *p*-values, famously pointing out that they depend on more extreme events that have not been observed: “What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.” (Jeffreys, 1961, p. 385). In a later paper, Jeffreys clarifies this statement: “I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.” (Jeffreys, 1980, p. 453).

In sum, Jeffreys was convinced that induction is an extended form of logic; that this “logic of partial beliefs” needs to treat degrees of belief as probabilities; that simple laws or hypotheses should be viewed as separate models that are allocated non-zero prior probabilities, and that a useful and logically consistent method of hypothesis testing needs to be comparative, and needs to be based on the data at hand rather than on data that were never observed. These convictions coalesced in Jeffreys’s development of the Bayes factor, an attempt to provide a consistent method of model selection and hypothesis testing that remedies the weaknesses and limitations inherent to *p*-value statistical hypothesis testing.

2.2.3 The Bayes factor hypothesis test

In reverse order, we elaborate on the way in which each of Jeffreys’s convictions motivated the construction of his Bayes factor alternative to the classical hypothesis test.

2.2.3.1 ad. Conviction iv: Classical tests are inadequate

Jeffreys’s development of a Bayesian hypothesis test was motivated in part by his conviction that the use of classical *p* values is “absurd”. Nevertheless, Jeffreys reported that the use of Bayes factors generally yields conclusions similar to those reached by means of *p* values: “As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher’s, and have not yet found a disagreement in the actual decisions reached” (Jeffreys, 1961, p. 393); thus, “In spite of the difference in principle between my tests and those

based on the P integrals (...) it appears that there is not much difference in the practical recommendations." (Jeffreys, 1961). However, Jeffreys was acutely aware of the fact that disagreements can occur (see also Edwards et al., 1963; Lindley, 1957). In psychology, these disagreements appear to arise repeatedly, especially for cases in which the p value is in the interval from .01 to .05 (Johnson, 2013; Wetzels et al., 2011).

2.2.3.2 ad. Conviction iii: The test of a general law requires it be given prior probability

Jeffreys first identified the null hypothesis with a separate null model \mathcal{M}_0 that represents a general law and pits it against the alternative model \mathcal{M}_1 which relaxes the restriction imposed by the law. For instance, for the t -test, \mathcal{M}_0 : normal data X with $\mu = 0$ –the law says that the population mean is zero– and \mathcal{M}_1 : normal data X that allows μ to vary freely. As we do not know whether the data were generated according to \mathcal{M}_0 or \mathcal{M}_1 we consider the model choice a random variable such that $P(\mathcal{M}_1) + P(\mathcal{M}_0) = 1$.

2.2.3.3 ad. Conviction ii: Induction requires a logic of partial belief

As the unknowns are considered to be random, we can apply Bayes' rule to yield posterior model probabilities given the observed data, as follows

$$P(\mathcal{M}_1 | d) = \frac{p(d | \mathcal{M}_1)P(\mathcal{M}_1)}{P(d)}, \quad (2.2.1)$$

$$P(\mathcal{M}_0 | d) = \frac{p(d | \mathcal{M}_0)P(\mathcal{M}_0)}{P(d)}, \quad (2.2.2)$$

where $p(d | \mathcal{M}_i)$ is known as the marginal likelihood which represents the “likelihood of the data being generated from model \mathcal{M}_i ”. By taking the ratio of the two expressions above, the common term $P(d)$ drops out yielding the key expression

$$\underbrace{\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_0 | d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}}. \quad (2.2.3)$$

This equation has three crucial ingredients. First, the prior odds quantifies the relative plausibility of \mathcal{M}_1 over \mathcal{M}_0 before any datum is observed. Most researchers enter experiments with prior knowledge, prior experiences, and prior expectations, and these can in principle be used to determine the prior odds. Jeffreys preferred the assumption that both models are equally likely a priori, such that $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$. This is consistent with the Wrinch-Jeffreys simplicity postulate in the sense that prior mass 1/2 is assigned to a parsimonious model (e.g., $\mathcal{M}_0 : \mu = 0$, the general law), and the remaining 1/2 is spread out over a larger model \mathcal{M}_1 where μ is unrestricted. In general then, the prior odds quantify a researcher's initial skepticism about the hypotheses under test. The second ingredient is the posterior odds, which quantifies the relative plausibility of \mathcal{M}_0 and

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

\mathcal{M}_1 after having observed data d . The third ingredient is the Bayes factor (Jeffreys, 1935): the extent to which the observed data d update the prior odds to the posterior odds. For instance, when $\text{BF}_{10}(d) = 9$, the observed data d are 9 times more likely to have occurred under \mathcal{M}_1 than under \mathcal{M}_0 ; when $\text{BF}_{10}(d) = 0.2$, the observed data d are 5 times more likely to have occurred under \mathcal{M}_0 than under \mathcal{M}_1 . The Bayes factor, thus, quantifies the relative probability of the observed data under each of the two competing hypotheses.

Typically, each model \mathcal{M}_i has unknown parameters θ_i that, in accordance to Jeffreys's second conviction, are considered as random with a density given by $\pi_i(\theta_i)$. By the law of total probability the “likelihood of the data being generated from model \mathcal{M}_i ” is then calculated by integrating out the unknown parameters within that model, that is, $p(d | \mathcal{M}_i) = \int f(d | \theta_i, \mathcal{M}_i) \pi_i(\theta_i) d\theta_i$, where $f(d | \theta_i, \mathcal{M}_i)$ is the likelihood, that is, the function that relates the observed data to the unknown parameters θ_i within model \mathcal{M}_i (e.g., Ly et al., 2017c; Myung, 2003). Hence, when we do not know which of two models ($\mathcal{M}_0, \mathcal{M}_1$) generated the observed data and both models contain unknown parameters, we have to specify two prior densities (π_0, π_1) from which we can construct a Bayes factor.

2.2.3.4 ad. Conviction i: Inference is inductive

The specification of the two prior distributions π_0 and π_1 is guided by two desiderata: predictive matching and information consistency. Predictive matching implies that the Bayes factor equals 1 when the data are completely uninformative; information consistency implies that the Bayes factor equals 0 or ∞ when the data are overwhelmingly informative. These desiderata ensure that the correct inference is reached in extreme cases, and in doing so they provide useful restrictions for the specification of the prior distributions.

To achieve the desired result that the Bayes factor equals $\text{BF}_{10}(d) = 1$ for completely uninformative data, the priors π_0 and π_1 need to be chosen such that the marginal likelihoods of \mathcal{M}_0 and \mathcal{M}_1 are predictively matched to each other, that is,

$$\int_{\Theta_0} f(d | \theta_0, \mathcal{M}_0) \pi_0(\theta_0) d\theta_0 = p(d | \mathcal{M}_0) = p(d | \mathcal{M}_1) = \int_{\Theta_1} f(d | \theta_1, \mathcal{M}_1) \pi_1(\theta_1) d\theta_1 \quad (2.2.4)$$

for every completely uninformative data set d .

On the other hand, when data d are overwhelmingly informative in favour of the alternative model the Bayes factor should yield $\text{BF}_{10}(d) = \infty$ or, equivalently, $\text{BF}_{01}(d) = 1/\text{BF}_{10}(d) = 0$, as this then yields $P(\mathcal{M}_1 | d) = 1$ for any prior model probability $P(\mathcal{M}_1) > 0$. A Bayes factor with this property is known to be information consistent.

2.3 Jeffreys's procedure for constructing a default Bayes factor

We now elaborate on Jeffreys's general procedure in constructing default Bayes factors –the specification of the two priors π_0 and π_1 – such that the procedure

fulfils the desiderata discussed above.

2.3.1 Step 1. Nest π_0 within π_1

In null hypothesis tests the model \mathcal{M}_1 can be considered an extension of \mathcal{M}_0 by inclusion of a new parameter, that is, $\theta_1 = (\theta_0, \zeta)$ where θ_0 denotes the common parameters and ζ the test-relevant parameter. Equivalently, \mathcal{M}_0 is said to be nested within \mathcal{M}_1 due to the connection $f(d | \theta_0, \mathcal{M}_0) = f(d | \theta_0, \zeta = 0, \mathcal{M}_1)$. Jeffreys exploited the connection between these two likelihood functions to induce a relationship between π_1 and π_0 . In general one has $\pi_1(\theta_0, \zeta) = \pi_1(\zeta | \theta_0)\pi_1(\theta_0)$, but due to the nesting Jeffreys treats the common parameters within \mathcal{M}_1 as in \mathcal{M}_0 , that is, $\pi_1(\theta_0) = \pi_0(\theta_0)$. Furthermore, when ζ can be sensibly related to θ_0 , Jeffreys redefines the test-relevant parameter as δ , and decomposes the prior as $\pi_1(\theta_0, \zeta) = \pi_1(\delta)\pi_0(\theta_0)$. For instance, in the case of the t -test Jeffreys focuses on effect size $\delta = \frac{\mu}{\sigma}$.

This implies that once we have chosen π_0 , we have then completely specified the marginal likelihood $p(d | \mathcal{M}_0)$ and can, therefore, readily calculate the denominator of the Bayes factor $\text{BF}_{10}(d)$ given data d . Furthermore, due to the nesting of π_0 within π_1 we can also calculate a large part of the marginal likelihood of \mathcal{M}_1 as

$$p(d | \mathcal{M}_1) = \int_{\Delta} \underbrace{\int_{\Theta} f(d | \theta_0, \delta, \mathcal{M}_1) \pi_0(\theta_0) d\theta_0}_{h(d | \delta)} \pi_1(\delta) d\delta, \quad (2.3.1)$$

where $h(d | \delta)$ is the test-relevant likelihood, a function that only depends on the data and the test-relevant parameter δ as the common parameters θ_0 are integrated out. The following two steps are concerned with choosing $\pi_1(\delta)$ such that the resulting Bayes factor is well-calibrated to extreme data.

2.3.2 Step 2. Predictive matching

Typically, a certain minimum number of samples n_{\min} is required before model \mathcal{M}_1 can be differentiated from \mathcal{M}_0 . All possible data sets with sample sizes less than n_{\min} are considered uninformative. For example, at least $n_{\min} = 2$ observations are required to distinguish $\mathcal{M}_0 : \mu = 0$ from \mathcal{M}_1 in a t -test. Specifically, confronted with a single Gaussian observation unequal to zero, for instance, $x_1 = 5$, lack of knowledge about σ within \mathcal{M}_0 means that we cannot exclude \mathcal{M}_0 as a reasonable explanation for the data.

Indeed, a member of \mathcal{M}_0 , a zero-mean normal distribution with a standard deviation of seven produces an observation less than five units away from zero with 53% chance. Similarly, lack of knowledge about σ also means that \mathcal{M}_1 cannot be excluded as a reasonable explanation of the data. To convey that –for the purpose of discriminating \mathcal{M}_0 from \mathcal{M}_1 – nothing is learned from any data set with a sample smaller than n_{\min} we choose $\pi_1(\delta)$ such that

$$p(d | \mathcal{M}_0) = p(d | \mathcal{M}_1) = \int_{\Delta} h(d | \delta) \pi_1(\delta) d\delta \quad (2.3.2)$$

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

for every data set d with a sample size less than n_{\min} . In sum, $\pi_1(\delta)$ is chosen such that when the data are completely uninformative we get $\text{BF}_{10}(d) = 1$.

2.3.3 Step 3. Information consistency

Even a limited number of observations may provide overwhelming support for \mathcal{M}_1 . In the case of the t -test, for instance, the support that an observed non-zero mean provides for \mathcal{M}_1 should increase without bound as the sample variance $s^2 \rightarrow 0$, for any sample size greater or equal to n_{\min} . Consequently, for data d with a sample size greater or equal to n_{\min} that point undoubtedly to \mathcal{M}_1 Jeffreys chose $\pi_1(\delta)$ such that $p(d | \mathcal{M}_1)$ diverges to infinity. That is, in order to achieve information consistency $p(d | \mathcal{M}_0)$ needs to be bounded and $\pi_1(\delta)$ needs to be chosen such that $p(d | \mathcal{M}_1) = \int_{\Delta} h(d | \delta) \pi_1(\delta) d\delta$ diverges to infinity for overwhelmingly informative data of any size n greater or equal to n_{\min} .

2.3.4 Summary

Jeffreys's procedure to construct a Bayes factor nests π_0 within π_1 , and therefore the choice of π_0 is the starting point of the method. The specification of π_0 yields $p(d | \mathcal{M}_0)$. Next, the test-relevant prior π_1 is chosen such that $p(d | \mathcal{M}_1)$ is well-calibrated to extreme data that are either completely uninformative or overwhelmingly informative. Together with π_0 , this calibrated test-relevant prior forms the basis for Jeffreys's construction of a Bayes factor.

As a default choice for π_0 , Jeffreys used his popular "Jeffreys's prior" on the common parameters θ_0 (Jeffreys, 1946). Derived from the likelihood function $f(d | \theta_0, \mathcal{M}_0)$, this default prior is parameterisation invariant, meaning that the same posterior is obtained regardless of how the parameters are represented (e.g., Ly et al., 2017c). Jeffreys's parameterisation-invariant priors are typically improper, that is, non-normalisable, even though they do lead to proper posteriors for the designs discussed below.

The specification of the test-relevant prior requires special care, as priors that are too wide inevitably reduce the weighted likelihood, resulting in a preference for \mathcal{H}_0 regardless of the observed data (Jeffreys-Lindley-Bartlett paradox; Bartlett, 1957; Jeffreys, 1961; Lindley, 1957; Marin and Robert, 2010). Consequently, Jeffreys's parameterisation-invariant prior cannot be used for the test-relevant parameter.

Note that Jeffreys's methodical approach in choosing the two priors π_0 and π_1 is fully based on the likelihood functions of the two models that are being compared; the priors do not represent substantive knowledge of the parameters within the model and the resulting procedure can therefore be presented as a reference analysis that may be fine-tuned in the presence of additional information. In the following two sections we illustrate Jeffreys's procedure by discussing the development of the default Bayes factors for two scenarios that are particularly relevant for experimental psychology: testing the nullity of a normal mean and testing the nullity of a correlation coefficient. Appendix 2.A provides a list of additional Bayes factors that are presented in ToP.

2.4 Jeffreys's Bayes factor for the test of the nullity of a normal mean: The Bayesian t -test

To develop the Bayesian counterpart of the classical t -test we first characterise the data and discuss how they relate to the unknown parameters within each model in terms of the likelihood functions. By studying the likelihood functions we can justify the nesting π_0 within π_1 , and identify data that are completely uninformative, as well as data that are overwhelmingly informative. The test-relevant prior is then selected based on the desiderata discussed above. We then apply the resulting default Bayes factor to an example data set on cheating and creativity. In addition, we develop the one-sided extension of Jeffreys's t -test, after which we conclude with a brief discussion.

2.4.1 Normal data

For the case at hand, experimental outcomes are assumed to follow a normal distribution characterised by the unknown population mean μ and standard deviation σ . Similarly, the observed data d from a normal distribution can be summarised by two numbers: the observed sample mean \bar{x} and the average sums of squares $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$; hence we write $d = (\bar{x}, s_n^2)$. The main difference between the null model $\mathcal{M}_0 : \mu = 0$ and its relaxation \mathcal{M}_1 is reflected in the population effect size, which is defined as $\delta = \frac{\mu}{\sigma}$, as σ provides a scale to the problem. This population effect size cannot be observed directly, unlike its sampled scaled version the t -statistic, i.e., $t = \frac{\sqrt{n}\bar{x}}{s_\nu}$, where s_ν refers to the sample standard deviation based on $\nu = n - 1$ degrees of freedom. Extreme data can be characterised by $|t| \rightarrow \infty$ or equivalently by $s_n^2 \rightarrow 0$ and it is used in the calibration step of the Bayes factor to derive the test-relevant prior. To improve readability we remove the subscript n when we refer to the average sum of squares $s^2 = s_n^2$.

2.4.2 Step 1. Nesting of π_0 within π_1

2.4.2.1 Comparing the likelihood functions

A model defines a likelihood that structurally describes how the observed data are related to the unknown parameters. The point null hypothesis \mathcal{M}_0 posits that $\mu = 0$, whereas the alternative hypothesis \mathcal{M}_1 relaxes the restriction on μ . Conditioned on the observations $d = (\bar{x}, s^2)$, the likelihood functions of \mathcal{M}_0 and \mathcal{M}_1 are given by

$$f(d | \sigma, \mathcal{M}_0) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2} [\bar{x}^2 + s^2]\right), \quad (2.4.1)$$

$$f(d | \mu, \sigma, \mathcal{M}_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2} [(\bar{x} - \mu)^2 + s^2]\right), \quad (2.4.2)$$

respectively. Note that $f(d | \sigma, \mathcal{M}_0)$ is a function of σ alone, whereas $f(d | \mu, \sigma, \mathcal{M}_1)$ depends on two parameters: σ and μ . As σ is a scaling parameter in both models, we can set $\pi_1(\mu, \sigma) = \pi_1(\mu | \sigma)\pi_0(\sigma)$. Jeffreys removed the scale from the problem by considering $\delta = \frac{\mu}{\sigma}$ as the test-relevant parameter and with $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$

we get the following marginal likelihood

$$p(d | \mathcal{M}_1) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \int_{-\infty}^\infty \exp \left(-\frac{n}{2} \left[\left(\frac{\bar{x}}{\sigma} - \delta \right)^2 + \left(\frac{s}{\sigma} \right)^2 \right] \right) \pi_1(\delta) d\delta \pi_0(\sigma) d\sigma, \quad (2.4.3)$$

for the encompassing model \mathcal{M}_1 .

2.4.2.2 The denominator of $\text{BF}_{10}(d)$

Jeffreys's default choice leads to $\pi_0(\sigma) \propto 1/\sigma$, the parameterisation-invariant prior that Jeffreys's would use to arrive at a posterior for σ within either model. This prior specification leads to the following marginal likelihood of \mathcal{M}_0

$$p(d | \mathcal{M}_0) = \begin{cases} \frac{1}{2|\bar{x}|} & n = 1, \\ \frac{\Gamma(\frac{n}{2})}{2(n\pi\bar{x}^2)^{\frac{n}{2}}} & n > 1 \text{ and } s^2 = 0, \\ \frac{\Gamma(\frac{n}{2})}{2(\pi ns^2)^{\frac{n}{2}}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} & n > 1 \text{ and } s^2 > 0, \end{cases} \quad (2.4.4), \quad (2.4.5), \quad (2.4.6)$$

where t is the observed t -value and ν the degrees of freedom defined as before. Hence, Eqns. (2.4.4, 2.4.5, 2.4.6) define the denominator of the Bayes factor $\text{BF}_{10}(d)$; Eq. (2.4.4) will be used for calibrating the Bayes factor $\text{BF}_{10}(d)$ to completely uninformative data, whereas Eq. (2.4.5) will be used for the calibration with respect to overwhelmingly informative data. Some statisticians only report the right term $\left(1 + \frac{t^2}{\nu}\right)^{-\frac{n}{2}}$ of Eq. (2.4.6), as the first factor also appears in the marginal likelihood of \mathcal{M}_1 and, thus, cancels out in the Bayes factor.

2.4.3 Step 2. Predictive matching: Symmetric $\pi_1(\delta)$

We now discuss how the test-relevant prior $\pi_1(\delta)$ can be chosen such that the resulting Bayes factor is well-calibrated. As elaborated above, we consider data sets with only one sample as completely uninformative in discriminating \mathcal{M}_0 from \mathcal{M}_1 . Jeffreys (1961, p. 269) studied Eq. (2.4.3) with $n = 1$, $\bar{x} > 0$, and, consequently, $s^2 = 0$, and concluded that $p(d | \mathcal{M}_1)$ is matched to Eq. (2.4.4) whenever $\pi_1(\delta)$ is symmetric around zero.

2.4.4 Step 3. Information consistency: Heavy-tailed $\pi_1(\delta)$

On the other hand, observed data $\bar{x} > 0, s^2 = 0$ with $n > 1$ can be considered overwhelmingly informative as the t -value is then infinite. To obtain maximum evidence in favour of the alternative we require that $\text{BF}_{10}(d) = \infty$. This occurs whenever the marginal likelihood of \mathcal{M}_1 is infinite and $p(d | \mathcal{M}_0)$ finite, see Eq. (2.4.5). Jeffreys (1961, pp. 269–270) showed that this is the case whenever the test-relevant prior $\pi_1(\delta)$ is heavy-tailed.

2.4.5 The resulting Bayes factor

Hence, a Bayes factor that meets Jeffreys's desiderata can be obtained by assigning $\pi_0(\sigma) \propto 1/\sigma$ and $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$, where $\pi_1(\delta)$ is symmetric around zero and heavy-tailed.

2.4.5.1 Jeffreys's default choice: The standard Cauchy distribution

The Cauchy distribution with scale γ is the most well-known distribution which is both symmetric around zero and heavy-tailed

$$\pi_1(\delta; \gamma) = \frac{1}{\pi\gamma \left(1 + \left(\frac{\delta}{\gamma}\right)^2\right)}. \quad (2.4.7)$$

As a default choice for $\pi_1(\delta)$, Jeffreys suggested to use the simplest version, the standard Cauchy distribution with $\gamma = 1$.

2.4.5.2 Jeffreys's Bayesian t -test

Jeffreys's Bayes factor now follows from the integral in Eq. (2.4.3) with respect to Cauchy distributions $\pi_1(\delta)$ divided by Eq. (2.4.6), whenever $n > 1$ and $s^2 > 0$. Jeffreys knew that this integral is hard to compute and went to great lengths to compute an approximation that makes his Bayesian t -test usable in practice. Fortunately, we can now take advantage of computer software that can numerically solve the aforementioned integral and we therefore omit Jeffreys's approximation from further discussion. By a decomposition of a Cauchy distribution we obtain a Bayes factor of the following form:

$$BF_{10;\gamma}(n, t) = \frac{\gamma \int_0^\infty (1+ng)^{-\frac{1}{2}} \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} (2\pi)^{-\frac{1}{2}} g^{-\frac{3}{2}} e^{-\frac{\gamma^2}{2g}} dg}{(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}}, \quad (2.4.8)$$

where g is an auxiliary variable that is integrated out numerically. Jeffreys's choice is obtained when $\gamma = 1$. The Bayes factor $BF_{10;\gamma=1}(n, t)$ now awaits a user's observed t -value and the associated n number of observations.

2.4.6 Example: The Bayesian between-subject t -test

To illustrate the default Bayesian t -test we extend Eq. (2.4.8) to a between-subjects design and apply the test to a psychological data set. The development above is easily generalised to a between-subject design in which observations are assumed to be drawn from two separate normal populations. To do so, we replace: (i) the value of t by the observed two-sample t value, (ii) the effective sample size by $n = \frac{n_1 n_2}{n_1 + n_2}$, and (iii) the degrees of freedom with $\nu = n_1 + n_2 - 2$ (Rouder et al., 2009).

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

Example 2.4.1 (Does cheating enhance creativity?). *Gino and Wiltermuth (2014, Experiment 2) reported that the act of cheating enhances creativity. This conclusion was based on five experiments. Here we analyse the results from Experiment 2 in which, having been assigned either to a control condition or to a condition in which they were likely to cheat, participants were rewarded for correctly solving each of 20 maths and logic multiple-choice problems. Next, participants' creativity level was measured by having them complete 12 problems from the Remote Association Task (RAT; Mednick, 1962).*

The control group featured $n_1 = 48$ participants who scored an average of $\bar{x}_1 = 4.65$ RAT items correctly with a sample standard deviation of $s_{n_1-1} = 2.72$. The cheating group featured $n_2 = 51$ participants who scored $\bar{x}_2 = 6.20$ RAT items correctly with $s_{n_2-1} = 2.98$. These findings yield $t(97) = 2.73$ with $p = .008$. Jeffreys's default Bayes factor yields $\text{BF}_{10}(d) \approx 4.6$, indicating that the data are 4.6 times more likely under \mathcal{M}_1 than under \mathcal{M}_0 . With equal prior odds, the posterior probability for \mathcal{M}_0 remains an arguably non-negligible 17%.

For nested models, the Bayes factor can also be obtained using the Savage-Dickey density ratio test (e.g., Dickey and Lientz, 1970; Wagenmakers et al., 2010; Marin and Robert, 2010). The Savage-Dickey test is based on the identity

$$\text{BF}_{10}(d) = \frac{\pi_1(\delta = 0)}{\pi_1(\delta = 0 | d)}. \quad (2.4.9)$$

One of the additional advantages of the Savage-Dickey test is that it allows the result of the test to be displayed visually, as the height of the prior versus the posterior at the point of test (i.e., $\delta = 0$). Fig. 2.1 presents the results from Experiment 2 of Gino and Wiltermuth (2014). \diamond

In this example, both the Bayesian and Fisherian analysis gave the same qualitative result. Nevertheless, the Bayes factor is more conservative, and some researchers may be surprised that, for the same data with $p = .008$ we get a posterior model probability of $P(\mathcal{M}_0 | d) = .17$, if the two hypotheses were equal probable a priori. Indeed, for many cases the Bayesian and Fisherian analyses disagree qualitatively as well as quantitatively (e.g., Wetzels et al., 2011).

2.4.7 The one-sided extension of Jeffreys's Bayes factor

Some reflection suggests that the authors' hypothesis from Example 2.4.1 is more specific – the authors argued that cheating leads to more creativity, not less. To account for the directionality of the hypothesis we need a one-sided adaptation of Jeffreys's Bayes factor $\text{BF}_{10;\gamma=1}(n, t)$. The comparison that is made is then between the model of no effect \mathcal{M}_0 and one denoted by \mathcal{M}_+ in which the effect size δ is assumed to be positive. We factorise $\text{BF}_{+0}(d)$ as

$$\text{BF}_{+0}(d) = \underbrace{\frac{p(d | \mathcal{M}_+)}{p(d | \mathcal{M}_1)}}_{\text{BF}_{+1}(d)} \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)}}_{\text{BF}_{10}(d)}, \quad (2.4.10)$$

where $\text{BF}_{+1}(d)$ is the Bayes factor that compares the unconstrained model \mathcal{M}_1 to the positively restricted model \mathcal{M}_+ (Morey and Wagenmakers, 2014; Mulder

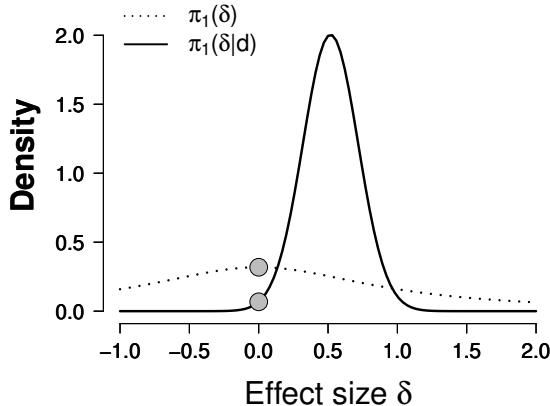


Figure 2.1: Posterior and prior distributions of the effect size for a two-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $BF_{10; \gamma=1} \approx 4.60$ equals the height ratio of the prior distribution $\pi_1(\delta)$ divided by the posterior distribution $\pi_1(\delta | d)$ at $\delta = 0$.

et al., 2010; Pericchi et al., 2008). The objective comparison between \mathcal{M}_+ and \mathcal{M}_1 is then to keep all aspects the same $\pi_+(\sigma) = \pi_1(\sigma) = \pi_0(\sigma)$, except for the distinguishing factor of δ being restricted to positive values within \mathcal{M}_+ . For the test-relevant prior distribution we restrict $\pi_1(\delta)$ to positive values of δ , which by symmetry of the Cauchy distribution means that $\pi_+(\delta)$ accounts doubly for the likelihood when δ is positive and nullifies it when δ is negative (Klugkist et al., 2005).

Example 2.4.2 (One-sided test for the Gino and Wiltermuth data). *For the data from Gino and Wiltermuth (2014, Experiment 2) the one-sided adaptation of Jeffreys's Bayes factor Eq. (2.4.8) yields $BF_{+0}(d) = 9.18$. Because almost all of the posterior mass is consistent with the authors' hypothesis, the one-sided Bayes factor is almost twice the two-sided Bayes factor. The result is visualised through the Savage-Dickey ratio in Fig. 2.2.* ◇

2.4.8 Discussion on the t -test

In this section we showcased Jeffreys's procedure in selecting the instrumental priors π_0, π_1 that yield a Bayes factor for grading the support that the data provide for \mathcal{M}_0 versus \mathcal{M}_1 . The construction of this Bayes factor began by assigning Jeffreys's parameterisation-invariant prior to the common parameters, that is, $\pi_0(\sigma) \propto 1/\sigma$. This is the same prior Jeffreys would use for estimating σ in either of the two models, when no doubt is present on the validity of the model itself. This prior on the common parameters then yields the denominator of the Bayes factor Eqns. (2.4.4, 2.4.5, 2.4.6). Jeffreys noted that when the test-relevant prior $\pi_1(\delta)$

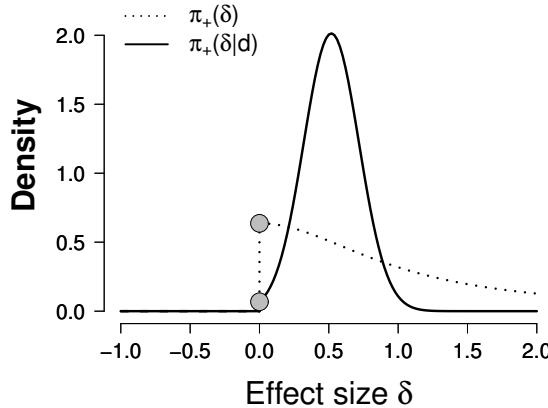


Figure 2.2: Posterior and prior distributions of effect size for a one-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $\text{BF}_{+0} = 9.18$ equals the height ratio of the prior distribution $\pi_+(\delta)$ over the posterior distribution $\pi_+(\delta | d)$ at $\delta = 0$. The prior distribution $\pi_+(\delta)$ is zero for negative values of δ . Furthermore, note that the prior distribution for $\delta \geq 0$ is twice as high compared to $\pi_1(\delta)$ in Fig. 2.1.

is symmetric and heavy tailed, the resulting Bayes factor is guaranteed to yield the correct conclusion for completely uninformative data and for overwhelmingly informative data. Jeffreys (1961, pp. 272–273) noted that the standard Cauchy prior for δ yields a Bayes factor Eq. (2.4.8) (with $\gamma = 1$) that aligns with this calibration.

It took several decades before Jeffreys's Bayes factor for the t -test was adopted by Zellner and Siow (1980), who generalised it to the linear regression framework based on a multivariate Cauchy distribution. One practical drawback of their proposal was the fact that the numerical integration required to calculate the Bayes factor becomes computationally demanding as the number of covariates grows.

Liang et al. (2008) proposed a computationally efficient alternative to the Zellner and Siow (1980) setup by first decomposing the multivariate Cauchy distribution into a mixture of gamma and normal distributions followed by computational simplifications introduced by Zellner (1986). As a result, the Bayes factor can be obtained from only a single numerical integral, regardless of the number of covariates. The form of the numerator in Eq. (2.4.8) is in fact inspired by Liang et al. (2008) and introduced to psychology by Rouder et al. (2009) and Wetzels et al. (2009). The combination $\pi_0(\sigma) \propto \sigma^{-1}$ and $\delta \sim \mathcal{C}(0, 1)$ was dubbed the JZS-prior in honour of Jeffreys, Zellner, and Siow; this is understandable in the framework of linear regression, although it should be noted that all ideas for the t -test were already present in the second edition of ToP (Jeffreys, 1948, pp. 242–248).

2.4.8.1 Model selection consistency

In addition to predictive matching and information consistency, Liang et al. (2008) showed that Zellner and Siow's 1980 generalisation of Jeffreys's work is also model selection consistent, which implies that as the sample size n increases indefinitely, the support that the data d provide for the correct data-generating model (i.e., \mathcal{M}_0 or \mathcal{M}_1) grows without bound. Hence, Jeffreys's default Bayes factor Eq. (2.4.8) leads to the correct decision whenever the sample size is sufficiently large. Jeffreys's procedure of assigning default priors for Bayesian hypothesis testing was recently generalised by Bayarri et al. (2012). We now turn to Jeffreys's development of another default Bayes factor: the test for the presence of a correlation.

2.5 Jeffreys's Bayes factor for the test of the nullity of a correlation

To develop the Bayesian correlation test we first characterise the data and discuss how they relate to the unknown parameters within each model in terms of the likelihood functions. By studying the likelihood functions we can justify the nesting of π_0 within π_1 and identify data that are completely uninformative and data that are overwhelmingly informative. As was done for the Bayesian t -test, the test-relevant prior is selected based on a calibration argument. The derivations and calibrations given here cannot be found in Jeffreys (1961), as Jeffreys appears to have derived the priors intuitively. Hence, we divert from the narrative of Jeffreys (1961, Paragraph 5.5) and instead: (a) explain Jeffreys's reasoning with a structure analogous to that of the previous section; and (b) give the exact results instead, as Jeffreys used an approximation to simplify the calculations. In effect, we show that Jeffreys's intuitive choice is very close to our exact result. After presenting the correlation Bayes factor we relate it to Jeffreys's choice and apply the resulting default Bayes factor to an example data set that is concerned with presidential height and the popular vote. In addition, we develop the one-sided extension of Jeffreys's correlation test, after which we conclude with a brief discussion.

2.5.1 Bivariate normal data

For the case at hand, experimental outcome pairs (X, Y) are assumed to follow a bivariate normal distribution characterised by the unknown population means μ_x, μ_y , standard deviations σ_x, σ_y of X and Y respectively. Within \mathcal{M}_1 the additional parameter ρ characterises the linear association between X and Y . To test the nullity of the population correlation ρ it is helpful to summarise the data for X and Y separately in terms of their respective sample means and average sums of squares: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$, respectively. The sample correlation coefficient r then defines an observable measure of the linear relationship between X and Y

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \quad (2.5.1)$$

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

This sample correlation coefficient r is an imperfect reflection of the unobservable population correlation coefficient ρ . In sum, the data can be summarised by the five quantities $d = (\bar{x}, s_x^2, \bar{y}, s_y^2, r)$.

The main difference between the null model \mathcal{M}_0 and \mathcal{M}_1 is reflected in the population correlation coefficient ρ , which cannot be observed directly, unlike its sampled version known as Pearson's r , Eq. (2.5.1). Extreme data can be characterised by $|r|=1$ and this is used in the calibration step of the Bayes factor to derive the form of the test-relevant prior.

2.5.2 Step 1. Nesting of π_0 within π_1

2.5.2.1 Comparing the likelihood functions

The point null hypothesis \mathcal{M}_0 assumes that the data follow a bivariate normal distribution with ρ known and fixed at zero. Hence, \mathcal{M}_0 depends on four parameters which we abbreviate as $\theta_0 = (\mu_x, \mu_y, \sigma_x, \sigma_y)$, while the alternative model \mathcal{M}_1 can be considered an extension of \mathcal{M}_0 with an additional parameter ρ , i.e., $\theta_1 = (\theta_0, \rho)$. These two bivariate normal models relate the observed data to the parameters using the following two likelihood functions

$$f(d | \theta_0, \mathcal{M}_0) = (2\pi\sigma_x\sigma_y)^{-n} \exp \left(-\frac{n}{2} \left[\left(\frac{\bar{x} - \mu_x}{\sigma_x} \right)^2 + \left(\frac{\bar{y} - \mu_y}{\sigma_y} \right)^2 \right] \right), \quad (2.5.2)$$

$$\times \exp \left(-\frac{n}{2} \left[\left(\frac{s_x}{\sigma_x} \right)^2 + \left(\frac{s_y}{\sigma_y} \right)^2 \right] \right),$$

$$f(d | \theta_1, \mathcal{M}_1) = (2\pi\sigma_x\sigma_y\sqrt{1-\rho^2})^{-n} \quad (2.5.3)$$

$$\times \exp \left(-\frac{n}{2(1-\rho^2)} \left[\frac{(\bar{x} - \mu_x)^2}{\sigma_x^2} - 2\rho \frac{(\bar{x} - \mu_x)(\bar{y} - \mu_y)}{\sigma_x\sigma_y} + \frac{(\bar{y} - \mu_y)^2}{\sigma_y^2} \right] \right),$$

$$\times \exp \left(-\frac{n}{2(1-\rho^2)} \left[\left(\frac{s_x}{\sigma_x} \right)^2 - 2\rho r \left(\frac{s_x s_y}{\sigma_x \sigma_y} \right) + \left(\frac{s_y}{\sigma_y} \right)^2 \right] \right),$$

respectively. Note that $f(d | \theta_0, \mathcal{M}_0) = f(d | \theta_0, \rho = 0, \mathcal{M}_1)$ and because the population correlation ρ is defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{E(XY) - \mu_x\mu_y}{\sigma_x\sigma_y}, \quad (2.5.4)$$

we know that ρ remains the same under data transformations of the form $\tilde{X} = aX - b$, $\tilde{Y} = cY - d$. In particular, we can take $b = \mu_x$, $d = \mu_y$, $a = 1/\sigma_x$, $c = 1/\sigma_y$ and conclude that ρ does not depend on these common parameters θ_0 . Hence, we nest π_0 within π_1 orthogonally, that is, $\pi_1(\theta_0, \rho) = \pi_1(\rho)\pi_0(\theta_0)$.

2.5.2.2 The denominator of $\text{BF}_{10}(d)$

Jeffreys's default choice leads to assigning $\pi_0(\theta_0)$ the joint prior $\pi_0(\mu_x, \mu_y, \sigma_x, \sigma_y) = 1 \cdot 1 \cdot \frac{1}{\sigma_x} \frac{1}{\sigma_y}$; this is the parameterisation-invariant prior that Jeffreys would use to update to the posterior for θ_0 within either model. When the averaged sum of

squares are both non-zero, this yields the following marginal likelihood of \mathcal{M}_0

$$p(d | \mathcal{M}_0) = 2^{-2} n^{-n} \pi^{1-n} (s_x s_y)^{1-n} \left[\Gamma\left(\frac{n-1}{2}\right) \right]^2. \quad (2.5.5)$$

Eq. (2.5.5) defines the denominator of the correlation Bayes factor $\text{BF}_{10}(d)$. Observe that this marginal likelihood does not depend on the sample correlation coefficient r .

2.5.3 Step 2. Predictive matching: Symmetric $\pi_1(\rho)$

2.5.3.1 Deriving the test-relevant likelihood function

We now discuss how the test-relevant prior $\pi_1(\rho)$ can be defined such that the resulting Bayes factor is well-calibrated. The conclusion is as before: we require $\pi_1(\rho)$ to be symmetric around zero. We discuss the result more extensively as it cannot be found in Jeffreys (1961). Furthermore, the test-relevant likelihood function reported by Jeffreys (1961, p. 291, Eq. 8) is in fact an approximation of the result given below.

Before we can discuss the calibration we first derive the test-relevant likelihood function by integrating out the common parameters θ_0 from Eq. (2.5.3) with respect to the parameterisation-invariant priors $\pi_0(\theta_0)$ as outlined by Eq. (2.3.1). This leads to the following simplification

$$p(d | \mathcal{M}_1) = p(d | \mathcal{M}_0) \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho, \quad (2.5.6)$$

where h is the test-relevant likelihood function that depends on n, r, ρ alone and is given by Eqns. (2.5.8, 2.5.9) below. The Bayes factor, therefore, reduces to

$$\text{BF}_{10}(d) = \frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)} = \frac{p(d | \mathcal{M}_0) \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho}{p(d | \mathcal{M}_0)} = \int_{-1}^1 h(n, r | \rho) \pi_1(\rho) d\rho. \quad (2.5.7)$$

Note that whereas $p(d | \mathcal{M}_0)$ does not depend on ρ or the statistic r , i.e., Eq. (2.5.5), the function h does not depend on the statistics $\bar{x}, s_x^2, \bar{y}, s_y^2$ that are associated with the common parameters. Thus, the evidence for \mathcal{M}_1 over \mathcal{M}_0 resides within n and r alone.

The test-relevant likelihood function $h(n, r | \rho)$ possess more regularities. In particular, it can be decomposed into an even and an odd function, that is, $h = A + B$, with A defined as

$$A(n, r | \rho) = (1 - \rho^2)^{\frac{n-1}{2}} {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{1}{2}; (r\rho)^2\right), \quad (2.5.8)$$

where ${}_2F_1$ is Gauss' hypergeometric function (see Appendix 2.B for details). Observe that A is a symmetric function of ρ when n and r are given. The second function B is relevant for the one-sided test and is given by

$$B(n, r | \rho) = 2r\rho(1 - \rho^2)^{\frac{n-1}{2}} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right]^2 {}_2F_1\left(\frac{n}{2}, \frac{n}{2}; \frac{3}{2}; (r\rho)^2\right), \quad (2.5.9)$$

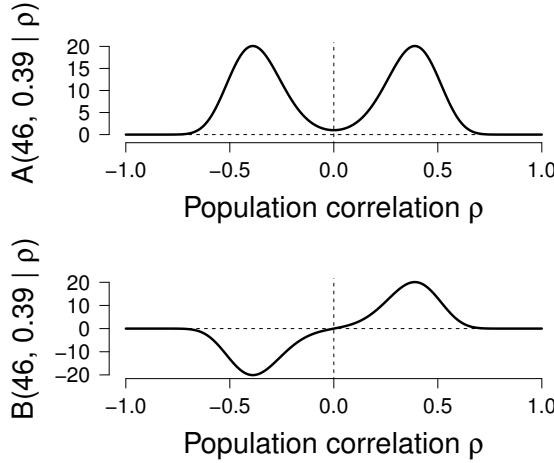


Figure 2.3: $A(n, r | \rho)$ is an even function of ρ , and $B(n, r | \rho)$ is an odd function of ρ . Together, A and B determine the function h from Eq. (2.5.7): $h(n, r | \rho) = A(n, r | \rho) + B(n, r | \rho)$. For this illustration, we used $n = 46$ and $r = 0.39$ based on the example data discussed below.

which is an odd function of ρ when n and r are given. Thus, the test-relevant likelihood function h that mediates inference about the presence of ρ from n and r is given by $h(n, r | \rho) = A(n, r | \rho) + B(n, r | \rho)$. Examples of the functions A and B are shown in Fig. 2.3.

2.5.3.2 Predictive matching and the minimal sample size of $n_{\min} = 3$

Interestingly, the predictive matching principle implies the use of a symmetric test-relevant prior as in the previous case. Note that we cannot infer the correlation of a bivariate normal distribution whenever we have only a single data pair (x, y) ; r is undefined when $n = 1$. Furthermore, when $n = 2$ we automatically get $r = 1$ or $r = -1$ regardless of the actually observations and how they were generated. As such, nothing is learned up to $n_{\min} = 3$ when testing the nullity of ρ . Hence, we have to choose $\pi_1(\rho)$ such that the resulting Bayes factor Eq. (2.5.7) equals one for $n = 1$ and $n = 2$ regardless of the actually observed r .

Using $n = 1$ in Eq. (2.5.8) and Eq. (2.5.9) we see that the reduced likelihood is the constant function, that is,

$$h(1, r | \rho) = A(1, r | \rho) + B(1, r | \rho) = 1 \quad (2.5.10)$$

for every ρ and r . From a consideration of Eq. (2.5.7) it follows that for a Bayes factor of one with $n = 1$, we require $\pi_1(\rho)$ to integrate to one (i.e., $\text{BF}_{10}(d) = \int_{-1}^1 \pi_1(\rho) d\rho = 1$), underscoring Jeffreys's claim that the test-relevant priors should be proper.¹ Similarly, for $n = 2$ we automatically obtain $|r| = 1$ and plugging this

¹Jeffreys rejected the parameterisation-invariant prior $\rho \propto (1 - \rho^2)^{-1}$ because it leads to

into Eq. (2.5.8) yields $A(2, |r|=1 | \rho) = 1$. Thus, with $\pi_1(\rho)$ a proper prior this yields a Bayes factor of $\text{BF}_{10}(d) = 1 + \int_{-1}^1 B(2, |r|=1 | \rho) \pi_1(\rho) d\rho$. To ensure that the Bayes factor equals one for data with a sample size of $n = 2$ we have to nullify the contribution of the function $B(2, |r|=1 | \rho)$. This occurs when $\pi_1(\rho)$ is symmetric around zero, since $B(2, r | \rho)$ is an odd function of ρ , see Fig. 2.3.

2.5.4 Step 3. Information consistency

On the other hand, a sample correlation $r = 1$ or $r = -1$ with $n \geq n_{\min} = 3$ can be considered overwhelmingly informative data in favour of the alternative model \mathcal{M}_1 . In our quest to find the right test-relevant prior that yields a Bayes factor that is information consistent, we consider the so-called symmetric stretched beta distributions given by

$$\pi_1(\rho; \kappa) = \frac{2^{\frac{\kappa-2}{\kappa}}}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})} (1 - \rho^2)^{\frac{1-\kappa}{\kappa}}, \quad (2.5.11)$$

where $\mathcal{B}(1/\kappa, 1/\kappa)$ is a beta function, see Appendix 2.C for details. Each $\kappa > 0$ yields a candidate test-relevant prior. Jeffreys's intuitive choice is represented by Eq. (2.5.11) with $\kappa = 1$, as this choice corresponds to the uniform distribution of ρ on $(-1, 1)$. Furthermore, κ can be thought of as a scale parameter of the prior as in Eq. (2.4.7). We claim that a Bayes factor based on a test-relevant prior Eq. (2.5.11) with $\kappa \geq 2$ is information consistent.

2.5.5 The resulting Bayes factor

To prove the information consistency claim, ρ is integrated out of the test-relevant likelihood with $h = A + B$ as discussed above, Eq. (2.5.7). This results in the following closed form Bayes factor:

$$\begin{aligned} \text{BF}_{10; \kappa}(n, r) &= \int_{-1}^1 h(n, r | \rho) \pi_1(\rho; \kappa) d\rho \\ &= \int_{-1}^1 A(n, r | \rho) \pi_1(\rho; \kappa) d\rho + \overbrace{\int_{-1}^1 B(n, r | \rho) \pi_1(\rho; \kappa) d\rho}^0 \\ &= \frac{2^{\frac{\kappa-2}{\kappa}} \sqrt{\pi}}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})} \frac{\Gamma\left(\frac{2+(n-1)\kappa}{2\kappa}\right)}{\Gamma\left(\frac{2+n\kappa}{2\kappa}\right)} {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{2+n\kappa}{2\kappa}; r^2\right), \end{aligned} \quad (2.5.12)$$

where the contribution of the B -function is nullified due to symmetry of the prior. We call Eq. (2.5.12) Jeffreys's exact correlation test, as we believe that Jeffreys would have derived this Bayes factor $\text{BF}_{10; \kappa}(n, r)$, if he had deemed it necessary to calculate it exactly.

unwelcome results when testing the null hypothesis $\rho = 1$. However, Robert et al. (2009) noted that such a test is rather uncommon as interest typically centres on the point null hypothesis $\mathcal{M}_0 : \rho = 0$.

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

Table 2.1 lists the Bayes factors for a selection of values for κ and n with $r = 1$ fixed; the results confirm that the Bayes factor is indeed information consistent when $\kappa \geq 2$. Note that Jeffreys's choice of $\kappa = 1$ does not lead to a Bayes factor which provides extreme support for \mathcal{M}_1 when confronted with data that are overwhelmingly informative (i.e., $r = 1$ and $n_{\min} = 3$). However, this Bayes factor does diverge when $n \geq 4$. Thus, Jeffreys's intuitive choice for κ misses the information consistency criterion by one data pair. The resulting Bayes factor $\text{BF}_{10;\kappa}(n,r)$ now awaits a user's observed r -value and the associated n number of observations. In what follows, we honour Jeffreys's intuition and showcase the correlation Bayes factor using Jeffreys's choice $\kappa = 1$.

Table 2.1: The Bayes factor $\text{BF}_{10;\kappa=2}$ is information consistent as it diverts to infinity when $r = 1$ and $n \geq 3$, while Jeffreys's intuitive choice $\text{BF}_{10;\kappa=1}$ does not do so until $n \geq 4$. Hence, Jeffreys intuitive choice $\kappa = 1$ misses the information consistency criterion by one observation. Furthermore, note the role of κ ; the smaller it is, the stronger the associated Bayes factor violatew the criterion of information consistency.

n	$\text{BF}_{10;\kappa=5}$	$\text{BF}_{10;\kappa=2}$	$\text{BF}_{10;\kappa=1}$	$\text{BF}_{10;\kappa=1/3}$	$\text{BF}_{10;\kappa=1/10}$
1	1	1	1	1	1
2	1	1	1	1	1
3	∞	∞	2	1.2	1.05
4	∞	∞	∞	1.75	1.17
5	∞	∞	∞	3.20	1.36

2.5.6 Example: The Bayesian correlation test

We now apply Jeffreys's default Bayesian correlation test to a data set analysed earlier by Stulp et al. (2013).

Example 2.5.1 (Do taller electoral candidates attract more votes?). *Stulp et al. (2013) studied whether there exists a relation between the height of electoral candidates and their popularity among voters. Based on the data from $n = 46$ US presidential elections, Stulp et al. (2013) reported a positive linear correlation of $r = .39$ between X , the relative height of US presidents compared to their opponents, and Y , the proportion of the popular vote. A frequentist analysis yielded $p = .007$. Fig. 2.4 displays the data. Based in part on these results, Stulp et al. (2013, p. 159) concluded that “height is indeed an important factor in the US presidential elections”, and “The advantage of taller candidates is potentially explained by perceptions associated with height: taller presidents are rated by experts as ‘greater’, and having more leadership and communication skills. We conclude that height is an important characteristic in choosing and evaluating political leaders.”*

For the Stulp et al. (2013) election data Jeffreys's exact correlation Bayes factor Eq. (2.5.12) yields $\text{BF}_{10;\kappa=1} = 6.33$, indicating that the observed data are

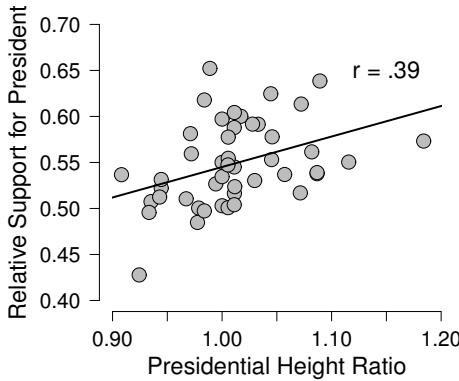


Figure 2.4: The data from $n = 46$ US presidential elections, showing the proportion of the popular vote for the president versus his relative height advantage against the closest competitor. The sample correlation equals $r = .39$, and, assuming an unrealistic sampling plan, the p -value equals .007. Jeffreys's default two-sided Bayes factor equals $\text{BF}_{10}(n = 46, r = .39) = 6.33$, and the corresponding one-sided Bayes factor equals $\text{BF}_{+0}(n = 46, r = .39) = 11.87$. See text for details.

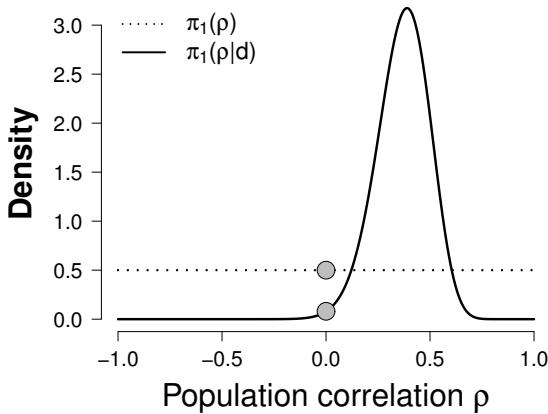


Figure 2.5: Posterior and prior distributions of the population correlation coefficient ρ for a two-sided default Bayes factor analysis of the height-popularity relation in US presidents Stulp et al. (2013). The Jeffreys default Bayes factor of $\text{BF}_{10; \kappa=1} = 6.33$ equals the height ratio of the prior distribution $\pi_1(\rho)$ over the posterior distribution $\pi_1(\rho | d)$ at $\rho = 0$.

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

6.33 times more likely under \mathcal{M}_1 than under \mathcal{M}_0 . This result is visualised in Fig. 2.5 using the Savage-Dickey density ratio test. With equal prior odds, the posterior probability for \mathcal{M}_0 remains an arguably non-negligible 14%. \diamond

2.5.7 The one-sided extension of Jeffreys's exact correlation Bayes factor

Whereas the function A fully determines the two-sided Bayes factor $\text{BF}_{10; \kappa}(n, r)$, the function B takes on a prominent role when we compare the null hypothesis \mathcal{M}_0 against the one-sided alternative \mathcal{M}_+ with $\rho > 0$.

To extend Jeffreys's exact correlation Bayes factor to a one-sided version, we retain the prior on the common parameters θ_0 . For the test-relevant prior $\pi_+(\rho; \kappa)$ we restrict ρ to non-negative values, which due to symmetry of $\pi_1(\rho; \kappa)$ is specified as

$$\pi_+(\rho; \kappa) = \begin{cases} 2\pi_1(\rho; \kappa) & \text{for } 0 \leq \rho \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (2.5.13)$$

Recall that A is an even function of ρ ; combined with the doubling of the prior for ρ this leads to a one-sided Bayes factor that can be decomposed as

$$\text{BF}_{+0; \kappa}(n, r) = \underbrace{\text{BF}_{10; \kappa}(n, r)}_{\int_0^1 A(n, r | \rho) \pi_+(\rho; \kappa) d\rho} + \underbrace{C_{+0; \kappa}(n, r)}_{\int_0^1 B(n, r | \rho) \pi_+(\rho; \kappa) d\rho}. \quad (2.5.14)$$

The function $C_{+0; \kappa}(n, r)$ can be written as

$$C_{+0; \kappa}(n, r) = \frac{2^{\frac{3\kappa-2}{\kappa}} r \kappa}{\mathcal{B}\left(\frac{1}{\kappa}, \frac{1}{\kappa}\right) \left((n-1)\kappa + 2\right)} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right]^2 {}_3F_2\left(1, \frac{n}{2}, \frac{n}{2}; \frac{3}{2}, \frac{2+\kappa(n+1)}{\kappa}; r^2\right), \quad (2.5.15)$$

where ${}_3F_2$ is a generalised hypergeometric function (Gradshteyn and Ryzhik, 2007, their Section 9.14) with three upper and two lower parameters.

The function $C_{+0; \kappa}(n, r)$ is positive whenever r is positive, since B as a function of ρ is then positive on the interval $(0, 1)$; consequently, for positive values of r the restricted, one-sided alternative hypothesis \mathcal{M}_+ is supported more than the unrestricted, two-sided hypothesis \mathcal{M}_1 , that is, $\text{BF}_{+0; \kappa}(n, r) > \text{BF}_{10; \kappa}(n, r)$. On the other hand, $C_{+0; \kappa}(n, r)$ is negative whenever r is negative; for such cases, $\text{BF}_{+0; \kappa}(n, r) < \text{BF}_{10; \kappa}(n, r)$.

Example 2.5.2 (One-sided correlation test for the US president data). As shown in Fig. 2.6, for the Stulp et al. (2013) data the one-sided Jeffreys's exact correlation Bayes factor Eq. (2.5.14) yields $\text{BF}_{+0; \kappa=1} = 11.87$, indicating that the observed data are 11.87 times more likely under \mathcal{M}_+ than under \mathcal{M}_0 . Because almost all posterior mass obeys the order-restriction, $\text{BF}_{+0} \approx 2 \times \text{BF}_{10}$ – its theoretical maximum. \diamond

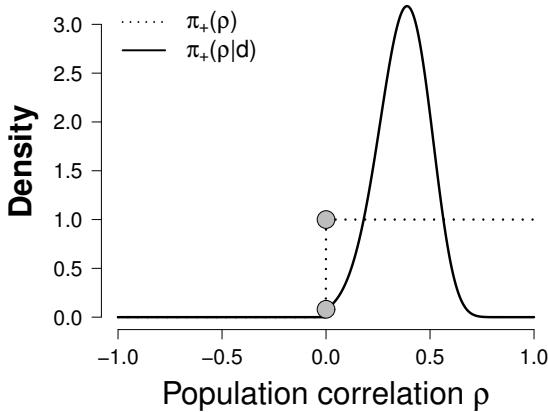


Figure 2.6: Posterior and prior distributions of the population correlation coefficient ρ for a one-sided default Bayes factor analysis of the height-popularity relation in US presidents Stulp et al. (2013). The Jeffreys default Bayes factor of $\text{BF}_{+0; \kappa=1} = 11.87$ equals the height ratio of the prior $\pi_+(\rho)$ over the posterior $\pi_+(\rho | d)$ at $\rho = 0$. The prior $\pi_+(\rho)$ is zero for negative values of ρ . Furthermore, note that the prior distribution $\pi_+(\rho)$ is twice as high for $\rho \geq 0$ compared to $\pi_-(\rho)$ in Fig. 2.5.

Using the same arguments as above, we can define the Bayes factor for a test between \mathcal{M}_- and \mathcal{M}_0 , which is in fact given by $\text{BF}_{-0; \kappa}(n, r) = \text{BF}_{+0; \kappa}(n, -r)$ due to the fact that B is an odd function of ρ . In effect, this implies that $\text{BF}_{+0; \kappa}(n, r) + \text{BF}_{-0; \kappa}(n, r) = 2 \times \text{BF}_{10; \kappa}(n, r)$, where the factor of two follows from symmetry of $\pi_-(\rho; \kappa)$ in the definition of $\pi_+(\rho; \kappa)$. Additional information on the coherence of the Bayes factor for order restrictions can be found in Mulder (2014) and Mulder (2016).

2.5.8 Discussion on the correlation test

As mentioned earlier, the previous analysis cannot be found in Jeffreys (1961) as Jeffreys did not derive the functions A and B explicitly. In particular, Jeffreys (1961, Eqns. (8, 9), p. 291) suggested that the integral of the likelihood Eq. (2.5.3) with respect to the parameterisation-invariant parameters $\pi_0(\theta_0)$ yields

$$h^J(n, r | \rho) = \frac{(1 - \rho^2)^{\frac{n-1}{2}}}{(1 - r\rho)^{\frac{2n-3}{2}},} \quad (2.5.16)$$

which in fact approximates the true test-relevant likelihood function $h = A + B$ very well for modest values of $|r|$ (cf. Jeffreys, 1961, p. 175) — this is illustrated in Fig. 2.7 which plots the error $h - h^J$. Specifically, the left panel of Fig. 2.7 shows that when $r = .39$, as in the example on the height of US presidents, there

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

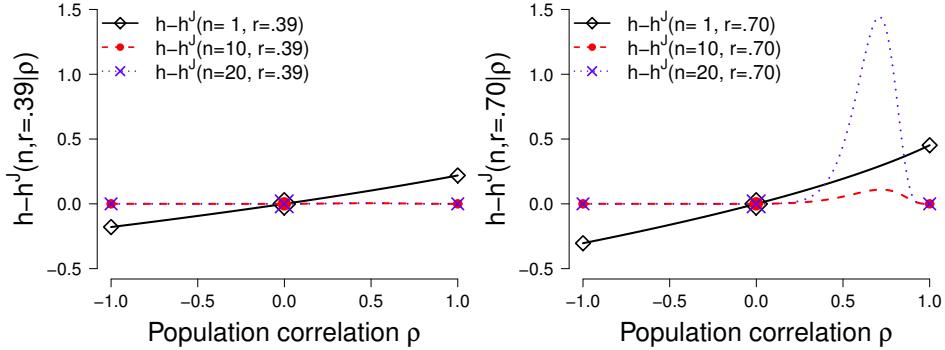


Figure 2.7: Error of approximation between the exact function h and Jeffreys's approximation h^J . The left panel shows that for a modest sample correlation (i.e., $r = .39$, as in the example on the height of US presidents) Jeffreys's approximation is quite accurate; moreover, the error decreases as n grows, and the curve of $n = 10$ overlaps with that of $n = 20$. However, the right panel shows that for a sample correlation of $r = .70$ the error increases with n , but only for some values of ρ . Furthermore, note that Jeffreys's approximation h^J does not yield $h^J(n = 1, r) = 1$ for every possible r .

is virtually no error when $n = 10$. The right panel of Fig. 2.7, however, shows that when $r = .70$, the error increases with n , but only for values of ρ from about .30 to about .95. From Jeffreys's approximation h^J one can define Jeffreys's integrated Bayes factor (Boekel et al., 2015; Keuken et al., 2017; Wagenmakers et al., 2016c):

$$\begin{aligned} \text{BF}_{10}^{\text{JI}}(n, r) &= \frac{1}{2} \int_{-1}^1 h_J(n, r | \rho) d\rho \\ &= \frac{\sqrt{\pi}}{2} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)} {}_2F_1\left(\frac{2n-3}{4}, \frac{2n-1}{4}; \frac{n+2}{2}; r^2\right). \end{aligned} \quad (2.5.17)$$

Jeffreys (1961, p. 175) noticed the resulting hypergeometric function, but as these functions were hard to compute, Jeffreys went on to derive a practical approximation for the users of his Bayes factor. The final Bayes factor that Jeffreys recommended for the comparison \mathcal{M}_1 versus \mathcal{M}_0 is therefore an approximation of an approximation and given as

$$\text{BF}_{10}^{\text{J}}(n, r) = \sqrt{\frac{\pi}{2n-3}} (1-r^2)^{\frac{4-n}{2}}. \quad (2.5.18)$$

For the US presidents data from Example 2.5.1 all three Bayes factors yield virtually the same evidence (i.e., $\text{BF}_{10, \kappa=1}(n = 46, r = .39) = 6.331$, $\text{BF}_{10}^{\text{JI}}(n = 46, r = .39) = 6.329$, and $\text{BF}_{10}^{\text{J}}(n = 46, r = .39) = 6.379$). Table 2.2 shows that the three Bayes factors generally produce similar outcomes, even for large values of r (cf. Robert et al., 2009). Jeffreys's approximate Bayes factor turns out to be

remarkably accurate, especially because there is rarely the need to determine the Bayes factor exactly. Jeffreys (1961, p. 432) remarks:

In most of our problems we have asymptotic approximations to K [i.e., BF_{01}] when the number of observations is large. We do not need K with much accuracy. Its importance is that if $K > 1$ the null hypothesis is supported by the evidence; if K is much less than 1 the null hypothesis may be rejected. But K is not a physical magnitude. Its function is to grade the decisiveness of the evidence. It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are 10^4 or 10^{10} to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice.

Table 2.2: A comparison of Jeffreys's exact Bayes factor (i.e., $\text{BF}_{10;\kappa=1}$) to Jeffreys's approximate integrated Bayes factor (i.e., $\text{BF}_{10}^{J,I}$) and to Jeffreys approximation of the approximate integrated Bayes factor (i.e., BF_{10}^J) reveals the high accuracy of the approximations, even for large values of r .

n	$\text{BF}_{10;\kappa=1}(n, .7)$	$\text{BF}_{10}^{J,I}(n, .7)$	$\text{BF}_{10}^J(n, .7)$
5	1.1	1.1	0.9
10	3.6	3.6	3.2
20	67.5	67.2	63.7

n	$\text{BF}_{10;\kappa=1}(n, .9)$	$\text{BF}_{10}^{J,I}(n, .9)$	$\text{BF}_{10}^J(n, .9)$
5	2.8	2.8	1.5
10	84.6	83.7	62.7
20	197,753.0	196,698.0	171,571.5

Hence, the main advantage of having obtained the exact Bayes factor based on the true test-relevant likelihood function h may be that it justifies Jeffreys's approximation $\text{BF}_{10}^J(n, r)$. The true function h also provides insight in the one-sided version of Jeffreys's test, and it provides a clearer narrative regarding Jeffreys's motivation in model selection and hypothesis testing in general. Moreover, it allows us to show that Jeffreys's exact Bayes factor is model selection consistent.

2.5.8.1 Model selection consistency

To show that Jeffreys's correlation Bayes factor is model selection consistent, we use the sampling distribution of the maximum likelihood estimate (MLE). As r is the MLE we know that it is asymptotically normal with mean ρ and variance $\frac{1}{n(1-\rho^2)^2}$, where ρ is the true value. In particular, when the data are generated under \mathcal{M}_0 , thus, $\rho = 0$, we know that $r \sim \mathcal{N}(0, \frac{1}{n})$ when n is large. In order to show that the support for a true \mathcal{M}_0 grows without bound as the number of

2. HAROLD JEFFREYS'S DEFAULT BAYES FACTOR HYPOTHESIS TESTS: EXPLANATION, EXTENSION, AND APPLICATION IN PSYCHOLOGY

data points n increases, the Bayes factor $\text{BF}_{10;\kappa}(n, r)$ needs to approach zero as n increases.

We exploit the smoothness of $\text{BF}_{10;\kappa}(n, r)$ by Taylor expanding it up to third order in r . By noting that the leading term of the Taylor expansion $\text{BF}_{10;\kappa}(n, 0)$ has a factor $\Gamma\left(\frac{(n-1)\kappa+2}{2\kappa}\right)/\Gamma\left(\frac{n\kappa+2}{2\kappa}\right)$ we conclude that it converges to zero as n grows. The proof that the Bayes factor $\text{BF}_{10;\kappa}$ is also model selection consistent under \mathcal{M}_1 follows a similar approach by a Taylor approximation of second order and consequently concluding that $\text{BF}_{10;\kappa}(n, r)$ diverges to ∞ as n grows indefinitely.

2.6 Conclusion

We hope to have demonstrated that the Bayes factors proposed by Harold Jeffreys have a solid theoretical basis, and, moreover, that they can be used in empirical practice to answer one particularly pressing question: What is the degree to which the data support either the null hypothesis \mathcal{M}_0 or the alternative hypothesis \mathcal{M}_1 ? As stated by Jeffreys (1961, p. 302):

“In induction there is no harm in being occasionally wrong; it is inevitable that we shall be. But there is harm in stating results in such a form that they do not represent the evidence available at the time when they are stated, or make it impossible for future workers to make the best use of that evidence.”

It is not clear to us what inferential procedures other than the Bayes factor are able to represent evidence for \mathcal{M}_0 versus \mathcal{M}_1 . After all, the Bayes factor follows directly from probability theory, and this ensures that it obeys fundamental principles of coherence and common sense (e.g., Wagenmakers et al., 2014).

It needs to be acknowledged that the Bayes factor has been subjected to numerous critiques. Here we discuss two. First, one may object that the test-relevant prior distribution for the parameter of interest has an overly large influence on the Bayes factor (Liu and Aitkin, 2008). In particular, uninformative, overly wide priors result in an undue preference for \mathcal{M}_0 , a fact that Jeffreys recognised at an early stage. The most principled response to this critique is that the selection of appropriate priors is an inherent part of model specification. Indeed, the prior offers an opportunity for the implementation of a substantively different model (Vanpaemel, 2010).

In this chapter, we showcased this ability when we adjusted the prior to implement a directional, one-sided alternative hypothesis. In general, the fact that different priors result in different Bayes factors should not come as a surprise. As stated by Jeffreys (1961, p. x):

“The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.”

The second critique is that in practice, all models are wrong. At first glance this appears not to be a problem, as the Bayes factor quantifies the support for \mathcal{M}_0 versus \mathcal{M}_1 , regardless of whether these models are correct. However, it is important to realise that the Bayes factor is a relative measure of support. The fact that $BF_{10} = 100,000$ indicates that \mathcal{M}_1 receives much more support from the data than does \mathcal{M}_0 , but this does not mean that \mathcal{M}_1 is any good in an absolute sense (e.g., Andraszewicz et al., 2015; Anscombe, 1973). In addition, it has recently been suggested that when both models are misspecified, the Bayes factor may perform poorly in the sense that it is too slow to select the best model (van Erven et al., 2012). However, the Bayes factor does have a predictive interpretation that does not depend on one of the models being true (Wagenmakers et al., 2006); similarly, the model preferred by the Bayes factor will be closest (with respect to the Kullback-Leibler divergence) to the true data-generating model (Berger, 1985; Jeffreys, 1980). More work on this topic is desired and expected.

In mathematical psychology, the Bayes factor is a relatively popular method of model selection, as it automatically balances the tension between parsimony and goodness-of-fit, thereby safeguarding the researcher against overfitting the data and preferring models that are good at describing the obtained data, but poor at generalising and prediction (Myung et al., 2000c; Myung and Pitt, 1997; Wagenmakers and Waldorp, 2006b). Nevertheless, with the recent exception of the Bayes factor *t*-test, the Bayes factors proposed by Jeffreys (1961) have not received much attention, neither by statisticians nor mathematical psychologists. One of the reasons for this unfortunate fact is that Jeffreys notation is more accustomed to philosophers of logic (Geisser, 1980). In order to make Jeffreys's work somewhat more accessible, Appendix 2.D provides a table with a modern-day translation of Jeffreys's notation. In addition, any scholar new to the work of Jeffreys is recommended to first read the extended modern summary by Robert et al. (2009).

We would like to stress that a Jeffreys Bayes factor is not a mere ratio of likelihood functions averaged with respect to a subjective elicited prior $\pi_i(\theta_i)$ obtained from a within-model perspective. Jeffreys's development of the Bayes factor resembles an experimental design for which one studies where the likelihood functions overlap, how they differ, and in what way the difference can be apparent from the data. These considerations then yield priors from which a Bayes factor needs to be computed. The computations are typically hard to perform and might not yield closed form results. These computational issues were a major obstacle for the Bayesian community, however, Jeffreys understood that closed form solutions are not always necessary for good inference; moreover, he was able to derive approximate Bayes factors, allowing his exposition of Bayesian inductive reasoning to transcend from a philosophical debate into practical tools for scientific scrutiny.

Modern-day statisticians and mathematical psychologists may lack Jeffreys's talent to develop default Bayes factors, but we are fortunate enough to live in a time in which computer-driven sampling methods known as Markov chain Monte Carlo (MCMC: e.g., Gamerman and Lopes, 2006; Gilks et al., 1996) are widely available. This removes the computational obstacles one needs to resolve after the priors are specified. These tools make Jeffreys's method of testing more attainable than ever before.

2.A The default Bayes factor hypothesis tests proposed by Jeffreys in ToP

Table 2.3: Default Bayes factor hypothesis tests proposed by Jeffreys (1961) in Chapter V of “Theory of Probability” (third edition).

Tests	Pages
Binomial rate	256 – 257
Simple contingency	259 – 265
Consistency of two Poisson parameters	267 – 268
Whether the true value in the normal law is zero, σ unknown	268 – 274
Whether a true value is zero, σ known	274
Whether two true values are equal, standard errors known	278 – 280
Whether two location parameters are the same, standard errors not supposed equal	280 – 281
Whether a standard error has a suggested value σ_0	281 – 283
Agreement of two estimated standard errors	283 – 285
Both the standard error and the location parameter	285 – 289
Comparison of a correlation coefficient with a suggested value	289 – 292
Comparison of correlations	293 – 295
The intraclass correlation coefficient	295 – 300
The normal law of error	314 – 319
Independence in rare events	319 – 322

2.B The hypergeometric function

The hypergeometric function (Oberhettinger, 1972, Section 15) with two upper parameters and one lower parameter generalises the exponential function as follows (Gradshteyn and Ryzhik, 2007, p 9.114):

$${}_2F_1(a, b; c; z) = 1 + \frac{a \cdot b}{c \cdot 1} z + \frac{a(a+1)b(b+1)}{c(c+1) \cdot 1 \cdot 2} z^2 \quad (2.B.1)$$

$$+ \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2) \cdot 1 \cdot 2 \cdot 3} z^3 + \dots \quad (2.B.2)$$

2.C The stretched beta density

By the change of variable formula, we obtain the stretched beta density of ρ on $(-1, 1)$ with parameters $\alpha, \beta > 0$

$$\frac{1}{2\mathcal{B}(\alpha, \beta)} \left(\frac{\rho+1}{2} \right)^{\alpha-1} \left(\frac{1-\rho}{2} \right)^{\beta-1}, \quad (2.C.1)$$

where $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function that generalises $\binom{n}{k}$ to real numbers. By setting $\beta = \alpha$ this yields the symmetric beta density of ρ on $(-1, 1)$ with parameters $\alpha > 0$

$$\frac{2^{-2\alpha+1}}{\mathcal{B}(\alpha, \alpha)}(1-\rho^2)^{\alpha-1}. \quad (2.C.2)$$

The reparametrisation we used in text is given by simply substituting $\alpha = 1/\kappa$ allowing us to interpret κ as a scale parameter.

2.D Translation of Jeffreys's notation in ToP

Table 2.4: Translation of the notation introduced by (Jeffreys, 1961, pp. 245–267). The treatment of α and β as new or old parameters differs from context to context in (Jeffreys, 1961).

Jeffreys's notation	Modern notation	Interpretation
q	\mathcal{M}_0	Null hypothesis or null model
q'	\mathcal{M}_1	Alternative hypothesis or alternative model
H		Background information (mnemonic: “history”)
$P(q H)$	$P(\mathcal{M}_0)$	Prior probability of the null model
$\int f(\alpha) d\alpha$	$\int \pi(\theta) d\theta$	Prior density on the parameter θ
$P(q' d\alpha H)$	$P(\mathcal{M}_1, \theta)$	Probability of the alternative model and its parameter
$P(q aH)$	$\pi_0(\theta_0 x)$	Posterior density on the parameter within \mathcal{M}_0
$P(q' d\alpha aH)$	$\pi_1(\theta_1 x)$	Posterior density on the parameter within \mathcal{M}_1
K	$\text{BF}_{01}(d)$	The Bayes factor in favour of the null over the alternative
α', β	$\theta_0 = \alpha, \theta_1 = (\frac{\alpha'}{\beta})$	“Alternative” parameter $\theta_1 = (\frac{\alpha'}{\beta})$ new parameter
$f(\beta, \alpha')$	$\pi_1(\zeta \theta_0)$	Prior of the new given the old prior within \mathcal{M}_1
$g_{\alpha\alpha} d\alpha'^2 + g_{\beta,\beta} d\beta^2$	$I(\vec{\theta})$	Fisher information matrix
$P(q, db H) = f(b) db$	$\pi_0(\theta_0)$	Prior density of the common parameters within \mathcal{M}_0
$P(q' db d\alpha H) = f(b) db d\alpha$	$\pi_1(\theta_1)$	Prior density of the parameters within \mathcal{M}_1
$P(\theta q, b, H)$	$f(d \theta_0, \mathcal{M}_0)$	The likelihood under \mathcal{M}_0
$P(\theta q', b, \alpha, H)$	$f(d \theta_0, \zeta, \mathcal{M}_1)$	Likelihood under \mathcal{M}_1
$P(q db \theta H)$	$\pi_0(\theta_0 d)$	Posterior of the parameters within \mathcal{M}_0
$P(q' db d\alpha \theta H)$	$\pi_1(\theta_1 d)$	Posterior of the parameters within \mathcal{M}_1

Chapter 3

An Evaluation of Alternative Methods for Testing Hypotheses, from the Perspective of Harold Jeffreys

Abstract

Our original article provided a relatively detailed summary of Harold Jeffreys's philosophy on statistical hypothesis testing. In response, Robert (2016) maintains that Bayes factors have a number of serious shortcomings. These shortcomings, Robert argues, may be addressed by an alternative approach that conceptualises model selection as parameter estimation in a mixture model. In a second comment, Chandramouli and Shiffrin (2016) seek to extend Jeffreys's framework by also taking into consideration probability mass functions that do not belong to the models under test. In this rejoinder we argue that Robert's (2016) alternative view on testing has more in common with Jeffreys's Bayes factor than he suggests, as they share the same "shortcomings". On the other hand, we show that the proposition of Chandramouli and Shiffrin (2016) to extend the Bayes factor is in fact further removed from Jeffreys's view on testing than the authors suggest. By elaborating on these points, we hope to clarify our case for Jeffreys's Bayes factors.

Keywords: Bayes factors, induction, model selection, replication, statistical evidence.

This chapter is published as: Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55. doi: <http://dx.doi.org/10.1016/j.jmp.2016.01.003>

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

3.1 Introduction

In our original article (Ly et al., 2016a) we outlined how Harold Jeffreys constructed his hypothesis tests. Jeffreys’s tests contrast a precise, point-null hypothesis M_0 versus a more general alternative hypothesis M_1 . Here the point-null hypothesis represents a general law, an invariance, or a categorical causal claim (e.g., “apple trees always bear apples”; “people cannot look into the future”; “Alzheimer’s disease is caused by a fungal infection of the central nervous system”), whereas the alternative hypothesis relaxes that law. Jeffreys’s tests require a thoughtful specification of the prior distribution for the parameter of interest, and much of Jeffreys’s work was concerned with providing good default specifications – “good” in the sense that they adhere to general common-sense desiderata (e.g., Bayarri et al., 2012). We are pleased that our summary attracted two comments by renowned researchers; below we respond to their ideas in a way that we hope is consistent with the overall philosophy of Harold Jeffreys himself.

3.2 Rejoinder to Robert

In general, Robert’s (2016) comments highlight the inevitable subtleties in constructing a Bayes factor. His alternative mixture model procedure is practical and may be immensely valuable for specific situations (i.e., hierarchical models) that are common in psychological research. Nevertheless, we believe Robert’s suggestion about the demise of the Bayes factor to be an overstatement.

3.2.1 Robert’s critique on the Bayes factor

Our understanding of Jeffreys’s method is partly based on the work by Robert and colleagues (2009), and it should, therefore, not come as a surprise that Robert’s view and ours overlap to a considerable degree. Robert’s arguments for dismissing the Bayes factor can be grouped in terms of (1) its usage in making decisions, and (2) the care that needs to be taken in choosing the priors.

3.2.1.1 First critique: The distinction between inference and decision making

We share Robert’s discontent with the statistical practice that emphasises all-or-none decisions at some arbitrary threshold, and we agree that scientific learning should instead be guided by a continuous measure of evidence. In the process of eviscerating p -value null hypothesis tests, Rozeboom (1960, pp. 422-423) already expressed a similar sentiment:

“The null-hypothesis significance test treats ‘acceptance’ or ‘rejection’ of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice

but determined solely by how likely it is, given the evidence, that the hypothesis is true.”

Our favourite continuous measure of evidence is of course a Bayes factor constructed from a pair of priors selected according to Jeffreys’s desiderata, which we simply refer to as a Jeffreys’s Bayes factor. It is important to note that this measure provides only the first of three Bayesian ingredients needed for decision making. The other two ingredients are the prior model probabilities (which, combined with the Bayes factor, yield posterior model probabilities) and the specification of a loss function (or equivalently, a utility function; Berger, 1985, Lindley, 1977, and Robert, 2007).

For instance, consider a Bayes factor of $\text{BF}_{10}(d) = 4.6$ for the observed data d . This Bayes factor can be converted into a posterior model probability of $P(\mathcal{M}_0 | d) = 0.17$ when we set $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$ (Ly et al., 2016a). One possible subsequent decision rule is then to accept $P(\mathcal{M}_1 | d)$ because it has the highest posterior model probability. We did not intend to suggest such a procedure, as the decision is clearly sensitive to the prior model probabilities. Furthermore, we do not recommend uniform prior model probabilities regardless of scientific context. In fact, when decision making is desired, the assignment of prior model probabilities is left to the substantive researcher. Such flexibility in assignment introduces subjectivity, and this may be seen either as a disadvantage or as an advantage. At any rate, prior model probabilities can be used to formalise the adage that “extraordinary claims require extraordinary evidence” (e.g., Wagenmakers et al., 2011). Moreover, the prior model probabilities can be used to address the problem of multiplicity (e.g., Jeffreys, 1961; Scott and Berger, 2010; Stephens and Balding, 2009). A similar argument applies to utility functions: these may be subjective and hard to elicit, but such difficulties do not sanction the practice of ignoring utility functions altogether, at least not when the purpose is to make decisions.

Thus, Robert worries that computation of Bayes factors may tempt users to make all-or-none decisions while disregarding prior model probabilities or loss functions. We agree with Robert that there is a considerable difference between inference and decision making, and that scientific learning should be guided by a continuous measure of evidence that incorporates what we have learned from the observed data. The Bayes factor is such a measure.

3.2.1.2 Second critique: The Jeffreys-Lindley-Bartlett paradox

We suspect that the Jeffreys-Lindley-Bartlett (henceforth JLB) paradox is central to Robert’s (1993; 2014) dismissal of the Bayes factor and it is the main motivation for the development of the mixture model alternative. We take a closer look at the JLB paradox and discuss two consequences foreseen by Jeffreys, who was keenly aware of the “paradox” from the very beginning (Etz and Wagenmakers, 2017).

First, the JLB paradox implies that we cannot use improper priors to construct a Bayes factor. For instance, to estimate μ within the normal model $\mathcal{M}_1 : X \sim \mathcal{N}(\mu, 1)$, we typically employ Jeffreys’s (1946) prior $\mu \propto 1$. The reason to do so stems from the fact that Jeffreys’s prior is parameterisation-invariant, leading to

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

a posterior that is independent on how researchers parameterise the problem (Ly et al., 2017c). The JLB paradox implies that we cannot use this same (estimation) prior on the test-relevant parameter for a Bayesian test. More specifically, when we pit the aforementioned model \mathcal{M}_1 against the null model $\mathcal{M}_0 : X \sim \mathcal{N}(0, 1)$ the improper prior $\pi_1(\mu) \propto 1$ then becomes useless. To see this we consider the Jeffreys's prior as the limit of proper priors $\mu \sim \mathcal{N}(0, g)$ with g tending to infinity. The Bayes factor for the observed data $d = (n, \bar{x})$ is then given by

$$\lim_{g \rightarrow \infty} \tilde{\text{BF}}_{10;g}(d) = \lim_{g \rightarrow \infty} \frac{\int \exp[-\frac{n}{2}(\bar{x} - \mu)^2] \exp[-\frac{1}{2g}\mu^2] d\mu}{\sqrt{2\pi g} \exp[-\frac{n}{2}\bar{x}^2]}, \quad (3.2.1)$$

$$= \lim_{g \rightarrow \infty} \frac{1}{\sqrt{1+ng}} \exp\left[\frac{g(n\bar{x})^2}{2(1+ng)}\right] = 0, \quad (3.2.2)$$

regardless of the fixed sample size n and the observed sample mean \bar{x} . As such, the Bayes factor constructed from the improper Jeffreys's prior will always favour the null model and this also holds for other improper priors. Moreover, Eq. (3.2.2) shows that for fixed data $d = (n, \bar{x})$ and a Bayes factor constructed from a normal prior with hyperparameter g we can obtain a Bayes factor in favour of the null hypothesis of arbitrary size (i.e., $\tilde{\text{BF}}_{10;g}(d) < 1$) simply by taking g large enough.

Hence, the JLB paradox effectively implies that a testing problem should be treated differently from one that is concerned with estimation. As such, when π_1 is interpreted as prior belief about the parameters θ_1 , in the example above $\theta_1 = \mu$, one's belief about the parameter then changes depending on whether one is concerned with testing or estimation. More generally, this difference is due to the fact that estimation is typically a *within*-model affair. Recall that a model \mathcal{M}_i specifies a relationship $f_i(d | \theta_i)$ that defines which parameters θ_i are relevant in the data generating process of the data d . Hence, the function f_i gives the (only) context in which the parameters θ_i can be perceived.

In essence, the function f_i justifies that it is meaningful to calculate a posterior distribution for the parameter. To underline this point we add subscripts to the parameters indicating model membership in the next example, by taking $\theta_0 = \sigma_0$ and $\theta_1 = (\mu_1, \sigma_1)$ for f_0 and f_1 both normals. For example, when we assume that $\mathcal{M}_0 : X \sim \mathcal{N}(0, \sigma_0^2)$ only a posterior for the standard deviation σ_0 is worthwhile to be pursued, as the posterior for the population mean remains zero, regardless of the data. Within \mathcal{M}_0 , the Jeffreys's prior for σ_0 is given by $\pi_0(\sigma_0) \propto \sigma_0^{-1}$, which can be updated to a posterior $\pi_0(\sigma_0 | d)$. On the other hand, under $\mathcal{M}_1 : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ we are dealing with two parameters of interest. Within \mathcal{M}_1 , the Jeffreys's prior for μ_1 is $\pi(\mu_1) \propto 1$, for σ_1 is $\pi_1(\sigma_1) \propto 1/\sigma_1$ and we take $\pi_1(\mu_1, \sigma_1) = \pi_1(\mu_1)\pi_1(\sigma_1)$. These priors can be updated to posteriors $\pi_1(\mu_1 | d)$ and $\pi_1(\sigma_1 | d)$. Even though the two priors $\pi_0(\sigma_0)$ and $\pi_1(\sigma_1)$ have the same form, they do not lead to the same posterior. In fact, due to the presence of μ_1 as a parameter, the posterior mean of $\pi_1(\sigma_1 | d)$ within \mathcal{M}_1 will be smaller or equal to the posterior mean of $\pi_0(\sigma_0 | d)$ within \mathcal{M}_0 . Thus, when we are interested in the standard error σ_i , it matters whether we believe that \mathcal{M}_0 holds true or whether the population mean μ_1 plays a role in the data generating process as specified by f_1 . The Bayes factor helps us distinguish which of the two models is better suited to the data and which posterior for σ_i we should report. Hence, testing is a *between*-model

matter. Jeffreys himself was very clear about the distinction between estimation and testing:

“We are now concerned with the more difficult question: in what circumstances do observations support a change of the form of the law itself? This question is really logically prior to the estimation of the parameters, since the estimation problem presupposes that the parameters are relevant.” (Jeffreys, 1961, p. 245)

Hence, testing implies that we are uncertain about which of the two functional relationships defined by the models \mathcal{M}_0 and \mathcal{M}_1 is adequate for the data under study. This uncertainty is expressed through the prior statement $P(\mathcal{M}_0), P(\mathcal{M}_1) > 0$ and when \mathcal{M}_0 and \mathcal{M}_1 are the only models under consideration we require that $P(\mathcal{M}_0) + P(\mathcal{M}_1) = 1$. The priors π_1, π_0 in a Bayes factor are, thus, chosen to guide scientific learning, that is, how one transitions from prior model odds to posterior model odds and are not designed to yield posteriors that are necessarily good for estimation. To simplify notation, we drop the subscripts indicating model membership when the context is clear.

Second, the separation of estimation and testing and the resulting separation of models led us to instantiate the hypotheses \mathcal{H}_i with their respective models \mathcal{M}_i as discussed in Ly et al. (2016a). In effect, we have different contexts in which the respective parameters exist and, therefore, a philosophical conundrum in what is meant by common parameters. The difference between the posteriors $\pi_0(\sigma | d)$ and $\pi_1(\sigma | d)$ discussed above showed that one should not be fooled by the fact that the Greek letters are identical. We therefore agree with Robert’s warning concerning the treatment of common parameters.

For the t -test the commonality between the two σ s within \mathcal{M}_0 and \mathcal{M}_1 is given by their meaning as a scaling parameter within either model. Furthermore, the nesting of $\pi_0(\sigma)$ as $\pi_1(\mu, \sigma) = \pi_1(\delta)\pi_0(\sigma)$ can be considered a practical choice. In effect, the Bayes factor $\text{BF}_{10}(d)$ is then given by the ratio of the following two marginal likelihoods

$$\begin{aligned} p(d | \mathcal{M}_1) &= (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \int_{-\infty}^\infty \exp\left(-\frac{n}{2} \left[(\frac{\bar{x}}{\sigma} - \delta)^2 + (\frac{s}{\sigma})^2 \right]\right) \pi_1(\delta) d\delta \pi_0(\sigma) d\sigma, \\ p(d | \mathcal{M}_0) &= (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n} \exp\left(-\frac{n}{2\sigma^2} [\bar{x}^2 + s^2]\right) \pi_0(\sigma) d\sigma, \end{aligned} \quad (3.2.3)$$

where $d = (n, \bar{x}, s^2)$.

With the nesting of π_0 within π_1 we made the following recommendation explicit: “It is to be understood that in pairs of equations of this type [such as Eq. (3.2.3)] the sign of proportionality indicates the same constant factor, which can be adjusted to make the total probability 1.” (Jeffreys, 1961, p. 247) More precisely, an improper prior $\pi_0(\sigma) \propto \sigma^{-1}$ has a suppressed normalisation constant $\pi_0(\sigma) = c_0\sigma^{-1}$ and we not only take $\pi_1(\sigma) = c_1\sigma^{-1}$ of the same form, but also choose to set $c_1 = c_0$, which allows us to use improper priors on the nuisance parameters (see Berger et al., 1998 for a theoretical justification). More examples of this type of nesting can be found in Dawid and Lauritzen (2001), Consonni and Veronese (2008), and references therein.

3.2.2 Jeffreys's common-sense desiderata

"Rejection of a null hypothesis is best when it is interocular". Edwards et al. (1963, p. 240)

In conclusion, the JLB paradox prohibits the usage of improper priors for testing, separates the estimation practice from a testing concern, and challenges the idea of common parameters. As noted above, we first require a justification before we can use the same prior on the nuisance parameters. After doing so, we then create an exception on the ban of improper priors allowing us to assign improper priors to the nuisance parameters, say, $\theta_0 = \sigma$. Furthermore, let δ denote the test-relevant parameter with, say, $\theta_1 = (\theta_0, \delta)$. Hence, after specifying Jeffreys's parameterisation-invariant priors on the nuisance parameters θ_0 , which we would use for estimation within each model, we only require to set the prior $\pi_1(\delta)$ in order to define the Bayes factor $\text{BF}_{10}(d)$. We suspect that Jeffreys's underlying reasons for the choice of $\pi_1(\delta)$ was to have a test that passes "the interocular traumatic test; you know what the data mean when the conclusion hits you between the eyes." Edwards et al. (1963, p. 217).

We believe that the information consistency criterion makes explicit which data hit us right between the eyes. This criterion leads to a Bayes factor that is consistent for a finite sample, a requirement that is much harder to be fulfilled than the asymptotic consistency criterion, at least for parametric models (e.g., Bickel and Kleijn, 2012, Yang and Le Cam, 2000). We agree with Robert that information consistency is in some cases an approximate statement. In particular, when the data are either distributed according to $\mathcal{M}_0 : X \sim \mathcal{N}(0, \sigma^2)$ or $\mathcal{M}_1 : X \sim \mathcal{N}(\mu, \sigma^2)$ then the interocular data set with $n > 2$, $\bar{x} \neq 0$ and, in particular, $s^2 = 0$ occurs with zero probability under both models, due to the assumption $\sigma > 0$. However, when \mathcal{M}_0 and \mathcal{M}_1 are the only two models under consideration, the observation $\bar{x} \neq 0$ with $n > 2$, in addition to $s^2 = 0$, then should lead to the logical exclusion of \mathcal{M}_0 , thus, $\text{BF}_{01}(d) = 0$.

To appreciate the information consistency criterion, we revisit the Bayesian t -test with Bayes factors $\tilde{\text{BF}}_{10;g}(d)$ that lacks this property by constructing it from $\pi_0(\sigma) \propto \sigma^{-1}$ and $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$ where $\pi_1(\delta)$ is normal around zero with variance g , i.e.,

$$\tilde{\text{BF}}_{10;g}(d) = (1 + ng)^{\frac{n-1}{2}} \left(\frac{1 + \frac{n\bar{x}^2}{ns^2}}{(1 + ng) + \frac{n\bar{x}^2}{ns^2}} \right)^{\frac{n}{2}}. \quad (3.2.4)$$

As before, letting g tend to infinity, while keeping n, \bar{x} and s^2 fixed, yields the JLB paradox, i.e., $\lim_{g \rightarrow \infty} \tilde{\text{BF}}_{10;g}(d) = 0$.

To simplify the discussion we suppose that g is set to one. The resulting Bayes factor $\tilde{\text{BF}}_{10;g=1}(d)$ is then asymptotically consistent. This means that if we repeatedly sample from the null model, we let n tend to infinity and simultaneously let $n\bar{x}^2/(ns^2) = t^2/(n-1)$ tend to zero yielding a Bayes factor of zero, where t is the usual t -statistic $t = \sqrt{n}\bar{x}/s_{n-1}$. Similarly, if we repeatedly sample from the alternative model, we let n tend to infinity and simultaneously let $t^2/(n-1)$ tend to infinity yielding a Bayes factor of infinity. Thus, this Bayes factor $\text{BF}_{10;g=1}(d)$ is able to detect the correct model when the number of data points tends to infinity.

The Bayes factor $\tilde{BF}_{10;g=1}(d)$, however, is not information consistent. For the *t*-test information consistency is concerned with having a fixed number of data points $n > 2$, an observed sample mean, say, $\bar{x} \neq 0$ and s^2 tending to zero. With g , n and \bar{x} fixed, this Bayes factor $\tilde{BF}_{10;g=1}(d)$ is a decreasing function of s^2 that attains its maximum when $s^2 = 0$. For instance, when $n = 4$, $\bar{x} = 7$ the maximum is then given by $\lim_{s^2 \rightarrow 0} \tilde{BF}_{10;g=1}(d) = 11.18$. Note that the data set with $n = 4$, $\bar{x} = 7$ and $s^2 \rightarrow 0$ is interocular as it leads to an observed sample effect size, an realisation of the *t*-statistic, that tends to infinity, which should therefore lead to infinite support for the alternative compared to the null model. The fact that the information inconsistent Bayes factor $\tilde{BF}_{10;g=1}(d)$ is bounded makes it hard to be interpret. For instance, the observations $n = 4$, $\bar{x} = 7$ and $s^2 = 1$ yields a Bayes factor of $\tilde{BF}_{10;g=1}(d) = 9.6$, which does not seem a lot of evidence against the null, but with respect to its maximum 11.81 might be considered substantial.

On the other hand, a Jeffreys's Bayes factor is by construction information consistent and has a supremum (i.e., maximum) at infinity, which makes it easier to be interpret. Jeffreys referred to this and other desiderata as common-sense as they came natural to him (Etz and Wagenmakers, 2017), but it took a long time before his intuition was formalised by Berger and Pericchi (2001) and extended by Bayarri et al. (2012).

Recall that information consistency in a *t*-test requires us to construct a Bayes factor from a heavy-tailed prior on δ and we agree with Robert that the Cauchy prior with scale $\gamma = 1$ is only one of many possible choices. This is why we included a robustness analysis in our open-source software package JASP (<https://jasp-stats.org/>). However, we believe that the merit of a Jeffreys's Bayes factor (with γ fixed) is due to the fact that it kickstarts scientific learning.

“In any of these cases it would be perfectly possible to give a form of $[\pi_1(\delta)]$ that would express the previous information satisfactorily, and consideration of the general argument of [Chapter] 5.0 will show that it would lead to common-sense results, but they would differ in scale.

As we are aiming chiefly at a theory that can be used in the early stages of a subject, we shall not at present consider the last type of case” (Jeffreys, 1961, p. 252).

Thus, Jeffreys was not opposed to incorporating previously acquired data in a Bayesian hypothesis test, but to do so he first designed a starting Bayes factor, for a first data set, say, d_{orig} . After observing d_{orig} , we can then straightforwardly update a Jeffreys's Bayes factor for a future, not yet observed, data set, say, d_{rep} . This informed Bayes factor $\text{BF}_{10}(d_{\text{rep}} | d_{\text{orig}})$ is then constructed from the priors $\pi_1(\theta_1 | d_{\text{orig}})$ and $\pi_0(\theta_0 | d_{\text{orig}})$. This idea forms the basis of the replication Bayes factors introduced in Verhagen and Wagenmakers (2014) and is further exploited in Ly et al. (2017b). Hence, the man who discovered the origin of the earth, thus, also provided us with the starting point for scientific learning.

3.2.3 Robert’s alternative approach

“Prior distributions must always be chosen with the utmost care when dealing with mixtures and their bearings on the resulting inference

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

assessed by a sensitivity study. The fact that some noninformative priors are associated with undefined posteriors, no matter what the sample size, is a clear indicator of the complex nature of Bayesian inference for those models” (Marin and Robert, 2014, p. 199)

As an alternative to Bayes factors, Robert (2016) suggests to use a mixture model approach elaborated upon in Kamary et al. (2014). The data generating process of a mixture model can be envisioned as a stepwise procedure. First, a membership variable z_j is realised; in a two-component mixture, z_j assumes either the value zero or one. Next, given the outcome $z_j = 0$ (or $z_j = 1$) a data point x_j is generated according to $\mathcal{M}_0 : X_j \sim f_0(x_j | \theta_0)$ (or $\mathcal{M}_1 : X_j \sim f_1(x_j | \theta_1)$). This means that the complete data should consist of n -pairs $(z_1, x_1), \dots, (z_n, x_n)$, but in reality we only have the observations $d = x_1, \dots, x_n$. As a result of not observing the membership variables z_j , the observations are perceived as if each of the data points were generated from the (arithmetic) mixture model $\mathcal{M}_a : X_j \sim (1 - \alpha)f_0(x_j | \theta_0) + \alpha f_1(x_j | \theta_1)$, where α is the mixture proportion. The artificial encompassing model \mathcal{M}_a therefore contains the two competing models, \mathcal{M}_0 and \mathcal{M}_1 , as special cases; when $\alpha = 0$ and $\alpha = 1$ respectively. Hence, to uncover whether the observations are more consistent with \mathcal{M}_0 or \mathcal{M}_1 , Kamary et al. (2014) suggest to focus on estimating α within the encompassing model \mathcal{M}_a .

Inferring α amounts to a missing data problem which is in principle computationally intensive as there are 2^n different combinations for the membership variables z_j s. Luckily, one can resort to a completion method pioneered by Diebolt and Robert (1994). When this stochastic exploration method yields n_0 and n_1 numbers of observations allocated to \mathcal{M}_0 and \mathcal{M}_1 , respectively, the posterior for α is then given by $\mathcal{B}(a + n_0, a + n_1)$, when we use a beta prior on the mixture proportion, $\alpha \sim \mathcal{B}(a, a)$. When n_0 is large and n_1 small or zero, the posterior for α then concentrates most of its mass near zero indicating more evidence for \mathcal{M}_0 as one would expect.

Kamary et al. (2014) note that the data generative view of the mixture model is theoretically justified, but that the resulting natural Gibbs sampler has convergence problems when the hyperprior a is smaller than one. To circumvent this problem, Kamary et al. (2014) propose to use a Metropolis-Hastings algorithm instead, and illustrate its use by examples followed by a proof that shows that the method is asymptotically consistent. Thus, the work by Kamary et al. (2014) impressively introduces an alternative view on testing, an algorithmic resolution, and a theoretical justification.

3.2.3.1 Testing versus estimation

We believe that the Kamary et al. (2014) mixture approach will be especially useful in psychological research. In particular, consider a hierarchical model where each participant’s performance x_j on a psychological task is captured by a particular model or strategy represented by f_i . The posterior for α then gives an indication of the prevalence of the model or strategy. When the posterior for α is near zero or near one, this suggests that one model or strategy is dominant; when the posterior for α is near $1/2$, this suggests that some participants are better described by one

strategy, and some are better described by another (for similar approaches see Friston and Penny, 2011; Lee et al., 2015).

The advantage of the mixture model approach is particularly acute when it is reasonable to assume that not all participants will follow one or the other strategy. In the special issue for *Journal of Mathematical Psychology* alone, the articles by Kary et al. (2016) and Turner et al. (2016) demonstrate considerable heterogeneity among participants: the behaviour of some participants is predicted much better by one model, the behaviour of other participants is predicted much better by the competing model, and the behaviour of a third set of participants is predicted by the models about equally well (see also Steingroever et al., 2016b).

The standard Bayes factor tests determines whether *all* participants are better predicted by model \mathcal{M}_0 or whether *all* participants are better predicted by model \mathcal{M}_1 . Therefore, one can construct situations in which the data support model \mathcal{M}_0 for 99 out of 100 participants, and nevertheless the Bayes factor strongly prefers model \mathcal{M}_1 . We believe that in these hierarchical scenarios, the mixture model approach is a valuable technique that can offer additional insight.

The above considerations suggest that the mixture approach relaxes Jeffreys's conceptualisation of a hypothesis test. More precisely, Jeffreys viewed the null hypothesis as a general law, which by definition implies that the membership variables z_j are either all zeroes or all ones. Note that by embedding the models into an artificial encompassing model, Kamary et al. (2014) transformed the testing problem into one of estimation. Jeffreys, however, did not feel that estimation is appropriate when the test of a general law is at hand:

“Broad used Laplace’s theory of sampling, which supposes that if we have a population of n members, r of which may have a property φ , and we do not know r , the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find all of them with φ , the posterior probability that all n are φ ’s is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been sampled. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919-1923.” (Jeffreys, 1980, p. 452)

Wrinch and Jeffreys (1919, 1921, 1923) argued that within an estimation framework, a general law such as \mathcal{H}_0 : “All swans are white” cannot gain much evidence until almost all swans have been inspected.¹ Moreover, common sense prescribes that the plausibility of a general law increases with every observation in accordance with the law, that is, $s = n$ number of successes within n trials. Jeffreys (1961, p. 256) operationalised the general law as a binomial model \mathcal{M}_0 with θ_0 fixed and its negation as the binomial model \mathcal{M}_1 with a θ free to vary. With a uniform prior on θ this then leads to a Bayes factor of $\text{BF}_{01}(d) = \frac{(n+1)!}{s!f!} \theta_0^s (1-\theta_0)^f$,

¹We now know that this particular statement \mathcal{H}_0 does not hold true, since Australia is home to many black swans. The statement itself however cannot be empirically discarded until the first exception is actually observed.

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

where n denotes the total number of trials, s the number of successes, and f the numbers of failures.

When only successes are observed (i.e., observations consistent with the general law $\mathcal{H}_0 : \theta_0 = 1$), the Bayes factor simplifies to $n + 1$; a single failure, on the other hand, indicates infinite evidence against the general law: the observation of a single black swan is interocular, as it conclusively falsifies the general law “all swans are white”. Hence, Jeffreys’s Bayes factor formalises inductive reasoning and the logic of proof by contradiction.

The discussion above indicates that the mixture model approach does not formalise inductive reasoning and the logic of proof by contradiction: after having observed 10,000 white swans, the observation of a single black swan will not greatly affect the mixture proportion – the mixture proportion still reflects the fact that there is a great preponderance of white swans. However, in Jeffreys conceptualisation, the single exception utterly destroys the general law.

Another concern with the mixture model approach is that it is relatively insensitive to the shape of the prior distributions. Of course, this is also its strength, as this is needed to avoid the JLB paradox. However, models that make correct predictions should receive more reward when these predictions are risky, and the degree of risk is partly encoded in the shape of the prior distributions. For instance, suppose we model a binomial parameter θ and assume that $\mathcal{M}_1 : \theta \sim U[1/2, 1]$ and $\mathcal{M}_2 : \theta \sim U[0, 1]$; further, suppose the observed data are highly consistent with the simpler model \mathcal{M}_1 . Because the predictions from \mathcal{M}_1 are twice as risky as those from \mathcal{M}_2 we would want to prefer \mathcal{M}_1 over \mathcal{M}_2 , and in fact, the Bayes factor in favour of \mathcal{M}_1 against \mathcal{M}_2 is asymptotically equal to 2 (e.g., Heck et al., 2015; Shiffrin et al., 2016).

3.2.4 Conclusion

Scientific learning involves more than just testing general laws and invariances. Estimation and exploration are important and the mixture approach has a lot to offer in this respect, particularly in hierarchical settings where the general law is unlikely to hold for all participants simultaneously. Other advantages of the mixture approach are apparent as well. For instance, Example 3.1 in Kamary et al. (2014) compares a Poisson distribution with parameter λ to a geometric distribution with parameter p (see Robert, 2015 for R code). The comparison begins by relating the parameterisations to each other by setting $p = (1 + \lambda)^{-1}$, which allows the use of the improper Jeffreys’s prior (with respect to the Poisson distribution) $\pi(\lambda) \propto \lambda^{-1}$ over the two models. Note how this procedure resembles Jeffreys’s recommendation for common parameters even though the arguments differ. Moreover, the resulting posterior $\pi(\lambda | d)$ is then calculated from the mixture of the likelihoods of both models. The simulations show that the mixture approach performs well. We do not know how a Jeffreys’s Bayes factor can be constructed to deal with a test between two models of different relational forms as Jeffreys was only concerned with nested model comparisons (e.g., Robert, 2016).

The mixture approach is not fully automatic, however, and requires some thoughts on how the priors should be chosen. In particular, one cannot naively use improper priors on the test-relevant parameters, as this may yield posteriors

that are also improper (Grazian and Robert, 2015). This was acknowledged by Robert (2016) who used an (arbitrary) standard normal prior on μ in a t -test. Our implementation of this recommendation leads to a posterior median ranging from 0.3 to 0.9, for the interocular data with $n = 4$, $\bar{x} = 7$ and $s^2 = 0$, while α should be 1.0 if it were information consistent. More recently, Kamary et al. (2016) proposed a noninformative reparametrisation for location-scale mixtures to resolve the aforementioned arbitrariness. Hence, as with a Jeffrey's Bayes factor, one should choose the priors carefully when one conceptualises model selection as parameter estimation in a mixture model.

Lastly, Robert notes that the mixture approach is superior to the Bayes factor as it leads to a faster accumulation of α to the null. The parametric convergence rate of \sqrt{n} follows immediate from casting the testing problem as one of estimation. Similarly, it should be noted that Johnson and Rossell (2010) also use the rate of convergence as a motivation for their Bayes factor approach. We are unsure whether this rate is relevant as we do not consider a testing problem as one of estimation. In the end the Bayes factor and the mixture approach of Kamary et al. (2014) simply answer different questions. The choice which method to use should not be based on the rate of convergence, but on the research question the user seeks to address.²

3.3 Rejoinder to Chandramouli and Shiffrin

Chandramouli and Shiffrin (2016) put forward a thought-provoking proposal which aims to explain and extend Bayesian induction using simple matrix algebra. We have given this novel idea considerable thought and outline some of our reservations below.³

We believe that Chandramouli and Shiffrin (henceforth C&S) put forward a belief propagation procedure that allows us to verify whether two given models, say, \mathcal{M}_1 and \mathcal{M}_2 align with a scientist's prior belief about the true data generating process $p^*(X)$. Instead of setting the priors onto the two given models \mathcal{M}_1 and \mathcal{M}_2 directly, C&S recommend to first elicit a scientist's prior belief about the true data generating $p^*(X)$ in the most general setting. This prior belief is then subsequently translated into priors on the models. Hence, the resulting prior model probabilities $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$ are *derived*.

In contrast, a Jeffreys's Bayes factor follows from a top-to-bottom procedure, where the top level is concerned with the comparison *between* two models (i.e., model classes) for which one has to (subjectively) *choose* prior model probabilities $P(\mathcal{M}_i)$. Based on top level desiderata, i.e., a coherent comparison between the two models, we then *derive* the pair of priors π_1 and π_2 on the lower level that are concerned with the parameters (i.e., model instances) *within* the models \mathcal{M}_1 and \mathcal{M}_2 respectively. In effect, the sole purpose of the pair π_1, π_2 is to mediate

²We thank Joris Mulder for attending us to this.

³The second and third authors are in a state of perpetual confusion regarding the details of the Chandramouli and Shiffrin proposal. All credit concerning this section goes to the first author, who, as such, takes full responsibility for any errors here. For a thorough understanding of our reply, we recommend to have the comment of Chandramouli and Shiffrin (2016) on hand.

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

scientific learning through the Bayes factor, that is, to update the prior model odds to posterior model odds.

On the other hand, the C&S induction scheme is a bottom-up approach based on the philosophy that the whole is the sum of its parts. At the lowest level, one has to *subjectively elicit* the scientist's prior belief about the true data generating process. The procedure then elaborates on how this lowest level belief can be used to *derive* the model instance priors π_1 and π_2 at the intermediate level. By aggregating the model instance priors of π_1 and π_2 we then get the prior model probabilities $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$ at the top level. As such, this method is not free from subjective input on the lowest level.

Our major concern with the C&S method is the lack of invariance, which stems from their recommendation to operationalise their procedure with a seemingly innocent looking finite-dimensional matrix with M rows and W number of columns.⁴ By using a finite-dimensional matrix, C&S basically made a choice in how they tackle the statistical modelling problem. The resulting model priors $P(\mathcal{M}_i)$ are sensitive to this choice. More specifically, by initialising their procedure with a finite-dimensional matrix, they use discretised approximations of quantities that are essentially continuous. The approximation error due to discretisation is non-negligible, as it permeates through all subsequent steps due to the bottom-up nature leading to an ill-defined Bayes factor.

In brief, we believe that the C&S approach has to overcome some challenges before their procedure can be perceived as an extension of traditional Bayes factors, let alone Jeffreys's Bayes factors. We have three remarks: (1) The C&S procedure is not invariant to how one discretises the statistical modelling problem; (2) the subjective assessment of the priors on the lowest level and the resulting prior model probabilities $P(\mathcal{M}_i)$ on the top level are, therefore, ill-defined, and (3) model selection based on posterior predictive p -statistics does not lead to a proper measure of evidence.

This paper continues as follows: We first apply the C&S induction scheme to a concrete example. Then we show that we get different results when we choose a different finite-dimensional matrix to operationalise the C&S induction scheme. Lastly, we argue that the implicit discretisation necessary for the finite-dimensional matrix is the main culprit of the resulting lack of invariance.

3.3.1 Running example

To illustrate why we believe that the C&S method is essentially a belief propagation procedure, we consider a random variable X with a finite number of outcomes W . This W is denoted as n in Chandramouli and Shiffrin (2016) and defines the number of columns in their matrix representations (i.e., their Figures 1 and 2). To simplify matters, we use an example (taken from Ly et al., 2017c) where X has $W = 3$ number of outcomes.

⁴We divert from the C&S notation, where the matrix is $M \times N$ dimensional, as the number of columns does *not* correspond with the number of samples in a data set. Instead, the number of columns refers to the number of possible outcomes a random variable can take on, we use w and W instead.

Example 3.3.1 (A Psychological Task with Three Outcomes). *In the training phase of a source-memory task, the participant is presented with two lists of words on a computer screen. List \mathcal{L} is projected on the left-hand side and list \mathcal{R} is projected on the right-hand side. In the test phase, the participant is then presented with two words, side by side, that can stem from either list, thus, ll, lr, rl, rr. At each trial, the participant is asked to categorise these pairs as either:*

- x_1 meaning both words come from the left list, i.e., “ll”,
- x_2 meaning the words are mixed, i.e., “lr” or “rl”,
- x_3 meaning both words come from the right list, i.e., “rr”.

Thus, the random variable X has $W = 3$ outcomes. To ease the discussion, we assume that the words presented to the participant are “rr”. \diamond

As model \mathcal{M}_1 we take the so-called individual-word strategy. A participant guided by this strategy will consider each word individually and compare it with list \mathcal{R} only. Within this model \mathcal{M}_1 , the parameter is given by $\theta_1 = \vartheta$, which we interpret as the participant’s “right-list recognition ability”. Hence, when the participant is presented with the pair “rr” she will respond x_1 with probability $(1 - \vartheta)^2$, thus, two failed recollections; x_2 with probability $2\vartheta(1 - \vartheta)$, thus, one failed and one successful recollection; x_3 with probability ϑ^2 , thus, two successful recollections.

More compactly, a participant guided by this strategy generates the outcomes $[x_1, x_2, x_3]$ with the following three probabilities $f_1(X | \vartheta) = [(1 - \vartheta)^2, 2\vartheta(1 - \vartheta), \vartheta^2]$, respectively. Note the data generative formulation. For instance, when the participant’s true ability is $\vartheta^* = 0.9$, the three outcomes $[x_1, x_2, x_3]$ are then generated with the three probabilities $f_1(X | 0.9) = [0.01, 0.18, 0.81]$ respectively. We call the function $f_i(X | \theta_i)$ with θ_i fixed a probability mass function (pmf) or model instance of \mathcal{M}_i .⁵ Hence, every ϑ in $(0, 1)$ yields a pmf that defines W number of probabilities. In effect, the model \mathcal{M}_1 consists of a collection of pmfs, which C&S refer to as a model class.

As a competing model \mathcal{M}_2 , we take the so-called only-mixed strategy. Within this model \mathcal{M}_2 , the parameter is given by $\theta_2 = \alpha$, which we interpret as the participant’s “mixed-list differentiability ability”. With probability α the participant first checks whether the presented pair of words is mixed. If she perceives it as mixed, she then produces the outcome x_2 with probability α . If she does not perceive the pair of words as mixed, the participant then randomly chooses x_1 or x_3 each with probability $(1 - \alpha)/2$.

More compactly, a participant guided by this strategy generates the outcomes $[x_1, x_2, x_3]$ with the following three probabilities $f_2(X | \alpha) = [(1 - \alpha)/2, \alpha, (1 - \alpha)/2]$, respectively. Again we formulated the model as a data generative process. For instance, when the participant’s true ability is $\alpha^* = 1/3$, the three outcomes $[x_1, x_2, x_3]$ are then generated with the same probability, i.e., $f_2(X | 1/3) =$

⁵C&S call the function $f_1(X | 0.9)$ a data distribution predicted by the model instance $\vartheta = 0.9$. When we use a capital X we mean the three probabilities simultaneously. On the other hand, a small letter x refers to the probability with which it is generated, for instance, $f_1(x_w | 0.9) = 0.18$ when $w = 2$.

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

$[1/3, 1/3, 1/3]$. Note that this last pmf $f_2(X | 1/3)$ is not in the collection of pmfs defined by \mathcal{M}_1 . Similarly, the pmf $f_1(X | 0.9)$ is not a member of the collection of pmfs defined by \mathcal{M}_2 .

The two models \mathcal{M}_1 and \mathcal{M}_2 share only one pmf (model instance), that is, the pmf indexed by $\vartheta = 0.5$ within \mathcal{M}_1 and, coincidentally, when $\alpha = 0.5$ within \mathcal{M}_2 . We use these two models \mathcal{M}_1 and \mathcal{M}_2 to explain the C&S belief propagation procedure.

3.3.2 Chandramouli and Shiffrin's procedure for induction

For a Bayesian analysis we need priors on the model instances, which we denote by $\pi_i(\theta_i)$ as we have done before,⁶ and the priors on the models $P(\mathcal{M}_i)$. Instead of doing so directly, C&S recommend to first (Step 1) elicit the scientist's prior belief about the true data generating process $p^*(X)$ in the most general setting. Next (Step 2) this subjectively chosen prior belief about $p^*(X)$ is used to derive the model instance priors $\pi_i(\theta_i)$ and, subsequently, the model class priors $P(\mathcal{M}_i)$. Lastly, (Step 3) C&S recommend to use posterior p -statistics for inference.

3.3.2.1 Step 1: Eliciting the prior on candidate true data generating pmfs

In our example, the true data generating pmf $p^*(X)$ defines three probabilities $p^*(X) = [p^*(x_1), p^*(x_2), p^*(x_3)]$ with which it generates the three outcomes $[x_1, x_2, x_3]$. For instance, a first candidate true data generating pmf could be $p(X | \psi_1) = [0.0, 0.0, 1.0]$, where ψ_1 is an index/label for later reference. A second candidate true data generating pmf could be $p(X | \psi_2) = [0.0, 0.1, 0.9]$ and so forth and so on. This method yields a candidate set of true pmfs that we depicted in Table 3.1. The “matrix” depicted in Table 3.1 is a simplification of the table in Figure 1 in Chandramouli and Shiffrin (2016) with $M = 66$ rows and $W = 3$ columns. Please ignore the quantities to the right of the double vertical line for the moment. Note that the number of rows $M = 66$ is a result of our arbitrary choice of using a step size of 0.1 on the probabilities. Furthermore, recall that the pmfs $p(X | \psi_m)$ are candidates for the true data generating pmf $p^*(X)$ and may not have any connection with the models \mathcal{M}_1 and \mathcal{M}_2 specified above. Of particular interest is the pmf $p(X | \psi_{62}) = [0.8, 0.1, 0.1]$, which is neither a member of \mathcal{M}_1 nor of \mathcal{M}_2 ,⁷ but because it defines a valid pmf it is, nonetheless, a candidate true data generating pmf.

Given this finite-dimensional matrix of Table 3.1, C&S then recommend to elicit a scientist's prior belief by setting prior beliefs $\lambda(\psi_m)$ for $m = 1, \dots, M$, thus,

⁶C&S denote this by $p_0(\theta_i)$. Instead, we use the Greek letter π_i to distinguish this model instance prior from the prior model probability $P(\mathcal{M}_i)$ on the top level. The subscript i refers to the model membership.

⁷A pmf of \mathcal{M}_1 with $f_1(x_1 | \vartheta) = 0.8$ requires $\vartheta \approx 0.11$. However, this automatically yields $f_1(x_2 | 0.11) = 0.19$. Hence, there is no ϑ in \mathcal{M}_1 that leads to the pmf indexed by ψ_{62} . Similarly, a pmf of \mathcal{M}_2 necessarily has $f_2(x_1 | \alpha) = f_2(x_3 | \alpha)$, which is clearly not the case for the pmf indexed by ψ_{62} .

Table 3.1: The matrix is a simplified version of the matrix found in Figure 1 of C&S with $M = 66$ and $W = 3$. The quantities under the columns with $D(\psi_m, \mathcal{M}_1)$ and $D(\psi_m, \mathcal{M}_2)$ at the top refer to the KL-divergences, see the main text. The parameter under θ_i refers to the model instance that the pmf $p(X | \psi_m)$ is allocated to within the model under \mathcal{M}_i . For example, the candidate true pmf $p(X | \psi_{18})$ is allocated to the model instance $f_1(X | \vartheta = 0.60)$ of model class \mathcal{M}_1 .

	x_1	x_2	x_3	$D(\psi_m, \mathcal{M}_1)$	$D(\psi_m, \mathcal{M}_2)$	θ_i	\mathcal{M}_i
ψ_1	0.0	0.0	1.0	0	0.693	$\vartheta = 1.00$	\mathcal{M}_1
ψ_2	0.0	0.1	0.9	0.002	0.624	$\vartheta = 0.95$	\mathcal{M}_1
ψ_3	0.0	0.2	0.8	0.011	0.555	$\vartheta = 0.90$	\mathcal{M}_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
ψ_{11}	0.0	1.0	0.0	0.693	0	$\alpha = 1.00$	\mathcal{M}_2
ψ_{12}	0.1	0.0	0.9	0.325	0.368	$\vartheta = 0.90$	\mathcal{M}_1
ψ_{13}	0.1	0.1	0.8	0.137	0.310	$\vartheta = 0.85$	\mathcal{M}_1
ψ_{14}	0.1	0.2	0.7	0.060	0.253	$\vartheta = 0.80$	\mathcal{M}_1
ψ_{15}	0.1	0.3	0.6	0.019	0.198	$\vartheta = 0.75$	\mathcal{M}_1
ψ_{16}	0.1	0.4	0.5	0.011	0.145	$\vartheta = 0.70$	\mathcal{M}_1
ψ_{17}	0.1	0.5	0.4	0.005	0.096	$\vartheta = 0.65$	\mathcal{M}_1
ψ_{18}	0.1	0.6	0.3	0.032	0.052	$\vartheta = 0.60$	\mathcal{M}_1
ψ_{19}	0.1	0.7	0.2	0.089	0.017	$\alpha = 0.70$	\mathcal{M}_2
ψ_{20}	0.1	0.8	0.1	0.193	0	$\alpha = 0.80$	\mathcal{M}_2
ψ_{21}	0.1	0.9	0.0	0.427	0.069	$\alpha = 0.90$	\mathcal{M}_2
ψ_{22}	0.2	0.0	0.8	0.500	0.193	$\alpha = 0.00$	\mathcal{M}_2
ψ_{23}	0.2	0.1	0.7	0.254	0.148	$\alpha = 0.10$	\mathcal{M}_2
ψ_{24}	0.2	0.2	0.6	0.133	0.104	$\alpha = 0.20$	\mathcal{M}_2
ψ_{25}	0.2	0.3	0.5	0.057	0.067	$\vartheta = 0.65$	\mathcal{M}_1
ψ_{26}	0.2	0.4	0.4	0.014	0.034	$\vartheta = 0.60$	\mathcal{M}_1
ψ_{27}	0.2	0.5	0.3	0.000	0.010	$\vartheta = 0.55$	\mathcal{M}_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
ψ_{61}	0.8	0.0	0.2	0.500	0.193	$\alpha = 0.00$	\mathcal{M}_2
ψ_{62}	0.8	0.1	0.1	0.137	0.310	$\vartheta = 0.15$	\mathcal{M}_1
ψ_{63}	0.8	0.2	0.0	0.011	0.555	$\vartheta = 0.10$	\mathcal{M}_1
ψ_{64}	0.9	0.0	0.1	0.325	0.368	$\vartheta = 0.10$	\mathcal{M}_1
ψ_{65}	0.9	0.1	0.0	0.003	0.624	$\vartheta = 0.05$	\mathcal{M}_1
ψ_{66}	1.0	0.0	0.0	0	0.693	$\vartheta = 0.00$	\mathcal{M}_1

on each candidate true data generating pmf $p(X | \psi_m)$.⁸ For example, $\lambda(\psi_{62}) = 0.7$ means that the scientist bestows a large portion of belief to the pmf indexed by ψ_{62} as being the true generating pmf $p^*(X)$. Furthermore, $\lambda(\psi_{61}) + \lambda(\psi_{62}) + \lambda(\psi_{63}) = 0.90$ means that the scientist is quite sure that the participant will generate the response x_1 with 80% chance. As λ represents the scientist's prior belief, we necessarily require that $\sum_{m=1}^M \lambda(\psi_m) = 1$.

⁸C&S denote this prior pmf probability as $p_0(\psi_m)$. Instead, we use the Greek letter λ to distinguish this prior pmf probability on the lowest level from the model instance prior $\pi_i(\theta_i)$ on the intermediate level and the prior model probabilities $P(\mathcal{M}_i)$ on the top level.

3.3.2.2 Step 2: Propagating the prior belief to yield the prior model probabilities

Once the prior beliefs $\lambda(\psi_m)$ about the true data generating $p^*(X)$ are chosen, C&S commence their belief propagation procedure by redistributing $\lambda(\psi_m)$ over the two models. Recall that a model (class) \mathcal{M}_i defines a collection of pmfs (model instances) denoted as $f_i(X | \theta_i)$. The allocation of the prior pmf belief of the first candidate true pmf in Table 3.1, that is, $\lambda(\psi_1)$, is easy, because the associated pmf $p(X | \psi_1) = [0.0, 0.0, 1.0]$ does not belong to \mathcal{M}_2 , but it is a member of \mathcal{M}_1 ; the pmf indexed by ψ_1 is a model instance of \mathcal{M}_1 when $\theta_1 = \vartheta = 1$. C&S therefore allocate the prior pmf probability $\lambda(\psi_1)$ to the model instance $\pi_1(\vartheta = 1)$ of \mathcal{M}_1 . On the other hand, the pmf $p(X | \psi_{62}) = [0.8, 0.1, 0.1]$ is neither a member of \mathcal{M}_2 nor does it belong to \mathcal{M}_1 . To nonetheless allocate this prior pmf belief $\lambda(\psi_{62})$ to a model instance of either \mathcal{M}_1 or \mathcal{M}_2 , C&S use a divergence measure denoted by D . For simplicity we take as D the Kullback-Leibler (KL) divergence, which is a measure of dissimilarity. The KL-divergence from a candidate true pmf indexed by ψ_m to a model instance of \mathcal{M}_i is defined as

$$D(\psi_m, \theta_i | \mathcal{M}_i) = \sum_{w=1}^W p(x_w | \psi_m) \log \frac{p(x_w | \psi_m)}{f(x_w | \theta_i)}, \quad (3.3.1)$$

and the larger this divergence, the more dissimilar the model instance $f_i(X | \theta_i)$ is from the candidate true data generating pmf $p(X | \psi_m)$. For example, a direct calculation shows that the KL-divergence between the candidate true $p(X | \psi_1)$ in Table 3.1 to the model instance of \mathcal{M}_1 with $\theta_1 = \vartheta = 1.0$ is given by $D(\psi_1, \theta_1 = 1.0 | \mathcal{M}_1) = 0$. The KL-divergence is zero if and only if the pmfs indexed by ψ_m and the model instance indexed by θ_i are exactly the same, hence, their dissimilarity is zero.

The KL-divergence between the candidate true $p(X | \psi_m)$ and a collection of pmfs defined by the model \mathcal{M}_i is given by $D(\psi_m, \mathcal{M}_i) = \min_{\theta_i} D(\psi_m, \theta_i | \mathcal{M}_i)$. That is, the dissimilarity between the candidate true data generating pmf ψ_m and the model \mathcal{M}_i is the smallest dissimilarity between $p(X | \psi_m)$ and the model instances $f_i(X | \theta_i)$ of model \mathcal{M}_i . For example, a direct calculation shows that the KL-divergence from the candidate true data generating pmf $p(X | \psi_{62})$ to \mathcal{M}_1 is given by $D(\psi_{62}, \mathcal{M}_1) = D(\psi_{62}, \theta_1 = 0.15 | \mathcal{M}_1) = 0.137$. Similarly, the KL-divergence between the same candidate true data generating pmf to \mathcal{M}_2 is given by $D(\psi_{62}, \mathcal{M}_2) = D(\psi_{62}, \theta_2 = 0.1 | \mathcal{M}_2) = 0.310$. Because the divergence from the candidate true pmf $p(X | \psi_{62})$ to \mathcal{M}_1 is smaller than the divergence to \mathcal{M}_2 , the C&S procedure implies that we should allocate all the prior pmf mass $\lambda(\psi_{62})$ to the prior model instance probability $\pi_1(\vartheta = 0.15)$ belonging to \mathcal{M}_1 .

We suspect that the underlying idea of this belief allocation procedure is based on the idea of chaining. Thus, if $\lambda(\psi_{62}) = 0.70$, the scientist has much fate in $p(X | \psi_{62})$ being the true data generating pmf. However, as $p(X | \psi_{62})$ is not in the model \mathcal{M}_1 nor in \mathcal{M}_2 , the C&S procedure then recommends to go for the next best thing; assigning the pmf prior $\lambda(\psi_{62})$ to the model instance that is most similar to $p(X | \psi_{62})$, in this case, $\pi_1(\vartheta)$ with $\vartheta = 0.15$.

This redistribution of the pmf prior $\lambda(\psi_m)$ to model instance priors can be

read from their table in Figure 1 in Chandramouli and Shiffrin (2016) from left to right.⁹

In our Table 3.1 the numbers under $D(\psi_m, \mathcal{M}_1)$ and $D(\psi_m, \mathcal{M}_2)$ represents the KL-divergence from the candidate true pmf indexed by ψ_m to the models \mathcal{M}_1 and \mathcal{M}_2 respectively. The parameter value under θ_i indicates which parameter value ϑ within \mathcal{M}_1 or α within \mathcal{M}_2 corresponds to the model instance that is closest to the pmf of ψ_m . The last column indicates whether the ψ_m is eventually allocated to \mathcal{M}_1 or \mathcal{M}_2 .

As in their table in Figure 1 of Chandramouli and Shiffrin (2016), note that there are multiple candidates ψ_m s allocated to certain parameter values in our Table 3.1. For example, the candidate pmfs indexed by ψ_3 and ψ_{12} are both allocated to the same model instance indexed by $\vartheta = 0.90$ within \mathcal{M}_1 . As such, C&S derive the prior on the model instances as $\pi_1(\vartheta) = \sum \lambda(\psi_m)$, where the sum is over the candidates ψ_m which have the same ϑ in the column under θ_i . For example, $\pi_1(\vartheta = 0.90) = \lambda(\psi_3) + \lambda(\psi_{12})$.

After allocating all the M number of prior pmf probability $\lambda(\psi_m)$ to the model instances of either model classes, we have $\pi_1(\vartheta_k)$ and $\pi_2(\alpha_{\tilde{k}})$ for $k = 1, \dots, K$ and $\tilde{k} = 1, \dots, \tilde{K}$. The K indicates the number of unique values of ϑ s in the column under θ_i . As there are multiple candidates allocated to certain parameter values we typically have $K + \tilde{K} < M$. With the model instance priors at hand, the C&S scheme tells us to aggregate them to yield prior model probabilities, i.e., $P(\mathcal{M}_1) = \sum_{k=1}^K \pi_1(\vartheta_k)$ and $P(\mathcal{M}_2) = \sum_{\tilde{k}=1}^{\tilde{K}} \pi_2(\alpha_{\tilde{k}})$. As a result of $\sum_{m=1}^M \lambda(\psi_m) = 1$ we have $P(\mathcal{M}_1) + P(\mathcal{M}_2) = 1$.

3.3.2.3 Step 3: Posterior predictive p -statistics

So far, we only discussed the C&S belief propagation procedure as a method to translate a scientist's prior belief $\lambda(\psi)$ about the true data generating $p^*(X)$ to prior beliefs on the model instances $\pi_i(\theta_i)$, which can then be used to define prior beliefs on the models $P(\mathcal{M}_i)$. These priors can be used for inference after we observe data d . As in C&S, we simplify the discussion by supposing that the data consist of one observation where the participant responded with x_1 .

To invert the data generative view of pmfs, we fix the data part of each pmf at the observation $p(X | \psi_m) = p(d | \psi_m)$ and consider the pmfs as a function of ψ_m , i.e., as likelihood functions. Bayes' rule then allows us to update the subjectively chosen pmf prior to a pmf posterior using all specified candidate likelihood functions indexed by the ψ_m s, that is, $\lambda(\psi_m | d) = p(d | \psi_m) \lambda(\psi_m) / C$, for $m = 1, \dots, M$, where the normalising constant C is given by $C = \sum_{m=1}^M p(d | \psi_m) \lambda(\psi_m)$. Recall that the rows $p(X | \psi_m)$, thus, the likelihood functions, themselves do not need to belong to the models \mathcal{M}_1 and \mathcal{M}_2 . In fact, most of them do not, as most of the entries under $D(\psi_m, \mathcal{M}_1)$ and $D(\psi_m, \mathcal{M}_2)$ are non-zero.

For inference concerning replication studies, C&S recommend using posterior predictive p -statistics. For example, the observations d_{orig} of the original experiment might suggest that a participant's "right-list recognition ability" ϑ is a half. To test whether this postulate $\vartheta = 0.5$ can be reproduced, C&S recommend to

⁹We are unsure what φ in their table indicates.

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

update the subjectively chosen pmf prior about the true $p^*(X)$ to a posterior yielding $\lambda(\psi_m | d_{\text{orig}})$. Recall that this posterior is also based on likelihood functions $p(d | \psi_m)$ that do not belong to \mathcal{M}_1 as discussed above. For example, if $\lambda(\psi_{62}) > 0$ then $p(X | \psi_{62}) = [0.8, 0.1, 0.1]$ in Table 3.1 is used as a likelihood to relate the observations d_{orig} to ψ_{62} . Because there is no ϑ that leads to $p(X | \psi_{62})$, see the footnote at the end of Section 3.3.2.1, the likelihood function at ψ_{62} does not and *cannot* extract information about ϑ from d_{orig} .

Nonetheless, C&S use the posterior $\lambda(\psi_m | d_{\text{orig}})$ to weight all the candidate true pmfs in Table 3.1 resulting in a posterior predictive

$$p(x_w | d_{\text{orig}}) = \sum_{m=1}^M p(x_w | \psi_m) \lambda(\psi_m | d_{\text{orig}}) \text{ for } w = 1, \dots, W. \quad (3.3.2)$$

This posterior predictive is used as a sampling distribution, i.e., it defines the probabilities with which new data are generated. If the actually observation d_{rep} is very improbable under this predictive, then the C&S procedure prescribes this as a failure of reproducibility. The problem with this prediction is that it is also calculated from the predictions of $p(X | \psi_{62})$, even though this pmf ψ_{62} has no connection to ϑ whatsoever.

In sum, it seems that the C&S recommendation for replication boils down to comparing the observed data d_{rep} in a replication attempt using the posterior predictive as a sampling distribution, which is based on irrelevant likelihood functions and subjective belief $\lambda(\psi_m)$. Moreover, by using the posterior predictive as a sampling distribution to assess replication, this method shares many pitfalls with common p -value tests and therefore does not quantify evidence (e.g., Bayarri and Berger, 2000; Wagenmakers, 2007).

3.3.2.4 C&S Bayes factors

Although C&S do not recommend to use Bayes factors for inference, they note that Bayes factors can be constructed from their belief propagation procedure. The main idea is to reuse the belief propagation procedure, but this time to redistribute the posterior beliefs $\lambda(\psi_m | d)$ about the true data generating $p^*(X)$ to posterior beliefs for the “model instances” $\pi_i(\hat{\theta}_i | d)$, which can then be used to define posterior beliefs on the “models” $P(\hat{\mathcal{M}}_i | d)$. We are reluctant to call $P(\hat{\mathcal{M}}_i | d)$ the posterior model probabilities, because they are calculated using likelihood functions that do not belong to \mathcal{M}_i (hence, the hats in our notation). There are now two ways to derive a Bayes factor based on the quantities resulting from the C&S belief propagation procedure.

The first method involves the ratio of the posterior and prior model odds, that is,

$$\hat{BF}_{12}(d) = \frac{P(\hat{\mathcal{M}}_1 | d) / P(\hat{\mathcal{M}}_2 | d)}{P(\mathcal{M}_1) / P(\mathcal{M}_2)}. \quad (3.3.3)$$

This Bayes factor depends on the subjectively chosen prior beliefs $\lambda(\psi_m)$ about $p^*(X)$, the chosen divergence measure D , and –most troublesome– on the collection of candidate likelihood functions $p(d | \psi_m)$ rather than on the likelihood that belong to the respective models.

The second method involves the ratio of marginal likelihoods, that is,

$$\tilde{BF}_{12}(d) = \frac{\sum_{k=1}^K f_1(d | \vartheta_k) \pi_1(\vartheta_k)}{\sum_{\tilde{k}=1}^{\tilde{K}} f_2(d | \alpha_{\tilde{k}}) \pi_2(\alpha_{\tilde{k}})}. \quad (3.3.4)$$

In contrast to $\hat{BF}_{12}(d)$, this Bayes factor is calculated from the likelihoods $f_i(d | \theta_i)$ that actually do belong to the respective models. Hence, $\hat{BF}_{12}(d)$ and $BF_{12}(d)$ will differ from each other.

We have some reservations about the Bayes factor as defined in Eq. (3.3.3) or Eq. (3.3.4) as a generalisation of traditional Bayes factors. First, a traditional Bayes factor leads to the same quantity whether it is computed as the ratio of the posterior and prior model odds, or as the ratio of marginal likelihoods. Second, a traditional Bayes factor would involve continuous integrals, whenever the parameters ϑ and α are free to vary in continuous intervals. The replacement of the integrals by finite sums is an artefact of only considering a finite number M of candidate true pmfs $p(X | \psi_m)$.

3.3.3 Lack of invariance

Our major concern with Bayes factors calculated from the C&S approach, however, is rooted in its operationalisation using a finite-dimensional matrix (e.g., Table 3.1), as it causes a lack of invariance affecting every step of their belief propagation procedure. As such, two scientist with the same subjective belief $\lambda(\psi)$ about the true $p^*(X)$ using the same divergence measure D , but with a different finite-dimensional matrix will calculate different Bayes factors.

We appreciate the attempt by C&S to assess how well models represent the true data generating process. Their procedure considers all possible data generating pmfs and as such can account for model misspecification. Although attractive, such an unrestrictive view leads to complications when one is concerned with testing models for which one has to set priors. The C&S recommendation is to do so subjectively, which we consider nigh impossible. More specifically, the collection of all possible data generative pmfs \mathcal{P} is typically hard to describe and without a proper description even harder to subjectively assign prior beliefs to. Our paper continuous as follows: (1) We first characterise \mathcal{P} and simplify it with a parameterisation; (2) a different parameterisation of \mathcal{P} is then given leading to a different finite-dimensional matrix. (3) In effect, this leads to different prior beliefs, and (4) different allocations, thus, different Bayes factors. (5) Lastly, we remark how this problem is related to the invariance problem already solved by Jeffreys (1946) and what his solution implies for the C&S procedure.

3.3.3.1 Characterising the collection of all possible pmfs

When X has $W = 3$ number of outcomes, its true distribution $p^*(X)$ can then be characterised by $W - 1 = 2$ parameters. Recall that a pmf for X then defines the three chances $p(X) = [p(x_1), p(x_2), p(x_3)]$ with which it generates the outcomes $[x_1, x_2, x_3]$. The pmf must therefore satisfy two conditions: (i) it has to be non-negative, thus, $0 \leq p(x_w)$ for each outcome x_w of X with $w = 1, \dots, W$, and (ii)

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

the probabilities must sum to one, i.e., $\sum_{w=1}^W p(x_w) = 1$. Note that this holds true for any candidate true pmf $p(X | \psi_m)$ in Table 3.1. We call the collection of functions for which the conditions (i) and (ii) hold the collection of all possible pmfs or the full model and denote it by \mathcal{P} . The collection \mathcal{P} has an uncountably infinite number of members, each capable of being the true data generating process $p^*(X)$. By using a finite-dimensional matrix such as the one in Table 3.1, C&S, thus, restrict their prior belief elicitation to only $M = 66$ candidate true pmfs.

To show that even for $W = 3$ the full model \mathcal{P} is uncountable, we first parameterise \mathcal{P} , that is, we identify each possible true pmf of \mathcal{P} with a two dimensional parameter $\psi = (b, c)$. Given any pmf $p(x) = [p(x_1), p(x_2), p(x_3)]$, we define $b = p(x_1)$, $c = p(x_2)$ and set $\psi = (b, c)$. This construction is essentially a function ξ that maps a member of the full model \mathcal{P} into a parameter space Ψ of dimension $W - 1 = 2$. Using the inverse parameterisation ξ^{-1} we can identify every parameter $\psi = (b, c)$, where (i') $0 \leq b, c$ and (ii') $b + c \leq 1$, with a pmf such that the three outcomes $[x_1, x_2, x_3]$ are generated with the probabilities $p(X | \psi) = [b, c, 1 - b - c]$. As there are an uncountable number of $\psi = (b, c)$ s for which the conditions (i') and (ii') holds, we conclude that there are also an uncountable number of pmfs $p(X | \psi)$ in the full model \mathcal{P} for which (i) and (ii) holds.

3.3.3.2 Different parameterisations, different representation of \mathcal{P} : A different set of candidate true pmfs

The aforementioned parameterisation $\xi : \mathcal{P} \rightarrow \Psi$ relates to the candidate true pmfs of Table 3.1 as we have actually chosen $\psi_1 = (0.0, 0.0)$, $\psi_2 = (0.0, 0.1), \dots, \psi_{62} = (0.8, 0.1)$, $\psi_{63} = (0.8, 0.2)$, $\psi_{64} = (0.9, 0.0)$, $\psi_{65} = (0.9, 0.1)$, $\psi_{66} = (1.0, 0.0)$. The resulting $M = 66$ number of columns is due to the dependence between b and c .

A different parameterisation $\tilde{\xi}$ from the full model \mathcal{P} to a parameter space $\tilde{\Psi}$ is based on a “stick-breaking” approach. Given a $p(X)$ we then choose $\tilde{b} = p(x_1)$, $\tilde{c} = p(x_2)/[1 - p(x_1)]$ and define $\tilde{\psi} = (\tilde{b}, \tilde{c})$.¹⁰ Using the inverse parameterisation $\tilde{\xi}^{-1}$ we can also identify every parameter $\tilde{\psi} = (\tilde{b}, \tilde{c})$, where (i'+ii') $0 \leq \tilde{b}, \tilde{c} \leq 1$, with a pmf such that the three outcomes $[x_1, x_2, x_3]$ are generated with the probabilities $p(X | \psi) = [\tilde{b}, (1 - \tilde{b})\tilde{c}, (1 - \tilde{b})(1 - \tilde{c})]$. Note that every parameter $\tilde{\psi}$ lies within the unit square $\tilde{\Psi} = [0, 1] \times [0, 1]$, and that \tilde{b} and \tilde{c} can be chosen independently from each other. Again, as there are an uncountable number of elements in the unit square, we have an uncountable collection of candidate true pmfs \mathcal{P} . With this stick-breaking representation of \mathcal{P} and a step size of 0.1 we get the matrix depicted in Table 3.2.

This new matrix differs substantially from the previous one. First, it has more rows, thus, a larger number of candidate true pmfs; $M = 111$ compared to $M = 66$ in Table 3.1. Second, there are more candidate pmfs that imply that the first response x_1 is generated with 80% chance; eleven in Table 3.2 compared to three in Table 3.1.

¹⁰This only works if $p(x_1) \neq 1$. When $p(x_1) = 1$, we simply set $\tilde{c} = 0$ and define $\tilde{\psi} = (1, 0)$.

Table 3.2: The matrix is a simplified version of the matrix found in Figure 1 of C&S based on the different parameterisation $\tilde{\xi}$ defined in text. Note how the pmf $p(X | \tilde{\psi}_{19})$ is allocated to \mathcal{M}_2 .

	x_1	x_2	x_3	$D(\psi_m, \mathcal{M}_1)$	$D(\psi_m, \mathcal{M}_2)$	θ_i	\mathcal{M}_i
$\tilde{\psi}_1 = (0.0, 0.0)$	0.00	0.00	1.00	0	0.693	$\vartheta = 1.00$	\mathcal{M}_1
$\tilde{\psi}_2 = (0.0, 0.1)$	0.00	0.10	0.90	0.003	0.624	$\vartheta = 0.95$	\mathcal{M}_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$\tilde{\psi}_{17} = (0.1, 0.5)$	0.10	0.45	0.45	0.000	0.120	$\vartheta = 0.68$	\mathcal{M}_1
$\tilde{\psi}_{18} = (0.1, 0.6)$	0.10	0.54	0.36	0.013	0.078	$\vartheta = 0.63$	\mathcal{M}_1
$\tilde{\psi}_{19} = (0.1, 0.7)$	0.10	0.63	0.27	0.046	0.041	$\alpha = 0.63$	\mathcal{M}_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$\tilde{\psi}_{88} = (0.7, 1.0)$	0.70	0.30	0.00	0.027	0.485	$\vartheta = 0.15$	\mathcal{M}_1
$\tilde{\psi}_{89} = (0.8, 0.0)$	0.80	0.00	0.20	0.500	0.193	$\alpha = 0.00$	\mathcal{M}_2
$\tilde{\psi}_{90} = (0.8, 0.1)$	0.80	0.02	0.18	0.393	0.212	$\alpha = 0.02$	\mathcal{M}_2
$\tilde{\psi}_{91} = (0.8, 0.2)$	0.80	0.04	0.16	0.315	0.233	$\alpha = 0.04$	\mathcal{M}_2
$\tilde{\psi}_{92} = (0.8, 0.3)$	0.80	0.06	0.14	0.248	0.256	$\vartheta = 0.17$	\mathcal{M}_1
$\tilde{\psi}_{93} = (0.8, 0.4)$	0.80	0.08	0.12	0.189	0.281	$\vartheta = 0.16$	\mathcal{M}_1
$\tilde{\psi}_{94} = (0.8, 0.5)$	0.80	0.10	0.10	0.137	0.310	$\vartheta = 0.15$	\mathcal{M}_1
$\tilde{\psi}_{95} = (0.8, 0.6)$	0.80	0.12	0.08	0.091	0.342	$\vartheta = 0.14$	\mathcal{M}_1
$\tilde{\psi}_{96} = (0.8, 0.7)$	0.80	0.14	0.06	0.053	0.378	$\vartheta = 0.13$	\mathcal{M}_1
$\tilde{\psi}_{97} = (0.8, 0.8)$	0.80	0.16	0.04	0.022	0.421	$\vartheta = 0.12$	\mathcal{M}_1
$\tilde{\psi}_{98} = (0.8, 0.9)$	0.80	0.18	0.02	0.002	0.474	$\vartheta = 0.11$	\mathcal{M}_1
$\tilde{\psi}_{99} = (0.8, 1.0)$	0.80	0.20	0.00	0.011	0.555	$\vartheta = 0.10$	\mathcal{M}_1
$\tilde{\psi}_{100} = (0.9, 0.0)$	0.90	0.00	0.10	0.325	0.368	$\vartheta = 0.10$	\mathcal{M}_1
$\tilde{\psi}_{100} = (0.9, 0.1)$	0.90	0.01	0.09	0.263	0.384	$\vartheta = 0.10$	\mathcal{M}_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$\tilde{\psi}_{110} = (0.9, 1.0)$	0.90	0.00	0.10	0.003	0.623	$\vartheta = 0.05$	\mathcal{M}_1
$\tilde{\psi}_{111} = (1.0, 0.0)$	1.00	0.00	0.00	0	0.693	$\vartheta = 0.00$	\mathcal{M}_1

3.3.3.3 Different representation, different prior beliefs

Expanding on these observations, we suspect that a scientist would subjectively set different prior beliefs depending on whether she is confronted with the matrix of Table 3.1 or with the matrix of Table 3.2. In particular, when confronted with the matrix of Table 3.1 the scientist might subjectively set $\lambda(\psi_{61}) = \lambda(\psi_{63}) = 0.1$ and $\lambda(\psi_{62}) = 0.7$ meaning that she is quite sure, that the participant will generate the response x_1 with 80% chance, i.e., $P(p^*(x_1) = 0.80) = 0.9$. To cohere to this belief the scientist would simply set $\lambda(\tilde{\psi}_{89}) = \lambda(\tilde{\psi}_{99}) = 0.1$ and $\lambda(\tilde{\psi}_{94}) = 0.7$ and, subsequently, set the prior belief of all the “in-between” pmfs that generate x_1 with 80% to zero in the Table 3.2. We highly doubt that any scientist would be so specific in formulating her prior beliefs and, thus, doubt that a subjective assessment of the prior beliefs will work here.

As an alternative, we might think that we are noninformative if we give each

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

candidate true pmf the same prior probability. This means that we then give each candidate true pmf of Table 3.1 a prior probability of $\lambda(\psi_m) = 1/66 \approx 0.0152$. The pmfs that the participant will generate the response x_1 with 80% chance then get a total prior probability of $3/66 \approx 0.0455$. On the other hand, in Table 3.2 a uniform prior on $\lambda(\tilde{\psi}_m) = 1/111 \approx 0.009$ and the pmfs that the participant will generate the response x_1 with 80% chance then gets prior probability of $11/111 \approx 0.099$. Hence, a different set of candidate true pmfs will lead to a different assessment of prior beliefs. This lack of invariance depends on how many and which true candidate pmfs are chosen from \mathcal{P} in constructing the finite-dimensional matrices of Table 3.1 and Table 3.2.

3.3.3.4 Different representation, different prior model probabilities, thus, different Bayes factors

Applying the C&S belief propagation procedure to the matrix of Table 3.1 yields different allocations, thus, different Bayes factors then when we use the matrix of Table 3.2. For example, a scientist might believe that the true data generating pmf is close to $p(X | \psi_{18}) = [0.1, 0.6, 0.3]$ of Table 3.1, thus, chooses $\lambda(\psi_{18}) = 0.50$. This prior belief then gets allocated to the model instance $f_1(X | \vartheta = 0.6)$ of \mathcal{M}_1 . Similarly, we would expect that the scientist would also set $\lambda(\psi_{19}) \approx 0.50$ when confronted with Table 3.2, because the candidate pmf $p(X | \tilde{\psi}_{19}) = [0.10, 0.63, 0.27]$ in the second matrix does not differ that much from the pmf $p(X | \psi_{18})$ of the first matrix. However, according to the second matrix the prior pmf probability $\lambda(\tilde{\psi}_{19})$ is then allocated to the model instance $f_2(X | \alpha = 0.63)$ of \mathcal{M}_2 . In effect, a different representation leads to a different belief allocation, thus, different priors $\pi_i(\theta_i)$, $P(\mathcal{M}_i)$ and different posteriors $P(\hat{\mathcal{M}}_i | d)$ and, consequently, different Bayes factors. As such, our understanding of the C&S belief propagation procedure leads to an inadequate definition of Bayes factors, which depends on how we choose to represent \mathcal{P} .

3.3.3.5 Jeffreys's prior and the C&S procedure

The reason for this lack of invariance is due to an error incurred from (1) the parameterisation ξ itself, and (2) the discretisation of the parameter space. For example, the matrices depicted in Table 3.1 and Table 3.2 were derived from the parameterisations ξ and $\tilde{\xi}$, respectively, followed by a discretisation of the parameter space with a step size of 0.1 in each coordinate. The first point can be repaired, as Jeffreys (1946) showed that Fisher information can be used to neutralise the parameterisation error. This solution is more commonly known as the Jeffreys's prior. In Ly et al. (2017c) we showed that the Jeffreys's prior on the parameters, say, $\psi = (b, c)$ in Ψ leads to a uniform prior on pmfs in \mathcal{P} . The second point however cannot be fixed.

To elaborate on this latter point, recall that the collection of all data generating pmfs \mathcal{P} is uncountably large, which means that the scientist's actual prior belief $\lambda(\psi)$ is a continuous quantity. By using a finite number M of candidate true data generating pmfs, the target continuous random variable $\lambda(\psi)$ is then approximated by a discretised version $\lambda(\psi_m)$. The corresponding discretisation

errors are comparable to the errors incurred when histograms are used to approximate a smooth density function. Moreover, because the actual belief about ψ is continuous, we have zero probability of having the true data generating process $p^*(X)$ being exactly equal to one of the finite number of candidate pmfs $p(X | \psi_m)$. As such, we cannot construct the actual belief $\lambda(\psi)$ from point masses. Note that this continuity issue was already alluded to in Section 3.3.3.3 as one would expect that if the pmfs indexed by $\tilde{\psi}_{89}, \tilde{\psi}_{94}, \tilde{\psi}_{99}$ in Table 3.2 are assigned some prior mass, the pmfs in between would also receive some prior mass. The implication is that the C&S procedure might only work if we use a “matrix” with an uncountable number of rows.

Furthermore, the discretisation leads to another type of approximation error that we refer to as geometric approximation error due to the chosen divergence measure D . This error was alluded to in Section 3.3.3.4, where a small change in the candidate true data generating pmf $p(X | \psi_{18}) = [0.1, 0.6, 0.3]$ to $p(X | \tilde{\psi}_{19}) = [0.10, 0.63, 0.27]$ leads to a completely different allocation of the prior belief; from a model instance of \mathcal{M}_1 to one of \mathcal{M}_2 . The geometrical interpretation stems from the fact that KL-divergence can be thought of as a generalisation of the Fisher information metric.¹¹ Moreover, it follows directly from the geometric interpretation that the C&S belief propagation procedure favours the more complex model, as it will attract a larger number of candidate data generating pmfs indexed by ψ_m , see Ly et al. (2017c). This a priori boosting of the more complex model is at odds with the simplicity postulate that seems to be central in the foundations of the C&S procedure, see Shiffrin et al. (2016).

The fact that we cannot construct the actual belief $\lambda(\psi)$ from point masses is at odds with the C&S idea that $P(\mathcal{M}_i)$ is the sum of its parts. This bottom-up view is what caused Shiffrin et al. (2016) to avoid overlapping models; when \mathcal{M}_1 and \mathcal{M}_2 share a pmf and the shared instance receives some prior mass, this prior mass will be accounted for twice. As a result, the prior model probabilities will then exceed one, i.e., $P(\mathcal{M}_1) + P(\mathcal{M}_2) > 1$. To deal with overlapping models Shiffrin et al. (2016) suggested to remove the common pmfs from the larger model. This idea is elaborated on with a toy example where \mathcal{M}_3 is a binomial model with the chance of success θ fixed at $\theta = 0.5$ and where \mathcal{M}_4 represents the binomial model in which θ is free to vary between zero and one. They then reformulate \mathcal{M}_4 as the binomial model $\tilde{\mathcal{M}}_4$ in which θ is free to vary between $(0, 0.49)$ and $(0.51, 1)$. This replacement of \mathcal{M}_4 by $\tilde{\mathcal{M}}_4$ leads to another approximation error. One solution would be to allow $\tilde{\mathcal{M}}_4$ to converge to \mathcal{M}_4 by allowing θ to be in $(0, 0.5 - \epsilon) \cup (0.5 + \epsilon, 1)$. This construction however depends on how ϵ goes to zero and induces the Borel-Kolmogorov paradox (e.g., Lindley, 1997; Wetzel et al., 2010a). This paradox is another indication of how the C&S belief propagation scheme depends on how we as scientists represent the problem in terms of the chosen parameterisation and, subsequently, discretise the parameter space.

In other words, we believe that the lack of invariance is inescapable when the C&S approach is operationalised with a finite-dimensional matrix leading to an over-simplification of the problem resulting in a representation that is not on par

¹¹The KL-divergence is not a metric in the formal sense, only its infinitesimal version can be related to the Fisher information as a metric, i.e., Jeffreys's prior.

3. AN EVALUATION OF ALTERNATIVE METHODS FOR TESTING HYPOTHESES, FROM THE PERSPECTIVE OF HAROLD JEFFREYS

with the sophisticated ideas behind the C&S approach.

3.3.4 Conclusion

Based on the different strategies used to set priors $\pi_i(\theta_i)$ within the models \mathcal{M}_i , we conclude that the C&S belief propagation procedure answers a different question than a traditional Bayes factor. We believe that C&S are mostly concerned with how a scientist's subjective knowledge of the true data generating $p^*(X)$ is permeated in the models \mathcal{M}_1 and \mathcal{M}_2 . Hence, C&S focus on checking whether the models \mathcal{M}_1 and \mathcal{M}_2 give a good representation of expert knowledge.

As such, we think that the C&S approach can be valuable at the preliminary stage of model building. In particular, by considering all possible data generating pmfs for the random variable X , the C&S procedure forces the statistician to focus on building a model that is relevant for the problem at hand, rather than being restricted by the standard models. We would like to emphasise that our remarks are not aimed at the aspiration of C&S to construct good models that mimic nature well.

Our major concern deals with the finite-dimensional representation that C&S use to operationalise their procedure and the recommendations to set $\lambda(\psi)$ subjectively. The idea to consider the full model \mathcal{P} is to account for misspecification; as a result, however, the subjective assessment of prior beliefs is nigh impossible. Note that the subjective belief $\lambda(\psi)$ is necessarily a continuous random variable, because the full model \mathcal{P} contains an uncountable number of candidate true pmfs $p(X | \psi)$. To make their procedure viable, C&S oversimplify the problem with a finite-dimensional matrix yielding approximation errors that cannot be ignored.

The problem worsens when X is also continuous. In that case, the full model should then be represented by a "matrix" with an uncountable number of rows and columns. Moreover, this full model is far too complex, as it does not even allow for consistent inference (Dvoretzky et al., 1956). This is why regularisation methods were invented and alternative models were proposed that grow with the number of samples (e.g., Bickel, 2006). The goal set by C&S to compare models in a totally unrestricted setting is ambitious and an active area of research that is progressing slowly, see Borgwardt and Ghahramani (2009), Ghosal et al. (2008), Holmes et al. (2015), Labadi et al. (2014), Salomond (2013) and Salomond (2014) for some recent results.

For estimation problems, one solution would be to forgo the finite matrix representation and consider the prior on \mathcal{P} as a continuous random variable instead. As a replacement of the subjective assessment, we then recommend Jeffreys's prior as it is uniform on \mathcal{P} when X has a finite number of outcomes W . A Jeffreys prior for the full model \mathcal{P} is viable when $W < \infty$, as the distribution of X is then at most a multinomial distribution with W categories. When X is continuous the Jeffreys prior can then be extended by a method described in Ghosal et al. (1997), which has been used successfully to justify Bayesian nonparametric estimation methods, see also Ghosal et al. (2000) and Kleijn and Zhao (2017). However, this replacement of the discretised $\lambda(\psi_m)$ by a continuous version $\lambda(\psi)$ is at odds with the philosophy that the prior on the whole, $P(\mathcal{M}_i)$, is a sum of its parts $\pi_i(\theta_i)$ as the individual model instances then necessarily receive zero mass. Furthermore,

we do not know how to translate a continuous $\lambda(\psi)$ on all pmfs \mathcal{P} to the model instances π_i of \mathcal{M}_i without an explicitly defined relationship between the true data generating $p^*(X)$ and the model instances of \mathcal{M}_i . In effect, we doubt that the C&S procedure extends traditional Bayes factors and that it is capable of yielding a Jeffreys's Bayes factors that formalises inductive reasoning and the logic of proof by contradiction. The reason for this doubt is due to the fact that C&S do not focus on the two models under test, instead, they embed these two models within a larger encompassing model as Robert did, see Section 3.2.

In conclusion, we believe that a Jeffreys's Bayes factor remains the preferred method of inference, because a Jeffreys's Bayes factor does not depend on how the full model \mathcal{P} is represented and discretised. Thus, it does not suffer from the lack of invariance as discussed above. Furthermore, a Jeffreys's Bayes factor does not require a subjectively elicitation of prior beliefs. Note that the Bayes factor focuses on comparing the models \mathcal{M}_1 and \mathcal{M}_2 , no reference is made to any true data generating process $p^*(X)$. Jeffreys was mostly concerned with quantifying the (relative) evidence provided by the observations for either model. The Bayes factor is not concerned with the true data generating process $p^*(X)$ and it does not aspire to do so. Both \mathcal{M}_1 and \mathcal{M}_2 could be poor descriptions of the true data generating pmf $p^*(X)$, but fortunately it has been shown that the model selected with a Bayes factor is the model closest to the true $p^*(X)$ in terms of KL-divergence (e.g., Dass and Lee, 2004). Hence, the model that is preferred by the Bayes factor will be able to generalise better to yet unseen data –a guarantee that aligns with the spirit of the C&S approach.

3.4 Conclusion

We would like to thank the authors of both comments for their stimulating remarks and for their creative alternatives and extensions to Jeffreys's Bayes factors. We hope that this discussion has resulted in a renewed appreciation for Harold Jeffreys's foundational contributions to model selection and hypothesis testing, and we look forward to future developments in this exciting and important area of research.

Part II

Bayes Factors for Common Designs

Chapter 4

Bayesian Inference for Kendall's Rank Correlation Coefficient

Abstract

This chapter outlines a Bayesian methodology to estimate and test the Kendall rank correlation coefficient τ . The nonparametric nature of rank data implies the absence of a generative model and the lack of an explicit likelihood function. These challenges can be overcome by modelling test statistics rather than data (Johnson, 2005). We also introduce a method for obtaining a default prior distribution. The combined result is an inferential methodology that yields a posterior distribution for Kendall's τ .

Keywords: Bayes factor, nonparametric inference.

4.1 Introduction

One of the most widely used nonparametric tests of dependence between two variables is the rank correlation known as Kendall's τ (Kendall, 1938). Compared to Pearson's ρ , Kendall's τ is robust to outliers and violations of normality (Kendall and Gibbons, 1990). Moreover, Kendall's τ expresses dependence in terms of monotonicity instead of linearity and is therefore invariant under rank-preserving transformations of the measurement scale (Kruskal, 1958; Wasserman, 2006). As expressed by Harold Jeffreys (1961, p. 231): “(...) it seems to me that the chief merit of the method of ranks is that it eliminates departure from linearity, and with it a large part of the uncertainty arising from the fact that we do not know any form of the law connecting X and Y ”. Here we apply the Bayesian inferential paradigm to Kendall's τ . Specifically, we define a default prior distribution

This chapter is published online as: van Doorn, J.B., Ly, A., Marsman, A., & Wagenmakers, E.-J. (2017). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*. doi: <http://dx.doi.org/10.1080/00031305.2016.1264998>

on Kendall's τ , obtain the associated posterior distribution, and use the Savage-Dickey density ratio to obtain a Bayes factor hypothesis test (Dickey and Lientz, 1970; Jeffreys, 1961; Kass and Raftery, 1995).

4.1.1 Kendall's τ

Let $X^n = (X_1, \dots, X_n)$ and $Y^n = (Y_1, \dots, Y_n)$ be two random vectors each containing measurements of the same n units. For example, consider the association between French and maths grades in a class of $n = 3$ children: Tina, Bob, and Jim; let $x^n = (8, 7, 5)$ be their observed grades for a French exam and $y^n = (9, 6, 7)$ be their realised grades for a maths exam. For $1 \leq i < j \leq n$, each pair (i, j) is defined to be a pair of differences $(x_i - x_j)$ and $(y_i - y_j)$. A pair is considered to be concordant if $(x_i - x_j)$ and $(y_i - y_j)$ share the same sign, and discordant when they do not. In our data example, Tina has higher grades on both exams than Bob, which means that Tina and Bob are a concordant pair. Conversely, Bob has a higher score for French, but a lower score for maths than Jim, which means Bob and Jim are a discordant pair. The observed value of Kendall's τ , denoted τ_{obs} , is defined as the difference between the number of concordant and discordant pairs, expressed as proportion of the total number of pairs:

$$\tau_{\text{obs}} = \frac{\sum_{1 \leq i < j \leq n}^n Q((x_i, y_i), (x_j, y_j))}{n(n-1)/2}, \quad (4.1.1)$$

where the denominator represents the total number of pairs and Q is the concordance indicator function:

$$Q((x_i, y_i)(x_j, y_j)) = \begin{cases} -1 & \text{if } (x_i - x_j)(y_i - y_j) < 0, \\ +1 & \text{if } (x_i - x_j)(y_i - y_j) > 0. \end{cases} \quad (4.1.2)$$

Table 4.1 illustrates the calculation for our small data example. Applying Eq. (4.1.1) gives $\tau_{\text{obs}} = 1/3$, an indication of a positive correlation between French and maths grades.

i	j	$x_i - x_j$	$y_i - y_j$	Q
1	2	8-7	9-6	1
1	3	8-5	9-7	1
2	3	7-5	6-7	-1

Table 4.1: The pairs (i, j) for $1 \leq i < j \leq n$ and the concordance indicator function Q for the data example where $x^n = (8, 7, 5)$ and $y^n = (9, 6, 7)$.

When $\tau_{\text{obs}} = 1$, all pairs of observations are concordant, and when $\tau_{\text{obs}} = -1$, all pairs are discordant. Kruskal (1958) provides the following interpretation of Kendall's τ : in the case of $n = 2$, suppose we bet that $y_1 < y_2$ whenever $x_1 < x_2$, and that $y_1 > y_2$ whenever $x_1 > x_2$; winning \$1 after a correct prediction and losing \$1 after an incorrect prediction, the expected outcome of the bet equals τ . Furthermore, Griffin (1958) has illustrated that when the ordered rank-converted

values of X are placed above the rank-converted values of Y and lines are drawn between the same numbers, Kendall's τ_{obs} is given by the formula: $1 - \frac{4z}{n(n-1)}$, where z is the number of line intersections; see Fig. 4.1 for an illustration of this method using our example data of French and maths grades. These tools allows us to straightforwardly and intuitively calculate and interpret Kendall's τ .

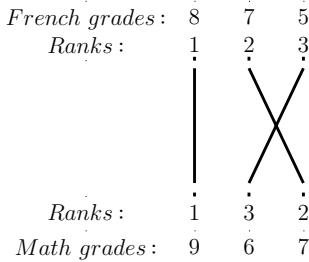


Figure 4.1: A visual interpretation of Kendall's τ_{obs} through the formula: $1 - \frac{4z}{n(n-1)}$, where z is the number of intersections of the lines. In this case, $n = 3$, $z = 1$, and $\tau_{\text{obs}} = 1/3$.

Despite these appealing properties and the overall popularity of Kendall's τ , a default Bayesian inferential paradigm is still lacking because the application of Bayesian inference to nonparametric data analysis is not trivial. The main challenge in obtaining posterior distributions and Bayes factors for nonparametric tests is that there is no generative model and no explicit likelihood function. In addition, Bayesian model specification requires the specification of a prior distribution, and this is especially important for Bayes factor hypothesis testing; however, for nonparametric tests it can be challenging to define a sensible default prior. Though recent developments have been made in two-sample nonparametric Bayesian hypothesis testing with Dirichlet process priors (Borgwardt and Ghahramani, 2009; Labadi et al., 2014) and Pólya tree priors (Chen and Hanson, 2014; Holmes et al., 2015), here we focus on a different approach, one that permits an intuitive and direct interpretation.

4.1.2 Modelling test statistics

In order to compute Bayes factors for Kendall's τ we start with the approach pioneered by Johnson (2005) and Yuan and Johnson (2008). These authors established bounds for Bayes factors based on the sampling distribution of the standardised value of τ , denoted by T^* , which will be formally defined in Section 4.2.1. Using the Pitman translation alternative, where a non-centrality parameter is used to distinguish between the null and alternative hypotheses (Randall and Wolfe,

1979), Johnson and colleagues specified the following hypotheses:

$$\mathcal{H}_0 : \theta = \theta_0, \quad (4.1.3)$$

$$\mathcal{H}_1 : \theta = \theta_0 + \frac{\Delta}{\sqrt{n}}, \quad (4.1.4)$$

where θ is the true underlying value of Kendall's τ , θ_0 is the value of Kendall's τ under the null hypothesis, and Δ serves as the non-centrality parameter which can be assigned a prior distribution. The limiting distribution of T^* under both hypotheses is normal distributed (Hotelling and Pabs, 1936; Noether, 1955; Chernoff and Savage, 1958), that is,

$$\mathcal{H}_0 : T^* \sim \mathcal{N}(0, 1) \quad (4.1.5)$$

$$\mathcal{H}_1 : T^* \sim \mathcal{N}(\frac{3\Delta}{2}, 1). \quad (4.1.6)$$

The prior on Δ is specified by Yuan and Johnson as

$$\Delta \sim \mathcal{N}(0, g), \quad (4.1.7)$$

where g is used to specify the expectation about the size of the departure from the null-value of Δ . This leads to the following Bayes factor:

$$BF_{01}(d) = \sqrt{1 + 9/4g} \exp\left(-\frac{gt^{*2}}{2g + 8/9}\right). \quad (4.1.8)$$

Next, Yuan and Johnson calculated an upper bound for $BF_{10}(d)$, thus, a lower bound on $BF_{01}(d)$, by maximising over the hyperparameter g .

4.1.3 Challenges

Although innovative and compelling, the approach advocated by Yuan and Johnson (2008) does have a number of non-Bayesian elements, most notably the data-dependent maximisation over the hyperparameter g that results in a data-dependent prior distribution. Moreover, the definition of \mathcal{H}_1 depends on n : as $n \rightarrow \infty$, \mathcal{H}_1 and \mathcal{H}_0 become indistinguishable and lead to an inconsistent inferential framework.

Our approach, motivated by the earlier work by Johnson and colleagues, sought to eliminate g not by maximisation but by a method we call "parametric yoking" (i.e., matching with a prior distribution for a parametric alternative). In addition, we redefined \mathcal{H}_1 such that its definition does not depend on sample size. As such, Δ becomes synonymous with the true underlying value of Kendall's τ when $\theta_0 = 0$.

4.2 Methods

4.2.1 Defining T^*

As mentioned above, Yuan and Johnson (2008) use the standardised version of τ , denoted T^* (Kendall, 1938) which is defined as

$$T^* = \frac{\sum_{1 \leq i < j \leq n}^n Q((X_i, Y_i), (X_j, Y_j))}{\sqrt{n(n-1)(2n+5)/18}}. \quad (4.2.1)$$

Here the numerator contains the concordance indicator function Q . Thus, T^* is not necessarily situated between the traditional bounds $[-1, 1]$ for a correlation; instead, T^* has a maximum of $\sqrt{\frac{9n(n-1)}{4n+10}}$ and a minimum of $-\sqrt{\frac{9n(n-1)}{4n+10}}$. This definition of T^* enables the asymptotic normal approximation to the sampling distribution of the test statistic (Kendall and Gibbons, 1990).

4.2.2 Prior distribution through parametric yoking

In order to derive a Bayes factor for τ we first determine a default prior for τ through what we term parametric yoking. In this procedure, a default prior distribution is constructed by comparison to a parametric alternative. In this case, a convenient parametric alternative is given by Pearson's correlation for bivariate normal data. Ly et al. (2016a) use a symmetric stretched beta prior distribution ($\alpha = \beta$) on the domain $(-1, 1)$, that is,

$$\pi(\rho) = \frac{2^{1-2\alpha}}{\mathcal{B}(\alpha, \alpha)} (1 - \rho^2)^{(\alpha-1)}, \quad \rho \in (-1, 1), \quad (4.2.2)$$

where \mathcal{B} is the beta function. For bivariate normal data, Kendall's τ is related to Pearson's ρ by Greiner's relation (Greiner, 1909; Kruskal, 1958):

$$\tau = \frac{2}{\pi} \arcsin(\rho). \quad (4.2.3)$$

We use this relationship to transform the beta prior in Eq. (4.2.2) on ρ to a prior on τ , which leads to

$$\pi(\tau) = \pi \frac{2^{-2\alpha}}{\mathcal{B}(\alpha, \alpha)} \cos\left(\frac{\pi\tau}{2}\right)^{(2\alpha-1)}, \quad \tau \in (-1, 1). \quad (4.2.4)$$

In the absence of strong prior beliefs, Jeffreys (1961) proposed a uniform distribution on ρ , that is, a stretched beta with $\alpha = \beta = 1$. This choice induces a non-uniform distribution on τ , i.e.,

$$\pi(\tau) = \frac{\pi}{4} \cos\left(\frac{\pi\tau}{2}\right). \quad (4.2.5)$$

In general, values of $\alpha > 1$ increase the prior mass near $\tau = 0$, whereas values of $\alpha < 1$ decrease the prior mass near $\tau = 0$. When the focus is on parameter estimation instead of hypothesis testing, we may follow Jeffreys (1961) and use a stretched beta prior on ρ with $\alpha = \beta = \frac{1}{2}$. As is easily confirmed by entering these values in Eq. (4.2.4), this choice induces a uniform prior distribution on Kendall's τ .¹ The parametric yoking framework can be extended to other prior distributions that exist for Pearson's ρ (e.g., the inverse Wishart distribution; Berger and Sun, 2008; Gelman et al., 2014), by transforming ρ with the inverse of the expression given in Eq. (4.2.3), namely,

$$\rho = \sin\left(\frac{\pi\tau}{2}\right). \quad (4.2.6)$$

¹ Additional examples and figures of the stretched beta prior, including cases where $\alpha \neq \beta$, are available online at <https://osf.io/b9qhj/>.

4.2.3 Posterior distribution and Bayes factor

Removing \sqrt{n} from the specification of \mathcal{H}_1 by substituting $\Delta\sqrt{n}$ for Δ , we get an (approximate) normal distribution for T^* under \mathcal{H}_1 with mean $\mu = \frac{3}{2}\Delta\sqrt{n}$ and standard deviation $\sigma = 1$, thus, the density of T^* at t^* is given by

$$f(t^* | \theta_0 + \Delta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}[t^* - \frac{3}{2}\Delta\sqrt{n}]^2\right). \quad (4.2.7)$$

Filling in the observed value for T^* and combining this normal likelihood function with the prior from Eq. (4.2.4) then yields a posterior distribution for Kendall's τ . Next, Bayes factors can be computed as

$$\text{BF}_{01}(d) = \frac{p(t^* | \theta_0)}{\int f(t^* | \theta_0 + \Delta) \pi(\Delta) d\Delta}, \quad (4.2.8)$$

which in the case of Kendall's τ translates to

$$\text{BF}_{01}(d) = \frac{\exp(-\frac{1}{2}t^{*2})}{\int_{-1}^1 \exp\left(-\frac{1}{2}[t^* - \frac{3}{2}\tau\sqrt{n}]^2\right) \pi \frac{2^{2\alpha}}{\mathcal{B}(\alpha, \alpha)} \cos(\frac{\pi\tau}{2})^{2\alpha-1} d\tau}. \quad (4.2.9)$$

4.2.4 Verifying the asymptotic normality of T^*

Our method relies on the asymptotic normality of T^* , a property established mathematically by Hoeffding (1948). For practical purposes, however, it is insightful to assess the extent to which this distributional assumption is appropriate for realistic sample sizes. By considering all possible permutations of the data, deriving the exact cumulative density of T^* , and comparing the densities to those of a standard normal distribution, Ferguson et al. (2000) concluded that the normal approximation holds under \mathcal{H}_0 when $n \geq 10$. But what if \mathcal{H}_0 is false?

Here we report a simulation study designed to assess the quality of the normal approximation to the sampling distribution of T^* when \mathcal{H}_1 is true. With the use of copulas, 100,000 synthetic data sets were created for each of several combinations of Kendall's τ and sample size n .² For each simulated data set, the Kolmogorov-Smirnov statistic was used to quantify the fit of the normal approximation to the sampling distribution of T^* .³ Fig. 4.2 shows the Kolmogorov-Smirnov statistic as a function of n , for various values of τ when data sets were generated from a bivariate normal distribution (i.e., the normal copula). Similar results were obtained using Frank, Clayton, and Gumbel copulas. As is the case under \mathcal{H}_0 (e.g., Ferguson et al., 2000; Kendall and Gibbons, 1990), the quality of the normal approximation increases exponentially with n . Furthermore, larger values of τ necessitate larger values of n to achieve the same quality of approximation.

The means of the normal distributions fit to the sampling distribution of T^* are situated at the point $\frac{3}{2}\Delta\sqrt{n}$. The data sets from this simulation can also be used to examine the variance of the normal approximation. Under \mathcal{H}_0 (i.e., $\tau = 0$),

²For more information on copulas see Nelsen (2006), Genest and Favre (2007), and Colonius (2016).

³R-code, plots, and further details are available online at <https://osf.io/b9qhj/>.

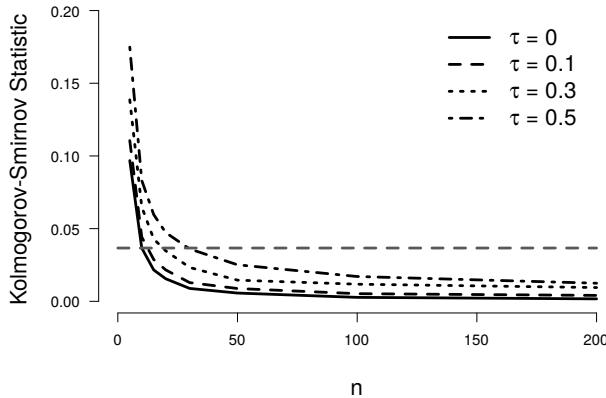


Figure 4.2: Quality of the normal approximation to the sampling distribution of T^* , as assessed by the Kolmogorov-Smirnov statistic. As n grows, the quality of the normal approximation increases exponentially. Larger values of τ necessitate larger values of n to achieve the same quality of approximation. The grey horizontal line corresponds to a Kolmogorov-Smirnov statistic of 0.038 (obtained when $\tau = 0$ and $n = 10$), for which Ferguson et al. (2000, p. 589) deemed the quality of the normal approximation to be “sufficiently precise for practical purposes”.

the variance of these normal distributions equals 1. As the population correlation grows (i.e., $|\tau| \rightarrow 1$), the number of permissible rank permutations decreases and so does the variance of T^* . The upper bound of the sampling variance of T^* is a function of the population value for τ (Kendall and Gibbons, 1990):

$$\sigma_{T^*}^2 \leq \frac{2.5n(1 - \tau^2)}{2n + 5}. \quad (4.2.10)$$

As shown in the online appendix, our simulation results provide specific values for the variance which respect this upper bound. This result has ramifications for the Bayes factor. As the test statistic moves away from 0, the variance falls below 1, and the posterior distribution will be more peaked on the value of the test statistic than when the variance is assumed to equal 1. This results in increased evidence in favour of \mathcal{H}_1 , so that our proposed procedure is somewhat conservative. However, for $n \geq 20$, the changes in variance will only surface in cases where there already exists substantial evidence for \mathcal{H}_1 (i.e., $\text{BF}_{10}(d) \geq 10$).

4.3 Results

4.3.1 Bayes factor behaviour

Now that we have determined a default prior for τ and combined it with the specified Gaussian likelihood function, computation of the posterior distribution and the Bayes factor becomes feasible. For an uninformative prior on τ (i.e.,

$\alpha = \beta = 1$), Fig. 4.3 illustrates $\text{BF}_{10}(d)$ as a function of n , for three values of τ_{obs} . The lines for $\tau_{\text{obs}} = 0.2$ and $\tau_{\text{obs}} = 0.3$ show that $\text{BF}_{10}(d)$ for a true \mathcal{H}_1 increases exponentially with n , as is generally the case. For $\tau_{\text{obs}} = 0$, the Bayes factor decreases as n increases.

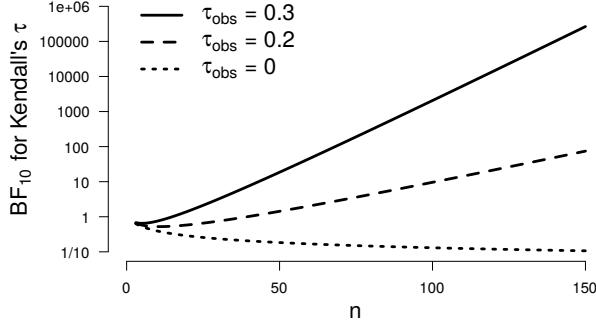


Figure 4.3: Relation between $\text{BF}_{10}(d)$ and sample size ($3 \leq n \leq 150$) for three values of Kendall's τ .

4.3.2 Comparison to Pearson's ρ

In order to put the result in perspective, the Bayes factors for Kendall's tau (i.e., $\text{BF}_{10}^\tau(d)$) can be compared to those for Pearson's ρ (i.e., $\text{BF}_{10}^\rho(d)$). The Bayes factors for Pearson's ρ are based on Jeffreys (1961), see also Ly et al., 2016a, who used the uniform prior on ρ . Fig. 4.4 shows that the relationship between $\text{BF}_{10}^\tau(d)$ and $\text{BF}_{10}^\rho(d)$ for normal data is approximately linear as a function of sample size. In addition, and as one would expect due to the loss of information when continuous values are converted to coarser ranks, $\text{BF}_{10}^\tau(d) < \text{BF}_{10}^\rho(d)$ in the case of evidence in favour of \mathcal{H}_1 (left panel of Fig. 4.4). When evidence is in favour of \mathcal{H}_0 , i.e. $\tau = 0$, $\text{BF}_{10}^\tau(d)$ and $\text{BF}_{10}^\rho(d)$ perform similarly (right panel of Fig. 4.4).

4.3.3 Real data example

Willerman et al. (1991) set out to uncover the relation between brain size and IQ. Across 20 participants, the authors observed a Pearson's correlation coefficient of $r = 0.51$ between IQ and brain size, measured in MRI count of grey matter pixels. The data are presented in the top left panel of Fig. 4.5. Bayes factor hypothesis testing of Pearson's ρ yields $\text{BF}_{10}^\rho(d) = 5.16$, which is illustrated in the middle left panel. This means that the data are 5.16 times as likely to occur under \mathcal{H}_1 than under \mathcal{H}_0 . When applying a log-transformation on the MRI counts (after subtracting the minimum value minus 1), however, the linear relation between IQ and brain size is less strong. The top right panel of Fig. 4.5 presents the effect of

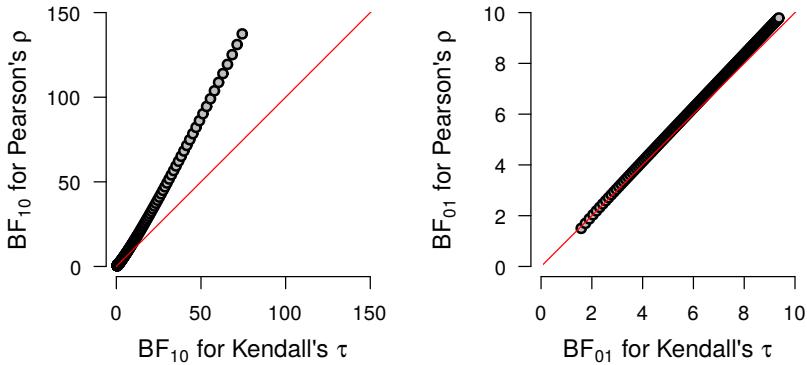


Figure 4.4: Relation between the Bayes factors for Pearson's ρ and Kendall's $\tau = 0.2$ (left) and Kendall's $\tau = 0$ (right) as a function of sample size (i.e., $3 \leq n \leq 150$). The data are normally distributed. Note that the left panel shows $\text{BF}_{10}(d)$ and the right panel shows $\text{BF}_{01}(d)$. The diagonal line indicates equivalence.

this monotonic transformation on the data. The middle right panel illustrates how the transformation decreases $\text{BF}_{10}^{\rho}(d)$ to 1.28. The bottom left panel presents our Bayesian analysis on Kendall's τ , which yields a $\text{BF}_{10}^{\tau}(d)$ of 2.17. Furthermore, the bottom right panel shows the same analysis on the transformed data, illustrating the invariance of Kendall's τ against monotonic transformations: the inference remains unchanged, which highlights one of Kendall's τ most appealing features.

4.4 Concluding comments

We outlined a nonparametric Bayesian framework for inference about Kendall's τ based on modelling test statistics and by assigning a prior by means of a parametric yoking procedure. The framework produces a posterior distribution for Kendall's τ , and –via the Savage-Dickey density ratio test– also yields a

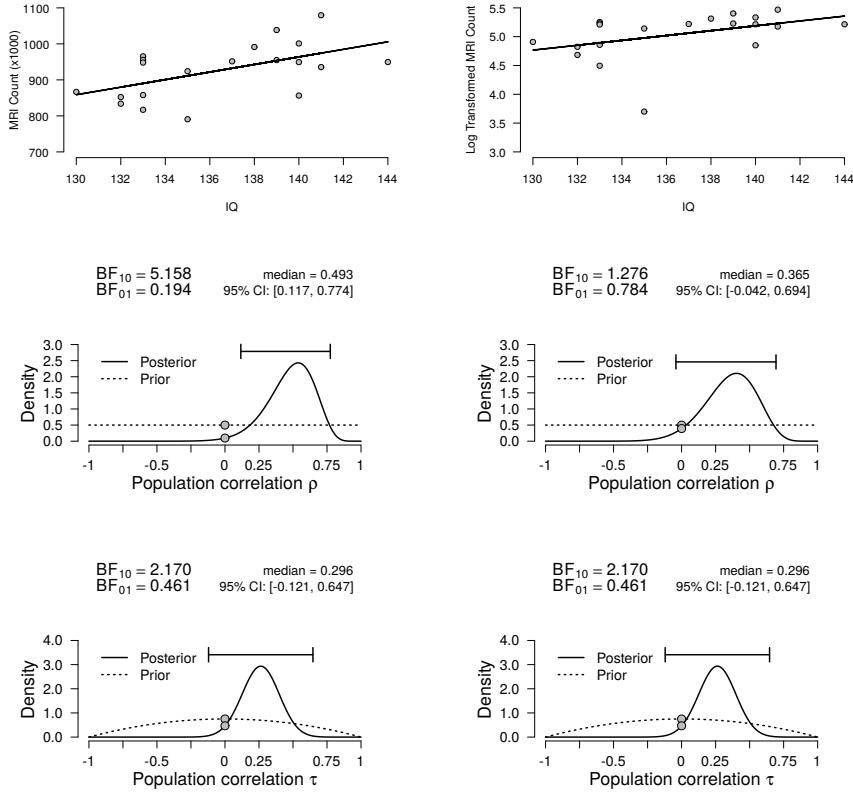


Figure 4.5: Bayesian inference for Kendall's τ illustrated with data on IQ and brain size (Willerman et al. 1991). The left column presents the relation between brain size and IQ, analysed using Pearson's ρ (middle panel) and Kendall's τ (bottom panel). The right column presents the results after a log transformation of brain size. Note that the transformation affects inference for Pearson's ρ , but does not affect inference for Kendall's τ .

Bayes factor that quantifies the evidence for the absence of a correlation.

Our general procedure (i.e., modelling test statistics and assigning a prior through parametric yoking) is relatively general and may be used to facilitate Bayesian inference for other nonparametric tests as well. For instance, Serfling (1980) offers a range of test statistics with asymptotic normality to which our framework may be expanded, whereas Johnson (2005) has explored the modelling of test statistics that have non-Gaussian limiting distributions.

Chapter 5

Informed Bayesian t -Tests

Abstract

Across the empirical sciences, few statistical procedures rival the popularity of the frequentist t -test. In contrast, the Bayesian versions of the t -test have languished in obscurity. In recent years, however, the theoretical and practical advantages of the Bayesian t -test have become increasingly apparent. Developed by Harold Jeffreys, the default Bayesian t -test assigns a zero-centred Cauchy prior distribution to the effect size parameter under the alternative hypothesis. This specification is relatively restrictive, and in scientific applications the researcher may possess expert knowledge that allows the formulation of an informed prior distribution that is centred away from zero. In order to facilitate a more informative Bayesian analysis of the comparison between two means we extend Harold Jeffreys's t -tests. This extension allows researchers to assign a shifted and scaled t prior distribution to the effect size parameter. We present expressions for the marginal posterior distribution of the effect size parameter and for the Bayes factor test; three real-data examples highlight the practical value of the approach.

Keywords: Bayes factor, informed hypothesis test, posterior distribution.

5.1 Introduction

The t -test is designed to assess whether two means differ from one another. The question is fundamental, and consequently the t -test has grown to be an inferential workhorse of the empirical sciences. The popularity of the t -test is underscored by considering the p -values published in eight major psychology journals from 1985 until 2013 (Nuijten et al., 2016); out of a total of 258,105 p -values, 26% tested the significance of a t statistic. For comparison, 4% of those p -values tested an r

This chapter is submitted for publication and also available as arXiv preprint:1705.01064 as: Gronau, Q.F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian t -tests.

statistic, 4% a z statistic, 9% a χ^2 statistic, and 57% an F statistic. Similarly, Wetzels et al. (2011) found 855 t -tests reported in 252 psychology articles, for an average of about 3.4 t -tests per article.

The popularity of the t -test has always concerned its classical or frequentist version. This article concerns the Bayesian version originally developed by Jeffreys (1948). Jeffreys's Bayesian t -test quantifies the predictive adequacy of two competing hypotheses: the null hypothesis \mathcal{H}_0 which states that the effect size δ is equal to zero (i.e., $\mathcal{H}_0 : \delta = 0$) versus the alternative hypothesis \mathcal{H}_1 which allows δ to vary freely. Jeffreys's test uses a zero-centred Cauchy prior distribution on δ with scale parameter $\gamma = 1$, i.e., $\delta \sim \text{Cauchy}(0, 1)$.

Jeffreys's Bayesian t -tests come with the well-known advantages of Bayesian inference: the ability to assess the relative plausibility of different values for δ by means of a posterior distribution; the ability to quantify evidence, both in favour of the null hypothesis and in favour of the alternative hypothesis; the ability to monitor the evidential flow as more observations become available; the ability to stop data collection whenever the evidence is compelling (Berger and Berry, 1988; Rouder, 2014; Savage, 1961); and the ability to provide composite estimates and predictions that take into account model uncertainty (Haldane, 1932; Hoeting et al., 1999). The fact that Jeffreys's t -tests have remained relatively unpopular among practitioners can be explained by a combination of factors: Jeffreys's writing is difficult to parse, Bayesian inference is generally excluded from the standard statistical curriculum offered in the empirical sciences, and the Bayesian t -tests were not available in mainstream statistical software packages. Recently, Rouder et al. (2009) have brought the Bayesian t -test into the limelight, with further impetus from the increased need for replication research to quantify evidence in favour of the null hypothesis (Marsman et al., 2017), and the fact that the test was subsequently implemented in user-friendly software packages (i.e., JASP Team, 2017; Morey and Rouder, 2015).

Jeffreys's Bayesian t -tests were meant to be “objective” in the sense of being applicable across a large range of different fields and phenomena. In Jeffreys's vision, possibly inspired by his experience as a physicist, the null hypothesis \mathcal{H}_0 represents an invariance or general law (i.e., $\delta = 0$), whereas the alternative hypothesis \mathcal{H}_1 relaxes the restriction imposed by \mathcal{H}_0 and reflects the intuition that the effect is likely to be small. Moreover, by centring the Cauchy prior on zero, Jeffreys avoided the complexity and context-dependency of having to specify another parameter (i.e., the most likely value of δ under \mathcal{H}_1). Nevertheless, a more informed specification of the prior distribution on δ , when possible, may increase the informativeness of the inference.

Here we present an extension of Jeffreys's default t -test that allows researchers to incorporate expert knowledge into the prior specification of the effect size parameter δ . Specifically, we consider two families of prior distributions for δ : the family of shifted and scaled t distributions (which includes Jeffreys's Cauchy prior as a special case), and the family of shifted and scaled normal distributions. For both families we derive the marginal posterior distribution of δ and the Bayes factor. For the normal family the solutions are completely analytic; for the t family the solutions contain a one-dimensional integral that can easily be evaluated numerically.

5.2 Jeffreys's default Bayes factor

Jeffreys developed Bayesian hypothesis tests for scenarios where the null hypothesis \mathcal{H}_0 represents a general law and the competing alternative hypothesis \mathcal{H}_1 relaxes the restriction imposed by \mathcal{H}_0 . For instance, in the one-sample t -test setting, \mathcal{H}_0 states that the population mean is equal to zero whereas the alternative hypothesis allows for a non-zero population mean.

To quantify the evidence that the data provide for or against a general law, Jeffreys (1961) developed a formal system of statistical inference with the following key equation (Wrinch and Jeffreys, 1921, p. 387):

$$\underbrace{\frac{P(\mathcal{H}_1 | d)}{P(\mathcal{H}_0 | d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d | \mathcal{H}_1)}{p(d | \mathcal{H}_0)}}_{\text{BF}_{10}(d)(d)} \underbrace{\frac{P(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \quad (5.2.1)$$

In his work, Jeffreys focused on the *Bayes factor* (Etz and Wagenmakers, 2017; Kass and Raftery, 1995; Ly et al., 2016a; Robert et al., 2009). The Bayes factor has an intuitive interpretation: when $\text{BF}_{10}(d) = 10$ this indicates that the data are 10 times more likely under \mathcal{H}_1 than under \mathcal{H}_0 ; when $\text{BF}_{10}(d) = .2$ this indicates that the data are 5 times more likely under \mathcal{H}_0 than under \mathcal{H}_1 .

Let \mathcal{H}_0 be specified by a series of nuisance parameters ζ and, importantly, a parameter of interest that is fixed at a single value of particular interest, $\theta = \theta_0$. The alternative hypothesis \mathcal{H}_1 is specified using similar nuisance parameters ζ , but in addition \mathcal{H}_1 releases the restriction on θ . In order to obtain the Bayes factor, the model parameters are integrated out

$$\text{BF}_{10}(d) = \frac{\int_{\Theta} \int_Z f(d | \theta, \zeta, \mathcal{H}_1) \pi(\theta, \zeta | \mathcal{H}_1) d\zeta d\theta}{\int_Z f(d | \theta = \theta_0, \zeta, \mathcal{H}_0) \pi(\zeta | \mathcal{H}_0) d\zeta}, \quad (5.2.2)$$

showing that the Bayes factor can be regarded as the ratio of two weighted averages where the weights correspond to the prior distribution for the parameters. Consequently, the choice of the prior distribution is crucial for the development of a Bayes factor hypothesis test.

Jeffreys's procedure for assigning prior distributions to the parameters exploits the fact that \mathcal{H}_0 is nested within \mathcal{H}_1 (obtained by setting $\theta = \theta_0$). For the nuisance parameters, Jeffreys used his famous parameterisation invariant “Jeffreys's” prior given by $\pi(\zeta) \propto \sqrt{\det I(\zeta)}$ where $I(\zeta)$ denotes the Fisher information matrix. For the test-relevant parameter θ , Jeffreys based his prior choice on two desiderata. The first desideratum, *predictive matching*, states that the Bayes factor should be perfectly indifferent (i.e., $\text{BF}_{10}(d) = 1$) in case the data are completely uninformative. The second desideratum, *information consistency*, states that the Bayes factor should provide infinite support for one of the two hypotheses in case the data are overwhelmingly informative (Bayarri et al., 2012; Jeffreys, 1942). In case of the t -test, these two desiderata led Jeffreys to a Cauchy prior distribution for the test-relevant effect size parameter (Ly et al., 2016a).

Jeffreys's development of the Bayes factor test was mainly concerned with situations where there is little prior knowledge available. However, it seems that he was not completely opposed to including existing expert prior knowledge into the

specification of the test-relevant prior distribution in case it is available (Jeffreys, 1961, p. 252):

“In any of these cases it would be perfectly possible to give a form of $f(a)$ [i.e., prior distribution for the test-relevant parameter, in our notation $\pi(\theta)$] that would express the previous information satisfactorily, and consideration of the general argument of 5.0 will show that it would lead to common-sense results, but they would differ in scale. As we are aiming chiefly at a theory that can be used in the early stages of a subject, we shall not at present consider the last type of case”

As Jeffreys alludes to, there may be situations where researchers have strong prior knowledge about the subject area. In this case, a procedure that admits a prior distribution for the test-relevant parameter which better reflects the prior state of knowledge might yield more informative tests of the researchers’ hypotheses than tests based on the default prior.

Note that Jeffreys’s default t -test assigns the crucial effect size parameter δ a zero-centred Cauchy distribution under \mathcal{H}_1 ; hence, under \mathcal{H}_1 the most likely effect size value is zero, a proposition that many researchers find problematic. Instead, these researchers may wish to centre their prior distribution on the non-zero value for δ that reflects their knowledge-based expectation about the effects at hand. Below we derive the posterior distributions and Bayes factors for two families of prior distributions on δ that allow changes in shift and scale. We first consider the one-sample and paired samples t -test and then consider the independent samples t -test.

5.3 One-sample and paired samples t -test

5.3.1 Model

In a one-sample t -test, n observations are assumed to be drawn from a normal distribution, that is, $y_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, 2, \dots, n$. Note that the paired samples t -test is a special case of this formalisation, since a paired samples t -test is equivalent to a one-sample t -test on the difference scores. We follow Jeffreys and reparameterise the model in terms of the dimensionless effect size $\delta = \mu/\sigma$ (Jeffreys, 1961; Ly et al., 2016a; Rouder et al., 2009; Wetzels et al., 2009). The null hypothesis states that the effect size is zero (i.e., $\mathcal{H}_0 : \delta = 0$) whereas the alternative hypothesis states that the effect size is non-zero (i.e., $\mathcal{H}_1 : \delta \neq 0$).

5.3.2 Prior distributions

Next we need to specify prior distributions for the parameters. For the parameter common to both models (i.e., σ^2), we use $\pi(\sigma^2) \propto \sigma^{-2}$, in line with Jeffreys’s approach. For the test-relevant effect size δ , we depart from Jeffreys’s default choice and consider an informed shifted and scaled t prior distribution, that is, $\delta \sim t(\mu_\delta, \gamma, \kappa)$, where μ_δ corresponds to the location parameter, γ to the scale parameter, and κ to the degrees of freedom.

Note that this specification includes Jeffreys's standard Cauchy prior distribution as a special case (i.e., when $\kappa = 1$). Additionally, we present the results for an informed normal prior of the form $\delta \sim \mathcal{N}(\mu_\delta, g)$ where μ_δ corresponds to the mean and g to the variance of the normal distribution.

5.3.3 Posterior distribution for effect size δ

After conducting a Bayesian t -test, the two quantities of primary interest are the posterior distribution for effect size δ under \mathcal{H}_1 and the Bayes factor (presented in the next section). For enhanced readability we suppress the conditioning on \mathcal{H}_1 and write the marginal posterior distribution for δ under \mathcal{H}_1 as

$$\pi(\delta | d) = \frac{\int_0^\infty f(d | \delta, \sigma^2) \pi(\sigma^2) d\sigma^2 \pi(\delta)}{p(d)}, \quad (5.3.1)$$

where d refers to the data. When we use a t prior of the form $t(\mu_\delta, \gamma, \kappa)$, we obtain

$$\pi(\delta | d) = \frac{\exp(-\frac{n}{2}\delta^2)(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}} [C + D] \frac{1}{\gamma} \frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi\kappa}\Gamma(\frac{\kappa}{2})} \left[1 + \frac{(\frac{\delta-\mu_\delta}{\gamma})^2}{\kappa}\right]^{-\frac{\kappa+1}{2}}}{\int_0^\infty (1 + ng)^{-\frac{1}{2}} \exp\left(-\frac{n\mu_\delta^2}{2(1+ng)}\right) \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} [A + B]\pi(g)dg}, \quad (5.3.2)$$

where $\pi(g)$ denotes an inverse-gamma distribution of the form $IG(\frac{\kappa}{2}, \gamma^2 \frac{\kappa}{2})$ given by

$$\pi(g) = \frac{(\gamma^2 \frac{\kappa}{2})^{\frac{\kappa}{2}}}{\Gamma(\frac{\kappa}{2})} g^{-\frac{\kappa}{2}-1} \exp(-\frac{\gamma^2 \kappa}{2g}), \quad (5.3.3)$$

and

$$A = \Gamma(\frac{\nu+1}{2}) {}_1F_1\left(\frac{\nu+1}{2}; \frac{1}{2}; \frac{n\mu_\delta^2 t^2}{2(1+ng)[(1+ng)\nu+t^2]}\right), \quad (5.3.4)$$

$$B = \frac{\sqrt{n}\mu_\delta t}{\sqrt{\frac{1}{2}(1+ng)[(1+ng)\nu+t^2]}} \Gamma(\frac{\nu+2}{2}) \quad (5.3.5)$$

$$\times {}_1F_1\left(\frac{\nu+2}{2}; \frac{3}{2}; \frac{n\mu_\delta^2 t^2}{2(1+ng)[(1+ng)\nu+t^2]}\right), \quad (5.3.6)$$

$$C = \Gamma(\frac{\nu+1}{2}) {}_1F_1\left(\frac{\nu+1}{2}; \frac{1}{2}; \frac{nt^2\delta^2}{2(\nu+t^2)}\right), \quad (5.3.7)$$

$$D = t\delta \sqrt{\frac{2n}{\nu+t^2}} \Gamma(\frac{\nu+2}{2}) {}_1F_1\left(\frac{\nu+2}{2}; \frac{3}{2}; \frac{nt^2\delta^2}{2(\nu+t^2)}\right), \quad (5.3.8)$$

where ${}_1F_1$ denotes the confluent hypergeometric function. Furthermore, t corresponds to the one-sample t -value defined as $t = \sqrt{n}\bar{y}/s$ where s corresponds to the unbiased sample standard deviation and $\nu = n - 1$ corresponds to the degrees of freedom of the t -test. The integral with respect to g in the denominator is due to the t prior representation as a scale-mixture of normal distributions as, for instance, applied in the well-known mixture of g priors (e.g., Liang et al., 2008). This

integral is independent of δ and thus only needs to be evaluated once, something which can easily be accomplished via numerical integration.

In order to derive this expression for the posterior distribution we made use of a lemma distilled from the Bateman project (Bateman et al., 1954; Ly et al., 2017d). Note that in order to calculate the posterior distribution, we only require the t -value and the sample size; this is convenient because it allows a computation of the posterior distribution of δ for a t -test reported in an empirical article which usually presents the t -value and the sample size, but not the raw data.

The posterior distribution for a normal prior of the form $\delta \sim \mathcal{N}(\mu_\delta, g)$ is obtained by removing the integral over g with respect to $\pi(g)$ in the denominator and replacing the shifted and scaled t prior distribution in the numerator by the normal prior distribution which yields:

$$\pi(\delta | d) = \frac{\exp(-\frac{n}{2}\delta^2)(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}} [C + D](2\pi g)^{-\frac{1}{2}} \exp(-\frac{1}{2g}(\delta - \mu_\delta)^2)}{(1 + ng)^{-\frac{1}{2}} \exp(-\frac{n\mu_\delta^2}{2(1+ng)})(1 + \frac{t^2}{\nu(1+ng)})^{-\frac{\nu+1}{2}} [A + B]} \quad (5.3.9)$$

where A, B, C , and D are defined as before. Detailed derivations can be found in the online appendix.

5.3.4 Bayes factor

In this section we present the Bayes factor for the informed generalisation of Jeffreys's one-sample t -test. The Bayes factor $\text{BF}_{10}(d)$ is given by the ratio of the *marginal likelihoods* of \mathcal{H}_1 and \mathcal{H}_0 . The marginal likelihoods are obtained by integrating out the model parameters with respect to their prior distributions

$$\text{BF}_{10}(d) = \frac{\int_0^\infty \int_{-\infty}^\infty f(d | \delta, \sigma^2) \pi(\delta) \pi(\sigma^2) d\delta d\sigma^2}{\int_0^\infty f(d | \delta = 0, \sigma^2) \pi(\sigma^2) d\sigma^2}. \quad (5.3.10)$$

The Bayes factor for a shifted and scaled t prior on δ is given by:

$$\text{BF}_{10}(d) = \frac{\int_0^\infty (1 + ng)^{-\frac{1}{2}} \exp\left(-\frac{n\mu_\delta^2}{2(1+ng)}\right) \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} [A + B] \pi(g) dg}{\Gamma(\frac{\nu+1}{2})(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}}, \quad (5.3.11)$$

the Bayes factor for a normal prior on δ is given by

$$\text{BF}_{10}(d) = \frac{(1 + ng)^{-\frac{1}{2}} \exp\left(-\frac{n\mu_\delta^2}{2(1+ng)}\right) \left(1 + \frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}} [A + B]}{\Gamma(\frac{\nu+1}{2})(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}}, \quad (5.3.12)$$

where A, B, ν, t , and $\pi(g)$ are defined as before. Note that, as for the posterior distribution of δ , the only information needed to compute this Bayes factor is the t -value and the sample size n . A detailed derivation is presented in the online appendix.

5.3.5 Example I: The crowd within effect

As an example application, consider the so-called *crowd within effect*. When people are asked to provide quantitative judgments, it has long been known that the average of the estimates across persons tends to be more accurate than the individual judgments; this is called the *wisdom of the crowd effect* (Galton, 1907; Surowiecki, 2004). Recently, Vul and Pashler (2008) showed that a similar effect may occur when a single individual is asked to provide a quantitative judgment twice: the average of two successive assessments was more accurate than the first or second assessment alone. Due to the similarity to the wisdom of the crowd effect, Vul and Pashler (2008) termed this phenomenon the *crowd within effect*.

This surprising effect was successfully replicated by Steegen et al. (2014). Here, we present an informed reanalysis of the replication study where we use our prior knowledge from the original study to specify the effect size prior distribution for the replication experiment, following Lindley's adage "today's posterior is tomorrow's prior" (Lindley, 1972).

The original experiment by Vul and Pashler featured an immediate and a delayed condition. In the immediate condition, participants provided judgments to eight questions such as "What percentage of the world's airports are in the United States?" and they were asked to make a second guess for each of the questions immediately after they had completed the questionnaire. In the delayed condition, participants provided the second judgments three weeks after completing the questionnaire. For our reanalysis, we focus on the results for the delayed condition. Furthermore, we only consider the comparison of the averaged estimate to the first guess. In the original experiment, a classical paired *t*-test indicated that the error of the average was smaller than the error of the first guess: $t(172) = 6.22, p < .001$. Steegen et al. successfully replicated this finding: $t(139) = 4.02, p < .001$. For our reanalysis, we proceeded as follows (Verhagen and Wagenmakers, 2014):

1. We analysed the original experiment using a zero-centred Cauchy prior for the effect size as proposed by Jeffreys (1948). However, instead of using the standard Cauchy distribution that was Jeffreys's default choice, we set the scale parameter of the Cauchy distribution to $1/\sqrt{2} \approx 0.707$, the present default choice in the field of psychology. This Cauchy prior is equivalent to a $t(0, 1/\sqrt{2}, 1)$ prior distribution for the effect size parameter δ .
2. Using the default Cauchy prior distribution, the paired *t*-test reported in the original article yields a Bayes factor of $BF_{10}(d) = 2,483,125$ indicating overwhelming evidence for \mathcal{H}_1 with $\delta \sim \text{Cauchy}(0, 1/\sqrt{2})$ over $\mathcal{H}_0 : \delta = 0$. This result can be obtained using the equations just presented, but the analysis is also easily conducted using the `BayesFactor` R package (Morey and Rouder, 2015) which also produces samples from the posterior distribution for the effect size parameter δ . We fitted a *t* distribution to these posterior samples which yielded the following distribution: $t(0.465, 0.078, 41.478)$.
3. Next, the posterior distribution for δ served as the prior distribution for the analysis of the replication experiment. Specifically, we contrasted the skeptic's $\mathcal{H}_0 : \delta = 0$ to the proponent's $\mathcal{H}_F : \delta \sim t(0.465, 0.078, 41.478)$.

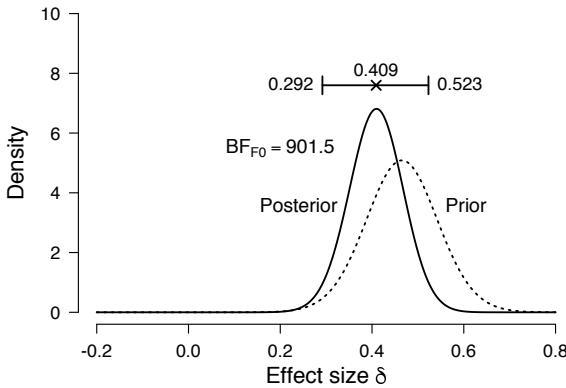


Figure 5.1: Results of the informed reanalysis of the crowd within effect replication study by Steegen et al. (2014). The dotted line corresponds to the informed $t(0.465, 0.078, 41.478)$ prior distribution based on our knowledge from the original study by Vul and Pashler (2008). The solid line corresponds to the posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favour of the informed alternative hypothesis over the null hypothesis equals $BF_{F0}(d) = 901.5$. Figure available at <https://tinyurl.com/kzrhad6> under CC license <https://creativecommons.org/licenses/by/2.0/>.

We obtain a Bayes factor of $BF_{F0}(d) = 901.5$ indicating that the data from Steegen et al. are about 900 times more likely under \mathcal{H}_F than under \mathcal{H}_0 . As a comparison, we also conducted the default analysis of the replication experiment (i.e., using again the zero-centred Cauchy prior with scale parameter $1/\sqrt{2}$ instead of the informed t prior distribution) which yielded a Bayes factor of $BF_{10}(d) = 170.2$.

Fig. 5.1 displays the informed prior distribution (dotted line), the posterior distribution after observing the replication experiment (solid line), a 95% posterior credible interval, the posterior median, and the Bayes factor. The posterior distribution is located near the informed prior distribution, albeit centred on a slightly smaller effect size value.

Fig. 5.2 displays the results for the default Cauchy prior distribution with a scale parameter of $1/\sqrt{2}$ (dotted line). Compared to the posterior distribution based on the informed t prior, the posterior distribution based on the default Cauchy prior is wider and shifted towards smaller effect size values. This reflects the influence of the prior distribution which, in case of the informed t prior, “pulls” the posterior towards larger values.

Another observation is that the Bayes factor in favour of the informed alternative hypothesis is larger than the Bayes factor in favour of the default alternative hypothesis. This can be explained by interpreting the Bayes factor as a measure

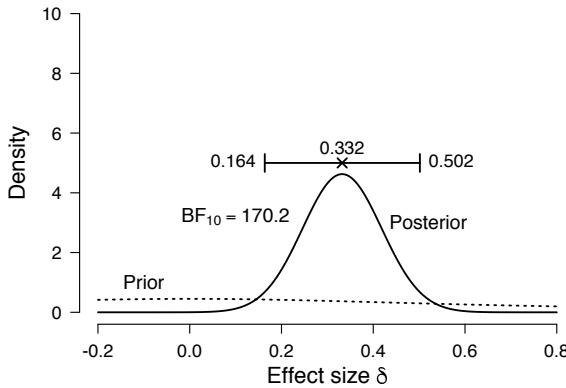


Figure 5.2: Results of the default analysis of the crowd within effect replication study by Steegen et al. (2014). The dotted line corresponds to the default Cauchy prior distribution with scale parameter $1/\sqrt{2}$. The solid line corresponds to the posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favour of the default alternative hypothesis equals $\text{BF}_{10}(d) = 170.2$. Figure available at <https://tinyurl.com/n7tkm9y> under CC license <https://creativecommons.org/licenses/by/2.0/>.

of the predictive success of two competing hypotheses. The informed alternative hypothesis makes riskier predictions than the default alternative hypothesis, but since the results are consistent with these predictions, the informed hypothesis is rewarded more. In fact, we can obtain the Bayes factor that compares the informed alternative hypothesis to the default alternative hypothesis by transitivity

$$\text{BF}_{F1}(d) = \underbrace{\frac{p(d | \mathcal{H}_F)}{p(d | \mathcal{H}_0)}}_{\text{BF}_{F0}(d)} \underbrace{\frac{p(d | \mathcal{H}_0)}{p(d | \mathcal{H}_1)}}_{\text{BF}_{01}(d)} = \frac{\text{BF}_{F0}(d)}{\text{BF}_{10}(d)}. \quad (5.3.13)$$

Hence, the Bayes factor in favour of the informed alternative hypothesis over the default alternative hypothesis equals $\text{BF}_{F1}(d) = 901.5/170.2 \approx 5.3$.

5.4 Two-sample t -test

5.4.1 Model

We extend our informed approach to the unpaired/independent samples version of Jeffreys's t -test. We assume that n_y observations are drawn from a normal distribution of the form $y_i \sim \mathcal{N}(\mu - \frac{\alpha}{2}, \sigma^2)$ where $i = 1, 2, \dots, n_y$ in the first group, and n_x observations are drawn from a normal distribution of the form

$x_j \sim \mathcal{N}(\mu + \frac{\alpha}{2}, \sigma^2)$ where $j = 1, 2, \dots, n_x$ in the second group. Similar to the one-sample t -test, we reparameterise the model in terms of the effect size $\delta = \alpha/\sigma$. As before, the null hypothesis $\mathcal{H}_0 : \delta = 0$ is pitted against the alternative hypothesis $\mathcal{H}_1 : \delta \neq 0$.

5.4.2 Prior distributions

The unpaired samples t -test features two nuisance parameters (i.e., the variance σ^2 and the grand mean μ). In line with Jeffreys's approach, we assign these parameters the (independent) parameterisation invariant prior $\pi(\mu, \sigma^2) \propto \sigma^{-2}$. For effect size δ , we again consider an informed shifted and scaled t distribution, that is $\delta \sim t(\mu_\delta, \gamma, \kappa)$, and an informed normal prior, that is $\delta \sim \mathcal{N}(\mu_\delta, g)$.

5.4.3 Posterior distribution for effect size δ and Bayes factor

Conveniently, for the independent samples t -test, the expressions for the posterior distribution and the Bayes factor are obtained by adjusting the respective expressions for the one-sample/paired samples t -test (i.e., Eqns. (5.3.2, 5.3.9), and Eqns (5.3.11, 5.3.12) in the following way:

1. Replace n by an “effective” sample size n_{eff} defined as $n_{\text{eff}} = \frac{n_y n_x}{n_y + n_x}$ (Rouder et al., 2009; Ly et al., 2016a).
2. Replace the one-sample/paired samples degrees of freedom $\nu = n - 1$ by the independent samples degrees of freedom $\nu = n_y + n_x - 2$.
3. Replace the one-sample/paired samples t -test t -value by the independent samples t -value defined as

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_y + n_x - 2} \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}}. \quad (5.4.1)$$

A detailed derivation is presented in the online appendix.

5.4.4 Example II: The facial feedback hypothesis

The *facial feedback hypothesis* states that affective responses can be influenced by one’s facial expression even when that facial expression is not the result of an emotional experience. In a seminal study, Strack et al. (1988) found that participants who held a pen between their teeth (inducing a facial expression similar to a smile) rated cartoons as more funny than participants who held a pen with their lips (inducing a facial expression similar to a pout).

In a recently published Registered Replication Report (Wagenmakers et al., 2016a), 17 labs worldwide attempted to replicate this finding using a preregistered and independently vetted protocol. A random-effects meta-analysis yielded an estimate of the mean difference between the “smile” and “pout” condition equal to 0.03 [95% CI: $-0.11, 0.16$]. Furthermore, one-sided default Bayesian unpaired

t -tests (using a Cauchy prior scale of $1/\sqrt{2}$) revealed that for all 17 studies, the Bayes factor indicated evidence in favour of the null hypothesis and for 13 out of the 17 studies, the Bayes factor in favour of the null was larger than 3. Overall, the authors concluded that “the results were inconsistent with the original result” (Wagenmakers et al., 2016a, p. 924).

Here we present an informed reanalysis of the replication studies based on a prior elicitation effort with Dr. Suzanne Oosterwijk, a social psychologist at the University of Amsterdam with considerable expertise in this domain.

5.4.4.1 Prior elicitation

Before commencing the elicitation process, we asked our expert to ignore the knowledge about the failed replication of Strack et al. (1988). Next, we stressed that the goal of the elicitation effort was to obtain an informed prior distribution for δ *under the alternative hypothesis*, that is, under the assumption that the effect is present. This was important in order to prevent unwittingly eliciting a prior that is a mixture between a point mass at zero and the distribution of interest. Then, we proceeded in steps of increasing sophistication. First, together with the expert we decided that the theory specified a direction, implying a one-sided hypothesis test. Next, we asked the expert to provide a value for the median of the effect size: this yielded a value of 0.35. Subsequently, we asked for values for the 33% and 66% percentile of the prior distribution for the effect size: this yielded values of 33%-tile = 0.25 and 66%-tile = 0.45. To finesse and validate the specified prior distribution we used the MATCH Uncertainty Elicitation Tool (<http://optics.eee.nottingham.ac.uk/match/uncertainty.php>), a web application that allows one to elicit probability distributions from experts (Morris et al., 2014). The complete elicitation effort took approximately one hour and resulted in a t distribution with location 0.350, scale 0.102, and 3 degrees of freedom.

5.4.4.2 Reanalysis of the Oosterwijk replication study

Having elicited an informed prior distribution for δ under \mathcal{H}_F , we now turn to a detailed reanalysis of the facial feedback replication attempt from Dr. Oosterwijk’s lab at the University of Amsterdam. Later, we summarise the results for all 17 replication attempts.

The alternative hypothesis is directional, that is, the teeth condition is predicted to result in relatively high funniness ratings, not relatively low funniness ratings. In order to respect the directional nature of the alternative hypothesis the two-sided informed t -test outlined above requires a correction. Specifically, the Bayes factor that compares an alternative hypothesis that only allows for positive effect size values to the null hypothesis can be computed via a simply identity that exploits the transitive nature of the Bayes factor (Morey and Wagenmakers,

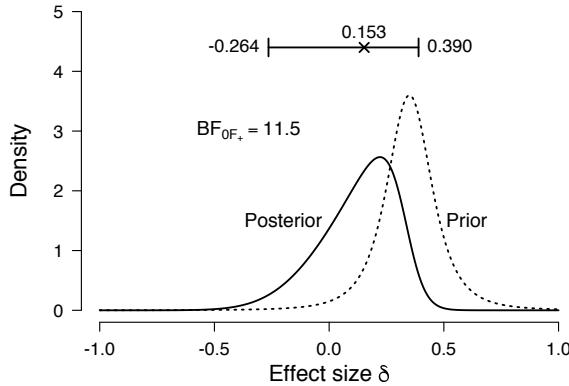


Figure 5.3: Results of an informed reanalysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the elicited $t(0.350, 0.102, 3)$ prior distribution. The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favour of the null hypothesis over the one-sided informed alternative hypothesis equals $\text{BF}_{0F_+}(d) = 11.5$. Figure available at <https://tinyurl.com/mk7uaxm> under CC license <https://creativecommons.org/licenses/by/2.0/>.

2014):

$$\text{BF}_{+0}(d) = \frac{\underbrace{p(d | \mathcal{H}_+)}_{\text{BF}_{+1}(d)} \underbrace{p(d | \mathcal{H}_1)}_{\text{BF}_{10}(d)}}{\underbrace{p(d | \mathcal{H}_1)}_{\text{BF}_{+1}(d)} \underbrace{p(d | \mathcal{H}_0)}_{\text{BF}_{10}(d)}} = \text{BF}_{+1}(d)\text{BF}_{10}(d). \quad (5.4.2)$$

We already showed how to obtain $\text{BF}_{10}(d)$, that is, the Bayes factor for the two-sided test of an informed alternative hypothesis; $\text{BF}_{+1}(d)$ can be obtained by simply dividing the posterior mass for δ larger than zero (obtained by numerically integrating Eq. (5.3.2)) by the prior mass for δ larger than zero. The Bayes factor hypothesis test that we report will respect the directional nature of the facial feedback hypothesis and include the correction term from Eq. (5.4.2).

Fig. 5.3 displays the results of the reanalysis of the data from the Oosterwijk lab. The prior and posterior distribution do not impose the directional constraint. The one-sided informed Bayes factor equals $\text{BF}_{0F_+}(d) = 11.5$ indicating that the data are about eleven times more likely under the null hypothesis than under the one-sided informed alternative hypothesis.

For comparison, Fig. 5.4 displays the results based on the default one-sided zero-centred Cauchy distribution with scale $1/\sqrt{2}$. The one-sided default Bayes factor equals $\text{BF}_{0+}(d) = 8.7$, indicating that the data are about 9 times more likely under the null hypothesis than under the one-sided default alternative hypothesis.

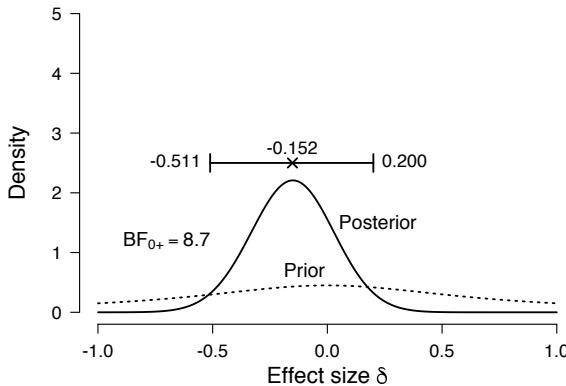


Figure 5.4: Results of the default analysis of the facial feedback hypothesis replication data from the Oosterwijk lab. The dotted line corresponds to the default Cauchy prior distribution with scale parameter $1/\sqrt{2}$. The solid line corresponds to the associated posterior distribution, with a 95% credible interval and the posterior median displayed on top. The Bayes factor in favour of the null hypothesis over the one-sided default alternative hypothesis equals $\text{BF}_{0+}(d) = 8.7$. Figure available at <https://tinyurl.com/mgs28ob> under CC license <https://creativecommons.org/licenses/by/2.0/>.

Hence, both the informed and the default Bayes factor yield the same qualitative conclusion, that is, considerable evidence for the null hypothesis. However, the unrestricted posterior distributions differ noticeably between the informed and the default analysis: the posterior median based on the informed prior specification is positive and equal to 0.153 (95% credible interval: $[-0.264, 0.390]$) whereas the posterior median based on the default prior distribution is equal to -0.152 (95% credible interval: $[-0.511, 0.200]$).

5.4.4.3 Reanalysis of all 17 facial feedback replication studies

Fig. 5.5 displays the (nondirectional) posterior distributions under the informed $\mathcal{H}_F : \delta \sim t(0.350, 0.102, 3)$ and under the default $\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, 1/\sqrt{2})$ for the reanalysis of all 17 replication studies. The posterior distributions for δ differ noticeably for the informed and the default prior specification. Under \mathcal{H}_F , the posterior distributions are shifted towards larger effect size values and the posteriors are more peaked than the ones corresponding to the default prior under \mathcal{H}_1 .

Fig. 5.6 displays the (directional) informed and default Bayes factors for the reanalysis of all 17 replication studies. Each dot corresponds to a study; the x -coordinate corresponds to the one-sided informed Bayes factor and the y -coordinate

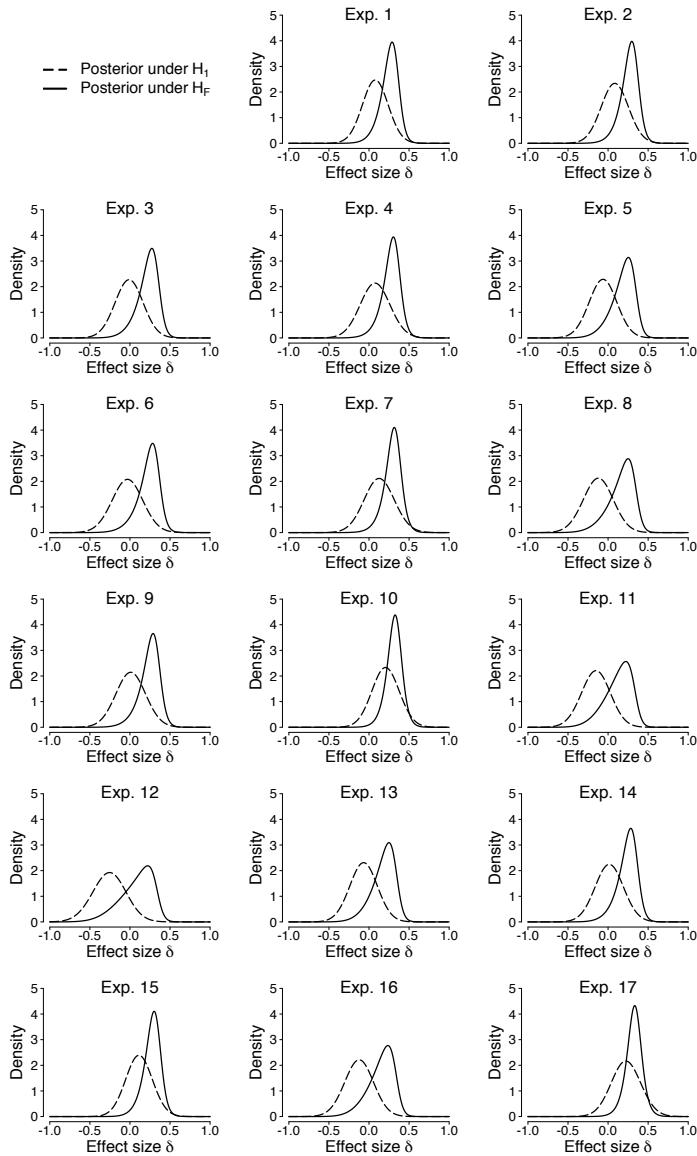


Figure 5.5: Comparison of the posterior distributions for δ under $\mathcal{H}_F : \delta \sim t(0.350, 0.102, 3)$ and $\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, 1/\sqrt{2})$ for the facial feedback hypothesis replication data from the 17 labs. Figure available at <https://tinyurl.com/17pxbno> under CC license <https://creativecommons.org/licenses/by/2.0/>.

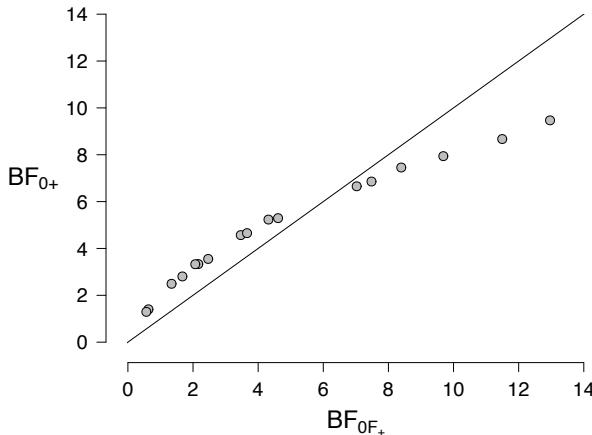


Figure 5.6: Comparison of the one-sided default and one-sided informed Bayes factor analysis of the facial feedback hypothesis replication data from the 17 labs. The x -coordinate corresponds to the one-sided informed Bayes factor and the y -coordinate corresponds to the one-sided default Bayes factor as reported in Wagenmakers et al. (2016a). Figure available at <https://tinyurl.com/ke4489k> under CC license <https://creativecommons.org/licenses/by/2.0/>.

corresponds to the one-sided default Bayes factor as reported in Wagenmakers et al. (2016a). The results are qualitatively similar. For Bayes factors smaller than about six, the one-sided default Bayes factor provides slightly more evidence for the null hypothesis than the one-sided informed Bayes factor. For Bayes factors larger than about six, this pattern is reversed.

In sum, for parameter estimation it matters whether the analysis is based on an informed prior distribution or a default prior distribution. In contrast, for hypothesis testing the results are relatively similar: both the informed Bayes factor and the default Bayes factor support the conclusion that the original study by Strack et al. (1988) could not be successfully replicated by the 17 labs involved in the replication attempt.

5.5 Example III: Reanalysis of 593 t -tests

For our final example we reanalyse the significant subset of the 855 t -tests collected by Wetzels et al. (2011) from the 2007 issues of two popular psychology journals (i.e., *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition*). The analysis of these 593 significant t -tests allows for an empirical investigation of the relation between the default Bayes factor and the informed Bayes factor.

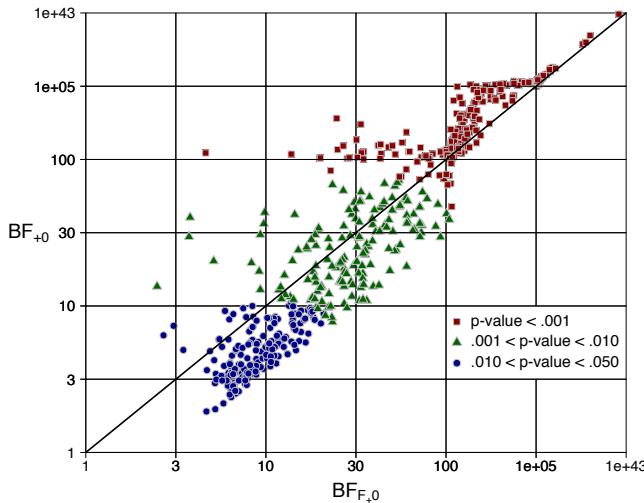


Figure 5.7: Comparison of the one-sided informed Bayes factor that specifies a zero truncated $t(0.350, 0.102, 3)$ prior for δ under the alternative hypothesis and the one-sided default Bayes factor that specifies a zero truncated $\text{Cauchy}(0, 1/\sqrt{2})$ prior for δ under the alternative hypothesis for the 593 significant t -tests reported in Wetzels et al. (2011). The corresponding two-sided p -values are indicated by different shapes and colours: red squares correspond to p -values smaller than .001; green triangles correspond to p -values between .001 and .01; blue circles correspond to p -values between .01 and .05. Figure available at <https://tinyurl.com/m5fatwu> under CC license <https://creativecommons.org/licenses/by/2.0/>.

For the informed Bayes factor, we adopt the expert prior from the previous example, that is, a $t(0.350, 0.102, 3)$ distribution for δ . This prior distribution was elicited within the context of the facial feedback hypothesis, but we believe it may serve as an informed prior for small-to-medium effect sizes across many areas of psychological research. For the default Bayes factor we again use the $\text{Cauchy}(0, 1/\sqrt{2})$ distribution for δ . Both Bayes factors are one-sided.

Fig. 5.7 displays the one-sided informed Bayes factors, $\text{BF}_{F,0}(d)$, and the default one-sided Bayes factors, $\text{BF}_{+,0}(d)$, for the 593 significant t -tests collected by Wetzels et al. (2011). The two-sided p -values are indicated by different shapes and colours. The overall pattern indicates a positive correlation between the informed and default Bayes factor. However, for Bayes factors that roughly correspond to t -tests with p -values in between .001 and .05, the informed Bayes factor generally provides more evidence for the alternative hypothesis than the default Bayes factor. For large Bayes factors that roughly correspond to p -values smaller than .001, this pattern is reversed.

This reversal can be explained by considering how well the informed alternative hypothesis has predicted the observed effect sizes. When considering the informed

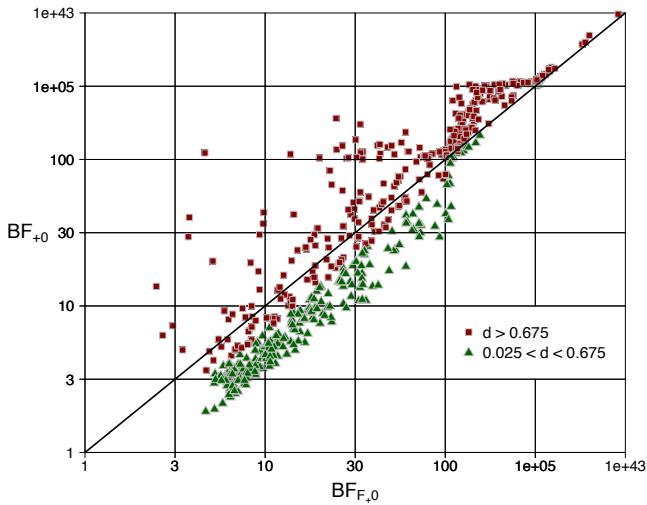


Figure 5.8: Comparison of the one-sided informed Bayes factor that specifies a zero truncated $t(0.350, 0.102, 3)$ prior for δ under the alternative hypothesis and the one-sided default Bayes factor that specifies a zero truncated Cauchy($0, 1/\sqrt{2}$) prior for δ under the alternative hypothesis for the 593 significant t -tests reported in Wetzels et al. (2011). The corresponding observed effect sizes are indicated by different shapes and colours: red squares correspond to observed effect sizes larger than 0.675; green triangles correspond to observed effect sizes between 0.025 and 0.675. Figure available at <https://tinyurl.com/mo5g9y2> under CC license <https://creativecommons.org/licenses/by/2.0/>.

$t(0.350, 0.102, 3)$ prior, the central 95% prior credible interval ranges from about 0.025 to 0.675. Hence, when the observed effect sizes fall in the predicted range (i.e., roughly from 0.025 to 0.675) the informed alternative hypothesis will perform relatively well; when the observed effect sizes are larger or smaller the default alternative hypothesis will perform relatively well.

Fig. 5.8 displays the same Bayes factors as Fig. 5.7 but indicates the observed effect sizes by different shapes and colours. For all observed effect sizes that fall within the interval that is plausible under the informed alternative hypothesis, $\text{BF}_{F,0}(d)$ provides more evidence for an effect than $\text{BF}_{+0}(d)$. Hence, this empirical investigation suggests that the informed generalisation of Jeffreys's t -tests can lead to more diagnostic tests.

5.6 Quantifying evidence for \mathcal{H}_0

A remaining question is how the informed Bayes factor compares to the default Bayes factor in case the null hypothesis is (approximately) true. To investigate this scenario we computed the maximum possible Bayes factor in favour of the null hypothesis as a function of the number of participants per group by fixing the

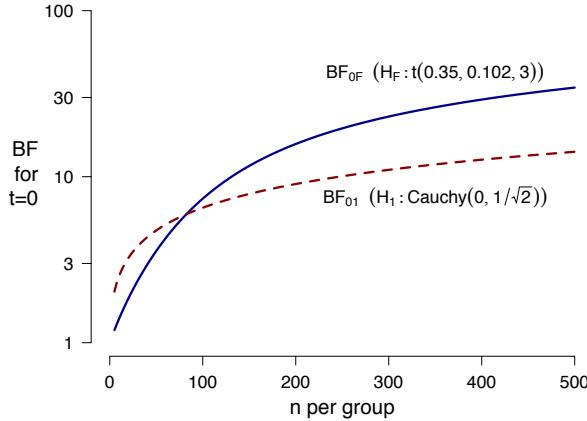


Figure 5.9: The maximum possible Bayes factor in favour of \mathcal{H}_0 as a function of the number of participants per group for unpaired t -tests. The dashed red line corresponds to the default Bayes factor that specifies a $\text{Cauchy}(0, 1/\sqrt{2})$ prior for δ under the alternative hypothesis; the solid blue line corresponds to the informed Bayes factor that specifies a $t(0.350, 0.102, 3)$ prior for δ under the alternative hypothesis. Figure available at <https://tinyurl.com/m92mvpz> under CC license <https://creativecommons.org/licenses/by/2.0/>.

t -value to zero. Fig. 5.9 displays the results. When the number of participants per group is small, the default Bayes factor provides more evidence for \mathcal{H}_0 than does the informed Bayes factor. However, when the number of participants per group is larger than 82, this pattern is reversed.

An intuitive explanation for the reversal is provided by the Savage-Dickey density ratio representation of the Bayes factor (Dickey and Lientz, 1970): the Bayes factor in favour of \mathcal{H}_0 equals the ratio of posterior to prior density for δ under the alternative hypothesis evaluated at the test value $\delta = 0$. When n is small, the posterior for δ under the informed hypothesis is similar to the informed prior whereas the less restrictive default prior will be updated more strongly by the data. Hence, in case \mathcal{H}_0 is true, the ratio of posterior to prior density evaluated at zero will be larger for the default alternative hypothesis than for the informed alternative hypothesis. However, when n grows large, the data start to overwhelm the prior so that the posterior distributions become more similar. In that case, the ratio of posterior to prior density will be larger for the informed alternative hypothesis since its prior density at zero is smaller than that of the default prior. In the limit as $n \rightarrow \infty$, the posterior distributions will be identical; in this case, the difference of the log Bayes factor for the informed and the default prior is equal to the difference of the log of the informed prior evaluated at zero and the log of the default prior evaluated at zero (e.g., Gronau and Wagenmakers, 2017).

5.7 Concluding comments

The comparison between two means is a quintessential inference problem. Harold Jeffreys believed that the only satisfactory solution lay in the application of a Bayes factor t -test (Rouder et al., 2009; Jeffreys, 1961; Ly et al., 2016a; Wetzels et al., 2009), the purpose of which is to quantify the relative predictive performance of two competing hypotheses: the null hypothesis \mathcal{H}_0 that represents an invariance or general law and the alternative hypothesis \mathcal{H}_1 that relaxes this restriction. However, in their current form, Jeffreys's tests do not allow researchers to incorporate expert knowledge in the specification of the prior distribution for effect size under the alternative hypothesis, as all of Jeffreys' priors are centred at zero.

In this article, we have presented an informed extension of Jeffreys's t -tests that admits the specification of prior distributions centred away from zero. This flexibility may encourage the use of prior distributions that better represent the predictions from the hypothesis under test, producing more diagnostic outcomes for the same data. Specifically, our final example showed that when the observed effect size falls within the region of plausible values predicted by the informed prior, the informed Bayes factor provides more evidence for an effect than does the default Bayes factor. Furthermore, we illustrated that when \mathcal{H}_0 is true and the sample size is sufficiently large, the informed Bayes factor can also provide more evidence for the nullity of an effect than does the default Bayes factor.

The merits and debits of an informed prior specification will always remain a topic of debate. In our opinion, the prior distribution is an integral part of the model specification (e.g., Vanpaemel, 2010) and should be adjusted to include existing knowledge and impose meaningful constraints. The perceived dangers of the subjectivity that accompanies an informed analysis can be mitigated by preregistering the prior distribution (Chambers, 2013) and by conducting sensitivity analyses in which different prior choices are explored. For instance, when it comes to small-to-medium effects in experimental psychology, we believe the Oosterwijk prior is eminently plausible; different specifications are certainly possible, but they will need to respect the basic constraints (i.e., a mean around .35, most mass between .1 and .6) so that the results will be relatively insensitive to these choices.

We hope that empirical scientists will feel encouraged to pay more attention to the predictions of the hypotheses they seek to test (Rouder et al., 2016a, 2016b). There is a price to be paid in order to apply an informed analysis –the method cannot be applied without thought– but the corresponding reward is a more diagnostic test.

The online appendix can be found on arXiv: <https://arxiv.org/abs/1704.02479>.

Chapter 6

A Limit-Consistent Bayes Factor for Testing the Equality of Two Poisson Rates

Abstract

To facilitate the selection of prior distributions with good properties we introduce the desideratum of *limit-consistency*. This desideratum is relevant for tests of equality between two processes, and it concerns the hypothetical scenario where data acquisition for one process is terminated early whereas data acquisition of the second process continues indefinitely. In such cases, the Bayes factor ought to approach a finite limit. We rederive Jeffreys's 1939 Bayes factor for the comparison between two Poisson rates and prove that it is not limit-consistent: as sample size for the uninterrupted process increases, support in favour of the null hypothesis eventually grows without bound. We generalise Jeffreys's approach by centring the alternative hypothesis around the value specified by the null hypothesis. We prove that the generalised version of Jeffreys's test is limit-consistent.

Keywords: Bayes factor, hypergeometric functions, statistical evidence, two-sample test.

6.1 Introduction

A homogenous Poisson process $Y_i(t_i)$ has rate λ_i if, for given $\lambda_i > 0$, the chance of observing $Y_i(t_i) = y_i$ after t_i units of time is

$$f(y_i | \lambda_i, t_i) = \frac{(\lambda_i t_i)^{y_i}}{y_i!} e^{-\lambda_i t_i}, \quad (6.1.1)$$

This chapter is under preparation as Ly, A., Raj, A., Marsman, M. & Wagenmakers, E.-J. (2017). A limit-consistent Bayes factor for testing the equality of two Poisson rates.

6. A LIMIT-CONSISTENT BAYES FACTOR FOR TESTING THE EQUALITY OF TWO POISSON RATES

where y_i is any non-negative integer. A famous use of the Poisson distribution is in the detection of radioactivity using a Geiger counter (Rutherford et al., 1910; Stirzaker, 2000).

Here we study a two-sample problem and focus on evaluating the hypothesis that two homogenous Poisson processes have equal rate with exposure times t_1 and t_2 not necessarily the same.

Throughout the text t_1 and t_2 represent time, but they could as well relate to measurement from different areas (e.g., Haight, 1967) or in the case of radioactivity refer to the different number of atoms in two specimens of rock. Indeed, measurements of Poisson processes with $t_1 = t_2$ are rare and, more often than not, $t_1 \neq t_2$.

The frequentist test for the equality of two Poisson rates has received considerable attention (e.g., Haight, 1967; Krishnamoorthy and Thomson, 2004; Przyborowski and Wilenski, 1940 for the $t_1 = t_2$ case, and Ng et al., 2007; Racliff, 1964; Shiue and Bain, 1982 for the $t_1 \neq t_2$ case). Here we focus on the Bayesian hypothesis test known as the Bayes factor.

The purpose of this paper is three-fold: Firstly, we introduce the desideratum of limit-consistency, relevant for the behaviour of any test that involves a comparison between two or more processes. Secondly, we rederive the Bayes factor proposed by Jeffreys for the two-sample Poisson problem (Jeffreys, 1939, pp. 211-212)¹ and prove that it violates limit-consistency. Thirdly, we propose a generalisation of Jeffreys's test that is limit-consistent.

6.1.1 Desiderata that facilitate the selection of prior distributions

Let \mathcal{M}_1 denote the model in which the rates λ_1 and λ_2 of the two Poisson processes are free to vary, and \mathcal{M}_0 the restriction of \mathcal{M}_1 such that $\lambda_1 = \lambda_2$. In the Bayesian setting the models are assigned prior model probabilities $0 < P(\mathcal{M}_0), P(\mathcal{M}_1) < 1$, which in light of the observed data d can be updated to posterior model probabilities using Bayes' theorem. Doing this for both models and subsequently taking the ratio of the result leads to the key expression

$$\underbrace{\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_0 | d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}}. \quad (6.1.2)$$

The term $\text{BF}_{10}(d)$ is known as the Bayes factor and equals the change from prior to posterior model odds brought about by the observed data d (Etz and Wagenmakers, 2017; Jeffreys, 1935; Kass and Raftery, 1995; Ly et al., 2016a).

Note that the Bayes factor $\text{BF}_{10}(d)$ does not depend on the prior model probabilities $P(\mathcal{M}_1)$ and $P(\mathcal{M}_0)$. However, the Bayes factor is the ratio of marginal likelihoods:

$$p(d | \mathcal{M}_i) = \int f(d | \theta_i, \mathcal{M}_i) \pi_i(\theta_i) d\theta_i, \quad (6.1.3)$$

¹The widely available third edition contains the same text (i.e., Jeffreys, 1961, pp. 267–268).

which shows that the Bayes factor does depend on the priors $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ that are assigned to the parameters within the two models.

For this reason, the selection of prior distributions demands careful consideration. Fortunately, general principles constrain the selection of prior distributions. For instance, priors of arbitrary width yield Bayes factors that favour the null model irrespective of the observed data (e.g., the Jeffreys-Lindley-Bartlett paradox Bartlett, 1957; Jeffreys, 1961; Lindley, 1957). Furthermore, the prior on the test-relevant parameter must be proper, as improper priors contain suppressed normalisation constants and may lead to Bayes factors of arbitrary value. Consequently, in Bayes factor hypothesis testing we cannot use popular “non-informative” priors selected by formal rules (Kass and Wasserman, 1996) such as the right-Haar prior (e.g., Berger et al., 1998; Ghosh, 2011) and Jeffreys’s parameterisation-invariant prior (e.g., Jeffreys, 1946; Ly et al., 2017c) which are both improper. We also require that a reasonable Bayes factor does not depend on the units of measurement that the researcher chooses to represent the data. Naturally, we also desire that the Bayes factors are calculable in the sense that both integrals in the numerator and denominator are solvable for any data set d .

Other desiderata unfortunately hold only for continuous random variables. For instance, the desideratum of *predictive matching* states that the Bayes factor ought to be perfectly indifferent, i.e., $\text{BF}_{10}(d) = 1$, in case the data are completely uninformative; the desideratum of *information consistency* states that the Bayes factor ought to provide infinite support for the alternative hypothesis in case the data are overwhelmingly informative (for a review see Bayarri et al., 2012; see also Bayarri and Berger, 2013). In case of discrete data it is not clear what constitutes completely uninformative and overwhelmingly informative.

Here we propose a new and relatively general desideratum that further constrains the selection of prior distributions: *limit-consistency*. This desideratum holds regardless of whether the data are discrete or continuous, and applies whenever the test at hand features a comparison between two or more processes or groups. Consider again a comparison between two Poisson processes and assume that the measurement of the first process is terminated early, whereas the measurement of the second process continues indefinitely. In the limit, knowledge about the second process will reach perfection, but knowledge about the interrupted process will remain incomplete. Consequently, there exists a bound on the level of evidence that can be obtained in a test that compares the two processes. As measurement for the second process continues, the Bayes factors ought to approach a finite limit.

The practical value of limit-consistency as a constraint on the selection of prior distributions will now be demonstrated through Harold Jeffreys’s test for the equality of two Poisson rates.

6.2 Jeffreys's Bayes factor for the comparison of two Poisson rates

Jeffreys's derivation starts with a rewrite of the joint distribution of the two processes $Y_1(t_1)$ and $Y_2(t_2)$, i.e.,

$$f(d | \lambda_1, \lambda_2, \mathcal{M}_1) = \frac{(\lambda_1 t_1)^{y_1}}{y_1!} e^{-\lambda_1 t_1} \frac{(\lambda_2 t_2)^{y_2}}{y_2!} e^{-\lambda_2 t_2} \quad (6.2.1)$$

in terms of the relative timed rate $\theta = \frac{\lambda_1 t_1}{\lambda_1 t_1 + \lambda_2 t_2}$ and the total timed rate $\zeta = \lambda_1 t_1 + \lambda_2 t_2$, that is,

$$f(d | \theta, \zeta, \mathcal{M}_1) = \underbrace{\binom{y.}{y_1} \theta^{y_1} (1 - \theta)^{y_2}}_{f(y_1 | y., \theta)} \underbrace{\frac{\zeta^y.}{y.!} e^{-\zeta}}_{f(y. | \zeta)} \quad (6.2.2)$$

where $y. = y_1 + y_2$ denotes the total number of observations across the two processes. Setting $\lambda_2 = \lambda_1$ shows that \mathcal{M}_0 can be perceived as a restriction of \mathcal{M}_1 with the relative timed rate θ known and fixed at $\theta_0 = \frac{t_1}{t.}$, where $t. = t_1 + t_2$, thus,

$$f(d | \zeta, \mathcal{M}_0) = f(y_1 | y., \frac{t_1}{t.}) f(y. | \zeta). \quad (6.2.3)$$

The factorisation of the two likelihood functions allowed Jeffreys to conceptualise the two-sample Poisson problem as a conditional binomial test, if the same prior on the common parameter ζ is chosen.² This can be achieved by assigning independent gamma priors $\lambda_i \sim \text{Gam}(\alpha_i, \beta t_i)$ in \mathcal{M}_1 and $\lambda_1 \sim \text{Gam}(\alpha., \beta t.)$ in \mathcal{M}_0 , where $\alpha. = \alpha_1 + \alpha_2$; the induced prior on the total timed rate is then $\zeta \sim \text{Gam}(\alpha., \beta)$ under both models.

Under the alternative hypothesis, the test relevant parameter θ receives a beta prior, that is,

$$\pi_\eta(\theta, \zeta | \mathcal{M}_1) = \underbrace{\frac{1}{\mathcal{B}(\alpha_1, \alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}}_{\text{Beta}(\theta; \alpha_1, \alpha_2)} \underbrace{\frac{\beta^\alpha.}{\Gamma(\alpha.)} \zeta^{\alpha.-1} e^{-\beta\zeta}}_{\text{Gam}(\zeta; \alpha., \beta)} \quad (6.2.4)$$

where $\eta = (\alpha_1, \alpha_2, \beta)$ denotes the hyperparameters and \mathcal{B} denotes the beta function. Hence, the prior factorises and, consequently, so do the marginal likelihoods:

$$p_\eta(d | \mathcal{M}_1) = p_{\alpha_1, \alpha_2}(y_1 | y.) p_{\alpha., \beta}(y.), \quad (6.2.5)$$

$$p_\eta(d | \mathcal{M}_0) = f(y_1 | y., \frac{t_1}{t.}) p_{\alpha., \beta}(y.), \quad (6.2.6)$$

where $f(y_1 | y., \frac{t_1}{t.})$ is given in Eq. (6.2.2), and where

$$p_{\alpha_1, \alpha_2}(y_1 | y.) = \binom{y.}{y_1} \frac{\mathcal{B}(y_1 + \alpha_1, y_2 + \alpha_2)}{\mathcal{B}(\alpha_1, \alpha_2)}, \quad (6.2.7)$$

$$p_{\alpha., \beta}(y.) = \frac{\Gamma(\alpha. + y.)}{\Gamma(\alpha.) y.!} \left(\frac{\beta}{1 + \beta} \right)^{\alpha.} \left(\frac{1}{1 + \beta} \right)^{y.}. \quad (6.2.8)$$

²With $t_1 = t_2$ we have $\theta_0 = \frac{1}{2}$ and note the resemblance to the frequentist conditional binomial test proposed by Przyborowski and Wilenski (1940), which was published one year after the first edition of Jeffreys's book.

Dividing Eq. (6.2.5) by Eq. (6.2.6) shows that

$$\text{BF}_{10; \alpha_1, \alpha_2}^J(d) = \frac{\mathcal{B}(y_1 + \alpha_1, y_2 + \alpha_2)}{\mathcal{B}(\alpha_1, \alpha_2)(\frac{t_1}{t_\cdot})^{y_1}(\frac{t_2}{t_\cdot})^{y_2}}, \quad (6.2.9)$$

where $d = (y_1, t_1, y_2, t_2)$. Jeffreys proposed to set $\alpha_1 = \alpha_2 = a$, i.e., $\text{BF}_{10; a}^J(d) = \text{BF}_{10; a, a}^J(d)$ with $a = 1$, that is,

$$\text{BF}_{10; 1}^J(d) = \frac{\mathcal{B}(y_1 + 1, y_2 + 1)}{(\frac{t_1}{t_\cdot})^{y_1}(\frac{t_2}{t_\cdot})^{y_2}}, \quad (6.2.10)$$

to compare the model \mathcal{M}_1 with differing Poisson rates against \mathcal{M}_0 in which the two rates are the same.

6.2.1 Properties of Jeffreys's Bayes factor $\text{BF}_{10; a}^J(d)$

6.2.1.1 Invariances

Observe that by setting $\lambda_i \sim \text{Gam}(\alpha_i, \beta t_i)$ and by specifying the test relevant parameter to be the unitless quantity θ , we have effectively assigned a beta prior $\text{Beta}(\alpha_1, \alpha_2)$ to θ . Within this framework, the setting $\alpha_1 = \alpha_2 = a$ leads to a Bayes factor $\text{BF}_{10; a}^J(d)$ that does not depend on how the processes are labeled, as the same output is obtained for $\tilde{d} = (y_2, t_2, y_1, t_1)$ and for d .

Furthermore, the measurement scale for the times t_1 and t_2 does not affect the outcome, as the Bayes factor $\text{BF}_{10; a}^J(d)$ only depends on the ratio $\frac{t_1}{t_\cdot}$.

6.2.1.2 Uninformativeness for balanced outcomes

Jeffreys's choice for $a = 1$ was inspired by the Bayes factor he developed for the binomial problem $X \sim \text{Bin}(\theta, n)$. Jeffreys (1961, p. 257) noted that when testing $\mathcal{H}_0 : \theta = 1/2$ against $\mathcal{H}_1 : \theta \in (0, 1)$ the Bayes factor $\text{BF}_{10; a}^J(x, n) = 1$ after a single observation $n = 1$ independently of whether we observe a success $x = 1$ or a failure $x = 0$. Jeffreys mentions that this behaviour also extends to the case when the number of observed successes x equals the number of failures $n - x$, (i.e., $n = 2x$), as an additional observation will once again not change the Bayes factor, meaning $\text{BF}_{10; a}^J(x, 2x) = \text{BF}_{10; a}^J(x, 2x+1) = \text{BF}_{10; a}^J(x+1, 2x+1)$. Note that this property holds for any other $a > 0$, but only if $\mathcal{H}_0 : \theta = 1/2$. When applied to the Poisson case, this means that when the two exposure times are the same (i.e., $t_1 = t_2$) and the observed counts are the same (i.e., $y_1 = y_2$), Jeffreys's Bayes factor does not change when a single additional count is added to one of the processes.

6.2.1.3 Limit-inconsistency

Unfortunately, Jeffreys's Bayes factor $\text{BF}_{10; a}^J(d)$ with $a > 0$ is limit-inconsistent, that is, $\text{BF}_{10; a}^J(d)$ does not stabilise once data collection of the first process is interrupted at t_1 and data acquisition of the second process continues indefinitely. The following property shows that Jeffreys's Bayes factor $\text{BF}_{10; a}^J(d)$ will eventually favour the simpler model \mathcal{M}_0 , regardless of the data.

6. A LIMIT-CONSISTENT BAYES FACTOR FOR TESTING THE EQUALITY OF TWO POISSON RATES

Property 6.2.1 ($\text{BF}_{10;a}^J(d)$ is limit-inconsistent). *Let $y_1, t_1, \lambda_2 > 0$ be fixed. Then for every $a > 0$ (i.e., a symmetric beta distribution on θ), Jeffreys's Bayes factor $\text{BF}_{10;a}^J(d)$ tends to zero as t_2 grows.* \diamond

Proof. We proof by contradiction and assume that for all t_2 the logarithm of the Bayes factor $\text{BF}_{10;a}^J(d)$ is bounded from below, thus, the existence of constant M such that

$$\log \text{BF}_{10;a}^J(d) = \log p_a(y_1 | y.) - \log f(y_1 | y., \frac{t_1}{t.}) \geq M \quad (6.2.11)$$

for all t_2 . To simplify matters we use $y_2 = E(Y_2(t_2)) = \lambda_2 t_2$. The asymptotic behaviour of $-\log f(y_1 | y., \frac{t_1}{t_1+t_2})$ for t_2 large can be described by the following two series

$$-y_1 \log(\frac{t_1}{t_1+t_2}) = y_1[\log(\frac{t_2}{t_1}) + \frac{t_1}{t_2}] + \mathcal{O}(\frac{1}{t_2^2}), \quad (6.2.12)$$

$$-\lambda_2 t_2 \log(\frac{t_2}{t_1+t_2}) = \lambda_2 t_1[1 - \frac{t_1}{2t_2}] + \mathcal{O}(\frac{1}{t_2^2}). \quad (6.2.13)$$

In addition, Stirling's approximation of the beta function implies that the logarithm of the Bayes factor behaves as

$$\begin{aligned} \log \text{BF}_{10;a}^J(d) &\sim \log \Gamma(y_1 + a) - (y_1 + a) \log(\lambda_2 t_2 + a) \\ &\quad + y_1 \log\left(\frac{t_2}{t_1}\right) + \lambda_2 t_1, \end{aligned} \quad (6.2.14)$$

when t_2 is large. The assumption that $\log \text{BF}_{10;a}^J(d)$ is bounded from below leads to a contradiction as a rewrite now shows that

$$\log \Gamma(y_1 + a) + \lambda_2 t_1 - M \geq y_1 \log(\lambda_2 t_1 + \frac{at_1}{t_2}) + a \log(\lambda_2 t_2 + a), \quad (6.2.15)$$

from which we can incorrectly conclude that the logarithm is a bounded function. Thus, $\text{BF}_{10;a}^J(d) \rightarrow 0$ as $t_2 \rightarrow \infty$. \square

Two concrete examples are given in Fig. 6.1. In both panels, the dashed line represents the logarithm of $\text{BF}_{10;a}^J(d)$ with $a = 1$. In the left panel $y_1 = t_1 = 1$ and $y_2 = \lambda_2 t_2$ with $\lambda_2 = 5$. The dashed line decreases, meaning that Jeffreys's Bayes factor eventually indicates evidence for the null hypothesis $\lambda_1 = \lambda_2$ as $\lambda_2 = 5$ is estimated more precisely but the estimate of λ_1 remains highly uncertain.

In the right panel $y_1 = t_1 = 1$ and $\lambda_2 = 1$. For small values of t_2 , Jeffreys's Bayes factor $\text{BF}_{10;a}^J(d)$ with $a = 1$ now conveys evidence in favour of the null, which is what is expected in this situation. However, the dashed line again decreases, showing that the evidence for the null grows without bound even though the information about the first process is limited and uncertain. Choosing another $a > 0$ does not solve the problem as the bias for the null model is driven by the term $a \log(\lambda_2 t_2 + a)$ in the asymptotic expansion in Eq. (6.2.15).

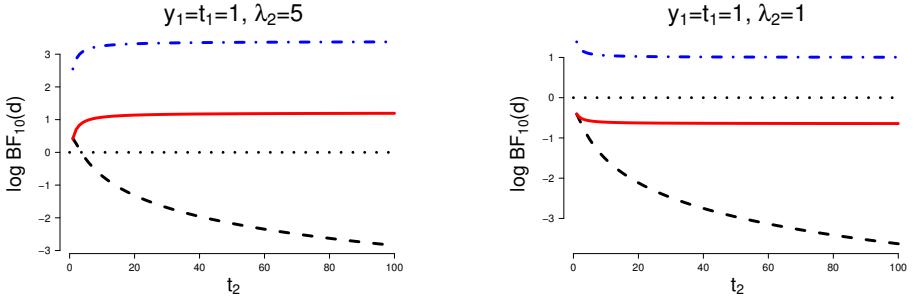


Figure 6.1: Jeffreys's Bayes factor $\text{BF}_{10;1}^J(d)$ (dashed lines in both panels) is limit-inconsistent and increasingly favours the null hypothesis as the exposure time of the second process lengthens. This bias can be eliminated by setting $a = 0$ in the prior distribution, but the resulting Bayes factor $\text{BF}_{10;0}^J(d)$ (dot-dashed lines) unduly favours the alternative model. The localised Bayes factor $\text{BF}_{10;a}^J(d)$ (solid line) is limit-consistent. Left panel: The log of the Bayes factor based on $y_1 = t_1 = 1$ and $\lambda_2 = 5$, which should yield some evidence in favour of the alternative hypothesis as t_2 grows. Right panel: The log of the Bayes factors based on $y_1 = t_1 = 1$ and $\lambda_2 = 1$, which should yield some evidence in favour of the null hypothesis as t_2 grows.

6.3 A limit-consistent Bayes factor for the comparison of two Poisson rates

One way to obtain a limit-consistent Bayes factor is by setting $a = 0$, as the bias term in the asymptotic expansion then cancels. The dot-dashed line in Fig. 6.1 confirms that $\text{BF}_{10;0}^J(d)$ stabilises as t_2 increases. The problem with $a = 0$, however, is that we then effectively use an improper prior with an unspecified normalisation constant on the test relevant parameter θ and this choice introduces new problems. Fig. 6.1 shows the undesirable consequence: in both panels, $\text{BF}_{10;0}^J(d)$ overvalues the support in favour of the alternative hypothesis; this is particularly poignant for the example shown in the right panel, where $y_1 = t_1 = 1$ and $\lambda_2 = 1$, which ought to result in evidence for the null hypothesis. Moreover, $a = 0$ yields infinite support for the alternative when $y_1 = 0$, $y_2 = 1$ and $t_1 = t_2$.

In order to obtain a Bayes factor that is limit-consistent we now consider the localised beta distribution that expands the beta distribution with an additional parameter θ_0 that allows the model \mathcal{M}_1 to be centred on the simpler model \mathcal{M}_0 . This centring occurs on the logit scale. Recall that the standard beta distribution $\text{Beta}(\alpha_1, \alpha_2)$ reparameterised as $\phi = \log(\frac{\theta}{1-\theta})$ is given by³

$$\int_{-\infty}^{\infty} \pi(\phi; \alpha_1, \alpha_2) d\phi = \frac{1}{\mathcal{B}(\alpha_1, \alpha_2)} \int_{-\infty}^{\infty} \frac{e^{\phi\alpha_1}}{(1+e^\phi)^{\alpha_1+\alpha_2}} d\phi. \quad (6.3.1)$$

³Thus, $\int d\phi = \int \theta^{-1}(1-\theta)^{-1} d\theta$ and equivalently, $\theta = \frac{e^\phi}{1+e^\phi}$ and therefore, $\int d\theta = \int \frac{e^\phi}{(1+e^\phi)^2} d\phi$.

6. A LIMIT-CONSISTENT BAYES FACTOR FOR TESTING THE EQUALITY OF TWO POISSON RATES

As ϕ ranges over the real line, its location can be shifted by ϕ_0 resulting in

$$\pi_{\alpha_1, \alpha_2}(\phi; \phi_0) = \frac{1}{\mathcal{B}(\alpha_1, \alpha_2)} \frac{e^{\alpha_1(\phi-\phi_0)}}{(1 + e^{\phi-\phi_0})^{\alpha_1+\alpha_2}}. \quad (6.3.2)$$

Back transforming this distribution with $\phi_0 = \log(\frac{\theta_0}{1-\theta_0})$ yields the localised beta distribution

$$\pi_{\alpha_1, \alpha_2}(\theta; \theta_0) = \underbrace{\frac{1}{\mathcal{B}(\alpha_1, \alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1} (\frac{1-\theta_0}{\theta_0})^{\alpha_1} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-(\alpha_1+\alpha_2)}}_{\text{Beta}(\theta; \alpha_1, \alpha_2)}, \quad (6.3.3)$$

where $\text{Beta}(\theta; \alpha_1, \alpha_2)$ refers to the beta density. Note that with $\theta_0 = 1/2$ (i.e., $t_1 = t_2$), $\alpha_1 = \alpha_2 = a$ and $a = 1$, we retrieve Jeffreys's choice on θ in the two-sample Poisson problem. Hence, a Bayes factor constructed from this prior retains the desirable behaviour of $\text{BF}_{10; a}^J(d)$ for $a > 0$ at $t_1 = t_2$. To derive this localised Bayes factor we have to calculate the marginal likelihood with this new prior.

Property 6.3.1 (The marginal likelihood of a binomially distributed random variable with a localised beta prior). *Let $f(y_1 | y., \theta)$ be the binomial pmf and θ distributed according to a localised beta prior. The marginal likelihood is then*

$$p_{\alpha_1, \alpha_2, \theta_0}(y_1 | y.) = \binom{y.}{y_1} \frac{\mathcal{B}(\alpha_1 + y_1, \alpha_2 + y_2)}{\mathcal{B}(\alpha_1, \alpha_2)} \times {}_2F_1(\alpha_1 + \alpha_2, \alpha_1 + y_1; \alpha_1 + \alpha_2 + y.; 2 - \frac{1}{\theta_0}) (\frac{1-\theta_0}{\theta_0})^{\alpha_1} \quad (6.3.4)$$

where ${}_2F_1(u, v; w; z)$ denotes Gauss' hypergeometric function. \diamond

Proof. With $C = \binom{y.}{y_1} (\frac{1-\theta_0}{\theta_0})^{\alpha_1} / \mathcal{B}(\alpha_1, \alpha_2)$, $u_1 = y + \alpha_1$, $u_2 = y_2 + \alpha_2$, $v = \alpha_1 + \alpha_2$ and by definition of the prior predictive, we have

$$p_{\alpha_1, \alpha_2, \theta_0}(y_1 | y.) = C \int_0^1 \theta^{u_1-1} (1-\theta)^{u_2-1} (1 - \theta[2 - \frac{1}{\theta_0}])^{-v} d\theta, \quad (6.3.5)$$

$$= C \mathcal{B}(u_1, u_2) {}_2F_1(v, u_1; u_1 + u_2; 2 - \frac{1}{\theta_0}). \quad (6.3.6)$$

The last equality follows from Euler's integral representation of the hypergeometric function (Abramowitz and Stegun, 1964, p. 558). \square

Hence, with the gamma prior $\zeta \sim \text{Gam}(\alpha., \beta)$ on the total timed rate as before, and a beta prior localised at $\theta_0 = (\frac{t_1}{t.})$ on the relative timed rate θ , we have the following Bayes factor for the two-sample Poisson problem:

$$\begin{aligned} \text{BF}_{10; \alpha_1, \alpha_2}(d) &= \frac{\mathcal{B}(\alpha_1 + y_1, \alpha_2 + y_2)}{\mathcal{B}(\alpha_1, \alpha_2) (\frac{t_1}{t.})^{y_1} (\frac{t_2}{t.})^{y_2}} \\ &\times {}_2F_1(\alpha_1 + \alpha_2, \alpha_1 + y_1; \alpha_1 + \alpha_2 + y_1 + y_2; \frac{t_1 - t_2}{t_1}) (\frac{t_2}{t_1})^{\alpha_1}. \end{aligned} \quad (6.3.7)$$

This is essentially Jeffreys's Bayes factor $\text{BF}_{10; \alpha_1, \alpha_2}^J(d)$, but with a correction factor for the localisation at $\frac{t_1}{t.}$.

6.3.1 Properties of the new Bayes factor $\text{BF}_{10;a}(d)$

Property 6.3.2 (Invariance). *With $\alpha_1 = \alpha_2 = a$ we have*

$$\text{BF}_{10;a}(d) = \text{BF}_{10;a}^J(d) \left(\frac{t_2}{t_1}\right)^a {}_2F_1(2a, a + y_1; 2a + y_1 + y_2; 1 - \frac{t_2}{t_1}) \quad (6.3.8)$$

$$= \text{BF}_{10;a}^J(d) \left(\frac{t_1}{t_2}\right)^a {}_2F_1(2a, a + y_2; 2a + y_1 + y_2; 1 - \frac{t_1}{t_2}) \quad (6.3.9)$$

which implies that this Bayes factor is invariant under relabeling and independent of the units for the times t_1 and t_2 . \diamond

Proof. Using Pfaff's transform (Gradshteyn and Ryzhik, 2007), we find

$${}_2F_1(2a, a + y_1; 2a + y_2; 1 - \frac{t_2}{t_1}) = \left(\frac{t_2}{t_1}\right)^{-2a} {}_2F_1(2a, a + y_2; 2a + n; 1 - \frac{t_1}{t_2}).$$

Multiplying both sides by $\text{BF}_{10;a}^J(d) \left(\frac{t_2}{t_1}\right)^a$ yields the assertion. \square

Property 6.3.3 (Limit-consistency). *Suppose that the data collection of the first process is halted at t_1 resulting in y_1 observations. Furthermore, let $y_2 = \lambda_2 t_2$ for some $\lambda_2 > 0$, then the Bayes factor $\text{BF}_{10;a}(d)$ converges to a limit as t_2 grows indefinitely. Thus,*

$$g_a(y_1, t_1, \lambda_2) = \lim_{t_2 \rightarrow \infty} \text{BF}_{10;a}(y_1, t_1, \lambda_2 t_2, t_2) \quad (6.3.10)$$

exists. For $a = 1$, the solution g_a can be well approximated by

$$\begin{aligned} \tilde{g}(y_1, t_1, \lambda_2) &= \log \Gamma(y_1 + a) + \lambda_2 t_1 - y_1 \log(\lambda_2 t_1) + a \log\left(\frac{\lambda_2}{t_1}\right) \\ &\quad + 2a \left[\frac{y_1 - a}{y_1 + \lambda_2 t_1 - a} + \log\left(\frac{t_1}{y_1 + \lambda_2 t_1 - a}\right) \right] \\ &\quad - (y_1 + a) \log\left(\frac{y_1 + \lambda_2 t_1 + a}{y_1 + \lambda_2 t_1 - a}\right) \\ &\quad - \frac{1}{2} \log\left(1 - \frac{8a(a + y_1)}{3a + \lambda_2 t_1 + y_1 + |y_1 + \lambda_2 t_1 - a|}\right) \end{aligned} \quad (6.3.11)$$

which we verified numerically. \diamond

Proof. To study the asymptotic behaviour of $\text{BF}_{10;a}(d)$ we consider Eq. (6.3.9) as the hypergeometric function with arguments smaller than one, thus, $1 - \frac{t_1}{t_2}$, are easier to handle as $t_1 \ll t_2$. We first provide some intuition.

Recall that the asymptotic behaviour of $\text{BF}_{10;a}^J(d)$ is given by

$$\log \text{BF}_{10;a}^J(d) \sim \log \Gamma(y_1 + a) - y_1 \log(\lambda_2 t_1) - a \log(\lambda_2 t_2 + a) + \lambda_2 t_1. \quad (6.3.12)$$

Hence, to show that $\log \text{BF}_{10;a}(d)$ stabilises, we have to show that the logarithms of the additional factors of Eq. (6.3.9), that is, $\left(\frac{t_1}{t_2}\right)^a$ and ${}_2F_1(2a, a + \lambda_2 t_2; 2a + n; 1 - \frac{t_1}{t_2})$ behave as $a \log(t_2)$ because this cancels out the bias-driving term $a \log(\lambda_2 t_2 + a)$. To see that this is possible, we consider the asymptotic behaviour of the hypergeometric function for the argument and the parameters separately.

6. A LIMIT-CONSISTENT BAYES FACTOR FOR TESTING THE EQUALITY OF TWO POISSON RATES

Suppose that the argument $z = 1 - \frac{t_1}{t_2}$ is fixed, then the parameters $v = a + \lambda_2 t_2$ and $w = 2a + y_1 + \lambda_2 t_2$ of the hypergeometric function ${}_2F_1(u, v; w; z)$ will be of the same order whenever t_2 is large (Temme, 2003). As a result, we obtain

$${}_2F_1(2a, a + \lambda_2 t_2; 2a + \lambda_2 t_2 + y_1; 1 - \frac{t_1}{t_2}) \approx (1 - z)^{-2a} = (\frac{t_1}{t_2})^{-2a}. \quad (6.3.13)$$

Multiplying both sides of Eq. (6.3.13) by $\text{BF}_{10; a}^J(d)(\frac{t_1}{t_2})^a$, taking the logarithm, and considering the asymptotic expansion with respect to t_2 shows that the bias-driving factor is adjusted to $-a \log(\lambda_2 t_1 + a \frac{t_1}{t_2})$, which suggests that $\log \text{BF}_{10; a}(d)$ indeed stabilises.

Similarly, suppose that the parameters $v = a + \lambda_2 t_2$ and $w = 2a + y_1 + \lambda_2 t_2$ are fixed, then for t_2 large, the argument of the hypergeometric function ${}_2F_1(u, v; w; z)$ can be taken to be one at the expense of a small approximation error. This roughly implies that for $y_1 > a$ we have

$${}_2F_1(2a, a + \lambda_2 t_2; 2a + y_1 + \lambda_2 t_2; 1) = \frac{\Gamma(2a + y_1 + \lambda_2 t_2)\Gamma(y_1 - a)}{\Gamma(y_1 + \lambda_2 t_2)\Gamma(y_1 + a)}, \quad (6.3.14)$$

whenever t_2 is large enough. Again, multiplying both sides by $\text{BF}_{10; a}^J(d)(\frac{t_1}{t_2})^a$ and writing out the beta function in $\text{BF}_{10; a}^J(d)$ then shows that

$$\text{BF}_{10; a}(d) \approx \mathcal{B}(\lambda_2 t_2 + a, y_1 - a)(\frac{t_1}{t_2})^a \quad (6.3.15)$$

whenever t_2 is large enough. An asymptotic expansion for t_2 large as in Prop. 6.2.1 then shows that

$$\log \text{BF}_{10; a}(d) \approx \log \Gamma(y_1 - a) + (a - y_1) \log(\lambda_2 t_1 + \frac{at_1}{t_2}) + \lambda_2 t_1 + \mathcal{O}(\frac{1}{t_2^2}), \quad (6.3.16)$$

which suggests that $\log \text{BF}_{10; a}(d)$ converges to a finite number as t_2 grows indefinitely.

For a rigorous proof of the result and the derivation of $g_a(y_1, t_1, \lambda_2)$, we used a Laplace approximation to the hypergeometric function ${}_2F_1(2a, a + \lambda_2 t_2; 2a + \lambda_2 t_2 + y_1; 1 - \frac{t_1}{t_2})$ as described by Butler and Wood (2002) at each fixed t_2 . We then used **Mathematica** to derive the limit which led to a function that spanned over four pages and therefore is not presented here.

By serendipity⁴ we were able to approximate the four-page equation by $\tilde{g}(y_1, t_1, \lambda_2)$ given above. Numerical experiments confirm that \tilde{g} approximates the true g_a well, see Table 6.1, and that g_a is in neighborhood of $\text{BF}_{10; a}(d)$ with t_2 large. The error that stands out occurs with $y_1 = 2$ and $t_1 = 5$, which leads to the correct limit $g_1(2, 5, 1) = 0.047$ and the approximated limit $\tilde{g}(2, 5, 1) = 0.019$.

□

The logarithm of the Bayes factor $\text{BF}_{10; a}(d)$ as a function of t_2 is depicted as the solid line in Fig. 6.1. In the left panel, an exact calculation shows that

⁴The replacement of $4ax(c - b)$ by $4a(c - a)$ in the definition of \hat{g} in Butler and Wood (2002, p. 1164).

Table 6.1: With $\lambda_2 = 1$ and relative error $\frac{g_1(y_1, t_1, \lambda_2) - \tilde{g}(y_1, t_1, \lambda_2)}{g_0(y_1, t_1, \lambda_2)}$ in percentage.

	$t_1 = 2$	$t_1 = 5$	$t_1 = 10$	$t_1 = 25$	$t_1 = 100$	$t_1 = 250$
$y_1 = 2$	-3.1	58.3	0.4	3e-2	4e-4	3e-5
$y_1 = 5$	-5.6	-0.1	4.4	2e-2	4e-4	3e-5
$y_1 = 10$	-5e-2	-0.8	-1e-2	2e-2	4e-4	3e-5
$y_1 = 25$	-1e-3	-3e-3	-7e-3	-7e-4	3e-4	2e-5
$y_1 = 100$	-3e-6	-8e-6	-2e-5	-5e-5	-8e-6	2e-5
$y_1 = 250$	-5e-8	-2e-7	-4e-7	-1e-6	-5e-6	-5e-7

Table 6.2: With $\lambda_2 = 5$ and relative error $\frac{g_1(y_1, t_1, \lambda_2) - \tilde{g}(y_1, t_1, \lambda_2)}{g_0(y_1, t_1, \lambda_2)}$ in percentage.

	$t_1 = 2$	$t_1 = 5$	$t_1 = 10$	$t_1 = 25$	$t_1 = 100$	$t_1 = 250$
$y_1 = 2$	-0.4	3e-2	3e-3	2e-4	2e-4	3e-6
$y_1 = 5$	4.4	2e-2	3e-3	2e-4	3e-6	2e-7
$y_1 = 10$	-1e-2	2e-2	3e-3	2e-4	3e-6	2e-7
$y_1 = 25$	-7e-3	-7e-4	3e-3	2e-4	3e-6	2e-7
$y_1 = 100$	-2e-5	-5e-5	-1e-4	-5e-3	2e-6	2e-7
$y_1 = 250$	-4e-7	-1e-6	-2e-6	-7e-7	2e-6	2e-7

Table 6.3: With $\lambda_2 = 0.001$ and relative error $\frac{g_1(y_1, t_1, \lambda_2) - \tilde{g}(y_1, t_1, \lambda_2)}{g_0(y_1, t_1, \lambda_2)}$ in percentage.

	$t_1 = 2$	$t_1 = 5$	$t_1 = 10$	$t_1 = 25$	$t_1 = 100$	$t_1 = 250$
$y_1 = 2$	-4e-2	-0.12	-0.27	-0.7	-3.2	-7.6
$y_1 = 5$	-3e-4	-9e-4	-2e-3	-6e-3	-3.2	0.1
$y_1 = 10$	-1e-5	-4e-5	-9e-5	-3e-4	-1e-3	-4e-3
$y_1 = 25$	-3e-7	-7e-7	-2e-6	-5e-6	-2e-5	-7e-5
$y_1 = 100$	-8e-10	-2e-9	-5e-9	-1e-8	-7e-8	-2e-7
$y_1 = 250$	-2e-11	-5e-11	-1e-10	-3e-10	-1e-9	-2e-9

$g_1(1, 1, 5) = 1.21$, whereas the approximation yields $\tilde{g}(1, 1, 5) = 1.15$; this means that when the first process yields $y_1 = t_1 = 1$ and stops, the evidence in favour of the alternative hypothesis is then bounded by $\text{BF}_{10;1}(d) \leq e^{1.21} \approx 3.35$. For the case depicted in the right panel, an exact calculation shows that $g_1(1, 1, 1) = -0.63$, whereas the approximation yields $\tilde{g}(1, 1, 1) = -0.90$; this means that when the first process yields $y_1 = t_1 = 1$ and stops, the evidence in favour of the null hypothesis is then bounded by $\text{BF}_{01;1}(d) \leq e^{0.63} \approx 2.46$.

6.4 Discussion

We proposed the new desideratum of *limit-consistency* that can help guide the specification of prior distributions for tests that involve a comparison of two or

6. A LIMIT-CONSISTENT BAYES FACTOR FOR TESTING THE EQUALITY OF TWO POISSON RATES

more processes or groups. As a concrete illustration of the added value of the desideratum we rederived Jeffreys's (1939) Bayes factor $\text{BF}_{10;a}^J(d)$ for the two-sample Poisson problem. We proved that this Bayes factor is not limit-consistent: when t_1 is fixed and t_2 grows indefinitely, the Bayes factor increasingly supports the null, regardless of the data. This implies that researchers who use Jeffreys's Bayes factor for the comparison of two Poisson rates can bias the evidence in favour of the null by selectively investing resources in data collection for one of the two processes. We then proposed a generalisation of Jeffreys's test that eliminates the bias-driving term; consequently, this localised Bayes factor is limit-consistent while retaining the positive features of Jeffreys's original test.

The proof of limit-consistency can perhaps be sharpened – a four-page long definition of the limit function $g_a(y_1, t_1, \lambda_2)$ based on **Mathematica** output is not intuitive, and the serendipitous approximation $\tilde{g}(y_1, t_1, \lambda_2)$ may warrant more research. Insights might be gained from the fact that the unnormalised posterior for θ with the localised beta prior as a function of t_2 is either log-concave or log-convex, depending on y_1, t_2 and λ_2 . Additional insight might be acquired from studying the differential equation corresponding to the hypergeometric function in the Bayes factor, or other saddle points methods as hinted at by Cvitković et al. (2017), López and Pagola (2011), and Temme (2003).

In our derivation of the localised Bayes factor $\text{BF}_{10;a}(d)$ we extended Jeffreys's proposal of using a beta distribution on the test-relevant parameter θ by adding a location parameter on the logit scale. By representing the problem this way we could use the methods and intuitions that Jeffreys developed for his Bayesian *t*-test (Jeffreys, 1948, pp. 242–248; Ly et al., 2016a, 2016b).

We believe that the localised Bayes factor $\text{BF}_{10;a}(d)$ for comparing two Poisson rates is consistent with Jeffreys's general philosophy of testing – more so, perhaps, than Jeffreys's own proposal from 1939. The desideratum of limit-consistency appears logical and compelling, and we hope that it can be helpful in a broad range of discrete data problems.

Part III

Scientific Learning with Bayes Factors

Chapter 7

Four Requirements for an Acceptable Research Programme

Abstract

In a recent article for *Basic and Applied Social Psychology*, Witte and Zenker (2016) proposed a research strategy that rests on the sequential evaluation of a point-alternative hypothesis. At first a large study is used to determine a “specific theoretical effect size” and then, in a series of follow-up studies, this estimated effect size is contrasted against an effect size of zero. The authors deem this strategy “free of various deficits that beset dominant strategies (e.g., meta-analysis, Bayes factor analysis)” and argue that its broad adoption constitutes “one way in which the confidence crisis may be overcome”.

We disagree with their research strategy as it does not go far enough. One should avoid hindsight bias and acknowledge uncertainty that comes with scientific learning. The four requirements given here provide the context in which Bayes factors can help empirical scientists learn from data.

Keywords: Crisis of confidence, exploratory versus confirmatory research, scientific learning.

We agree with Witte and Zenker (2016) that it can be useful to test an alternative hypothesis that is constructed, in part or in whole, from earlier data (e.g., Verhagen and Wagenmakers, 2014; Wagenmakers et al., 2016c). We also agree that it can be informative to take into account a sequence of studies as it unfolds over time (e.g., Scheibehenne et al., 2016). In this comment, however, we focus mainly on areas of disagreement, which centre on what we believe to be mistakes and omissions. First we address the mistakes and discuss how, in our opinion, Witte

This chapter is published as Marsman, M., Ly, A., & Wagenmakers, E.-J. (2016). Four requirements for an acceptable research programme. *Basic and Applied Social Psychology*, 38(6), 308–312. doi: <http://dx.doi.org/10.1080/01973533.2016.1221349>.

and Zenker (2016) fell prey to two common fallacies: the power fallacy and the fallacy of the transposed conditional. Even for experienced scholars, these fallacies may be difficult to recognise. Second, we address the omissions and discuss four requirements for an acceptable research programme.

7.1 The power fallacy

On repeated occasions, Witte and Zenker (2016) lament the lack of statistical power while at the same time boasting about the strength of statistical evidence. This confused interpretation of the data can be overcome by recognising that power and evidence are inherently different concepts. Before we start, let's take for granted that the desired test is between $\mathcal{H}_0 : \delta = 0$ versus a point-alternative $\mathcal{H}_1 : \delta = 0.30$.

Now power is a pre-data concept, a metric constructed by averaging across all possible data sets that could be obtained in the envisioned experiment. *A priori* and *on average* –with respect to all possible data sets– experiments designed with low power are unlikely to yield a significant outcome given that \mathcal{H}_1 is true. In contrast, evidence is a post-data concept. In this specific scenario it is given by the likelihood ratio, the relative probability of the observed data under the competing hypotheses. The likelihood ratio considers only the data that have in fact been obtained.

As discussed elsewhere in detail, after the data have been observed, data that could have been observed –but were not– are evidentially irrelevant (e.g., Berger and Wolpert, 1988; Bayarri et al., 2016; Wagenmakers et al., 2015a; Wagenmakers et al., 2017c). Basically, our pre-data state of knowledge has changed by the observation of the data, and after the data have arrived our post-data state of knowledge is all that ought to matter.

When the pre-data concept of power is erroneously used for post-data purposes –such as inference and the quantification of evidence–, this entails a deliberate loss of important information, namely the actual outcomes of the experiment.

7.2 The fallacy of the transposed conditional

Witte and Zenker (2016) correctly point out that the Bayes factor is the probability of the data under \mathcal{H}_0 versus \mathcal{H}_1 (Wagenmakers et al., 2016b). They also acknowledge that the Bayes factor and the likelihood ratio are “quantitatively” equivalent whenever the hypotheses are both simple (i.e., consisting of a single specified point value for effect size). However, Witte and Zenker (2016) argue that by simply changing the nomenclature¹ –from Bayes factors to likelihood ratios– allows them to interpret the likelihood ratio as the relative plausibility of the hypotheses. So even though what is calculated is the relative probability of the data given the hypotheses, the result is interpreted as the relative probability of the hypotheses given the data. By doing so Witte and Zenker (2016) commit the fallacy of the transposed conditional.

¹ “What's in a name? That which we call a rose by any other name would smell as sweet” – Juliet, Act 2 Scene 2

Unfortunately, in statistical inference there is no such thing as a free lunch (Rouder et al., 2016a). Any time one wishes to assign probabilities to parameters or models, one is automatically committed to the Bayesian framework (Ly et al., 2016a, 2016b). Specifically, the only way to obtain a posterior probability is by using the data to update a prior probability. Bayes factors quantify the extent to which the data change the prior model odds to posterior model odds, and as such they can be considered the relative evidence that the data provide for the models under consideration. The Bayes factor is therefore only one ingredient for inference. The other ingredient is the prior model odds. One is licensed to interpret Bayes factors (or likelihood ratios, for simple models) as posterior odds, but only when the prior odds equals 1, and *not* when the prior odds is ignored.

To appreciate the importance of the prior odds, consider the competing models \mathcal{H}_1 : “people have extra-sensory perception (ESP)” versus \mathcal{H}_0 : “people do not have ESP”. Few researchers would seriously entertain equal prior odds in this case. Moreover, suppose the likelihood ratio for an ESP experiment yielded a factor of 30 in favour of ESP; do we conclude from this that the ESP hypothesis is 30 times more likely than the null hypothesis? Of course we do not, and if the authors methodology were to sanction this inference (which it does not), then this would be a compelling argument against their methodology instead of a compelling argument for ESP. Extraordinary claims require extraordinary evidence, and in order to assess the posterior plausibility of ESP one needs to combine the evidence from the data (i.e., the Bayes factor) with the prior plausibility of the ESP phenomenon (Wagenmakers et al., 2015b).

7.3 Requirements of a research programme

A research programme that can cure the current “crisis of confidence” (Pashler and Wagenmakers, 2012) needs to be more ambitious than the approach proposed by Witte and Zenker (2016). Below we outline four key requirements and point to the relevant literature.

7.3.1 I. Preregistration

Philosophers, psychologists, physicians, and physicists have long argued that empirical research needs to respect the distinction between work that is exploratory or hypothesis-generating and work that is confirmatory or hypothesis-testing, and that this needs to be done by preregistering the analysis plan in all of its details (e.g., Barber, 1976; Chambers, 2013; Feynman, 1998; Goldacre, 2009; Peirce, 1878,8; Wagenmakers et al., 2012).

These theoretical arguments have garnered empirical support in the sense that preregistered replications rarely support the original effects (e.g., Nosek and Lakens, 2014; Open Science Collaboration, 2012). Without preregistration, researchers can easily and unwittingly fall prey to hindsight bias and confirmation bias. In our opinion, any research programme that does not include preregistration is seriously incomplete.

7.3.2 II. Transparency

In reproducible research, transparency is essential. Indeed, one can argue that preregistration falls under the general heading of transparency as well. Here we use transparency to refer to open materials, open data, and open analysis code. Recent initiatives such as TOP (Transparency and Openness Promotion, Nosek et al., 2015), PRO (The Peer Reviewers' Openness Initiative, Morey et al., 2016), and the Center for Open Science badges for good academic behaviour (Kidwell et al., 2016) aim and change the dominant culture so that openness becomes the norm, not the exception.

In our own work, we have developed the open-source statistical software program JASP (jasp-stats.org; JASP Team, 2017). In JASP, users can save data, analysis input, analysis output, and analysis annotations in a single .jasp file.² When this file is uploaded to the Open Science Framework, the OSF JASP pre-viewer allows anybody with an online browser to inspect the annotated output, even without having JASP installed.

7.3.3 III. Comprehensive knowledge updating

A mature research programme allows knowledge to be updated as new data come in (Scheibehenne et al., 2016). This requirement is fulfilled by Witte and Zenker (2016), but only in part: what is updated is the likelihood ratio, but not the value of the parameter. In other words, based on the initial study, Witte and Zenker (2016) committed themselves 100% to the single point estimate $\delta = 0.30$. This violates what Lindley termed “Cromwell’s rule”. Cromwell famously told the Church of Scotland “I beseech you, in the bowels of Christ, think it possible you may be mistaken”. Cromwell’s rule states that one should not categorically rule out anything, for this makes it impossible to learn. As explained by Lindley (1985), “So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved.”

Occasionally there are reasons to violate Cromwell’s rule. For instance, one may wish to evaluate the relative adequacy of the predictions from a theoretically meaningful hypothesis – perhaps a general law or invariance (Rouder et al., 2009), or perhaps a physical law involving gravity or the speed of light. In the current example, however, the point estimate of 0.30 is devoid of theoretical content; the effect size could differ from one context to the next, or it could be lower or higher. The original data set suggested $\delta = 0.30$, but what if a second, much larger data set³ had suggested $\delta = 0.10$? This value is still consistent with the general theory of there being an effect, only it is a little smaller than suggested in the original study. The likelihood ratio would have favoured \mathcal{H}_0 , but at the same time it would be obvious that \mathcal{H}_0 is false. This is the equivalent of the Lindley’s astronaut scenario.

²The analysis output may also be saved separately.

³For concreteness and to avoid ambiguity, let’s say one thousand times as large.

The correct way to update knowledge is to update both the plausibility of competing models and the plausibility of the parameters within those models.⁴ This implies that we also need priors on the parameters within the models. The introduction of these priors have led to much debate in the statistical community at first, as they were perceived as highly subjective. However, it has since been mathematically proved that the influence of the prior on the posterior washes out easily with enough data (e.g., Bickel and Kleijn, 2012; Kleijn and van der Vaart, 2012; van der Vaart, 1998) for the regular models typically used in the psychological sciences. As such, rather than using a point estimate of the parameter from a first data set as a point alternative hypothesis, we propose to use the posterior of the effect size instead. By using the posterior as a prior in the next study, we incorporate all the relevant information from the first data set for inference in a next experiment. Hence, subjectivity simply refers to the incorporation of previously collected data rather than an opinion. This method of extracting information from one study to another is further explored in Ly et al. (2017b), Verhagen and Wagenmakers (2014), and Wagenmakers et al. (2016c), and by doing so we adhere to the laws of probability. Hence, our proposition of using Bayesian methods leads to a principled method of learning. Moreover, it automatically gives us posteriors that can be readily used to quantify the uncertainty of our inference.

7.3.4 IV. Acknowledging uncertainty

In our experience, researchers strongly desire unambiguous yes/no answers, even when these are unavailable due to the stochastic nature of the data. Paradoxically, the noisier the data, the stronger this desire seems to become.

The decision-making framework of null hypothesis significance testing (NHST) offers some certainty: if $p < .05$, we may “reject the null hypothesis”. This is fulfilling, because by making a decision we have swept all of the existing uncertainty under the rug. There is no more need to debate the outcome any longer, the researcher may feel, because we were sanctioned to make a Decision to Reject the Null Hypothesis. After the Gordian knot has been cut, it is futile to argue about other possible decisions that could have been made. This way, NHST offers an illusion of certainty, and with it the protection again critique and self-doubt.

Unfortunately there are several problems with the decision-making framework of null hypothesis significance testing. The list is endless, but here we highlight the following concerns:

1. Utilities are ignored. If the purpose of statistical inference in academia is to make decisions, then one needs to specify utilities or loss functions associated to the potential outcomes (e.g., Lindley, 1985). Without utilities there can be no sensible decision making.
2. Scientists often do not make decisions. One of our favourite quotations is from Rozeboom (1960, p. 420): “The null-hypothesis significance test treats

⁴Point hypotheses are a good approximation to posterior distributions that are highly peaked, but in the case of Witte and Zenker (2016) we see no compelling reason in this case to violate Cromwell’s rule and update knowledge only partially.

acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data—i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.*”

3. The *p*-value from the framework of null hypothesis significance testing –upon which the Decision to Reject the Null Hypothesis is based– is “violently biased against the null hypothesis.” (Edwards, 1965, p. 400; see also Berger and Delampady, 1987; Edwards et al., 1963; Johnson, 2013; Marsman and Wagenmakers, 2017; Sellke et al., 2001; Wetzels et al., 2011). For these and other reasons we sympathise with the *p*-value ban in *Basic and Applied Social Psychology* (Trafimow and Marks, 2015).

Instead of using ad-hoc decision rules for seeking certainty where there is none, it is better to acknowledge and quantify uncertainty. If a Bayes factor indicates that the data are 4 times more likely under \mathcal{H}_1 than under \mathcal{H}_0 , this does not mean that \mathcal{H}_0 has been refuted, or that \mathcal{H}_1 is true. Authors should make claims that are in accordance with the strength of evidence in the data – often, this means that the claims should be more modest. In turn, editors and reviewers should reward such modesty, not punish it.

7.4 Concluding comments

We proposed four requirements for an acceptable research programme, which we believe to be at odds with Witte and Zenker’s (2016) proposal. Specifically, their proposal fails to acknowledge uncertainty and does not result in coherent knowledge updates. This is because Witte and Zenker (2016) sweep the prior model probabilities under the rug and violate the laws of probability by using an intermediate estimate as a point alternative hypothesis. Moreover, by using a point alternative hypothesis, Witte and Zenker (2016) ignore the uncertainty with which the alternative was specified.

For comprehensive knowledge updating, that is, statistical learning, we have to adhere to the laws of probability, the same way the motion of stars has to obey the laws of physics. Our advocacy for Bayesian methods in psychology is, in essence, a call to adopt a principled method of learning. This call is neither new nor controversial, as Bayesian methods have been adopted in fields such as econometrics, statistics and computer science with great success.

The reward for adopting Bayesian methods in psychology is substantial: not only do our conclusions adhere to the laws of probability, but we also obtain automatic uncertainty quantification in terms of posterior distributions. These posteriors provide a full summary of the previous data sets and can be transformed into so-called posterior predictives which give an indication of how our previous findings generalise to new experiments (Liu and Aitkin, 2008). The posterior predictive as a measure of replicability will be better in predicting future

outcomes compared to Witte and Zenker's (2016) approach as was shown in Wagenamakers et al. (2006). Their loss in performance is due to their commitment to a single point alternative hypothesis, thus, their disregard of the uncertainty in their intermediate step and, therefore, their violation of the laws of probability.

The proposed four requirements for an acceptable research programme are relatively straightforward to execute, but they imply that researchers acknowledge and counteract fundamental human biases and desires. Implementing the programme therefore requires a change in academic culture. Academic culture is difficult to change, but the past five years have demonstrated that it can be done. Driven by the combined efforts from researchers, journals, funders, and institutes (especially the Center for Open Science), there has been a dramatic and positive reorientation of academic values. The caterpillar known as psychological science has finally started its metamorphosis, and only the future will show whether the butterfly is willing to learn from the data that were actually observed.

Chapter 8

Bayesian Reanalyses from Summary Statistics: A Guide for Academic Consumers

Abstract

Across the social sciences, researchers have overwhelmingly used the classical statistical paradigm to draw conclusions from data, often focusing heavily on a single number: p . Recent years, however, have witnessed a surge of interest in an alternative statistical paradigm: Bayesian inference, in which probabilities are attached to parameters and models. We feel it is informative to provide statistical conclusions that go beyond a single number, and –regardless of one’s statistical preference– it can be prudent to report the results from both the classical and the Bayesian paradigm. In order to promote a more inclusive and insightful approach to statistical inference we show how the open-source software program JASP (jasp-stats.org) provides a set of comprehensive Bayesian reanalyses from just a few commonly-reported summary statistics such as t and N . These Bayesian reanalyses allow researchers –and also editors, reviewers, readers, and reporters– to quantify evidence on a continuous scale, assess the robustness of that evidence to changes in the prior distribution, and gauge which posterior parameter ranges are more credible than others. The procedure is illustrated using the seminal Festinger and Carlsmith (1959) study on cognitive dissonance.

Keywords: Bayes factor, data visualisation, effect size, hypothesis testing, p -value.

This chapter is submitted for publication and also available as PsyArXiv preprint: <https://osf.io/7dzmk> as: Ly, A., Raj, A., Marsman, M., Etz, A., & Wagenmakers, E.-J. (2017). Bayesian reanalyses from summary statistics: A guide for academic consumers.

8.1 Introduction

Classical null hypothesis statistical testing (NHST) allows researchers to evaluate scientific propositions in a seemingly straightforward manner: whenever the p -value falls below a threshold α (usually set to .05) researchers feel licensed to reject the null hypothesis that the effect is absent and embrace the alternative hypothesis that the effect is present. For example, in the results section one may encounter conclusions such as “overall classification accuracy was greater than chance”, “the analysis revealed a main effect of the manipulation”, and “the correlation was significant”; in the discussion section, these statements are abstracted from the standard NHST framework even further, conveying the impression that whenever $p < .05$, the data strongly favour the alternative hypothesis over the null hypothesis of no effect.

The field’s mechanistic use of p -values appears to be at odds with the recent warning issued by the *The American Statistical Association* (ASA; Wasserstein and Lazar, 2016, p. 131): “The widespread use of ‘statistical significance’ (generally interpreted as ‘ $p \leq 0.05$ ’) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.” Indeed, p -values have been critiqued on numerous grounds (e.g., Nickerson, 2000; Rouder et al., 2016a; Wagenmakers et al., 2017b). One widely appreciated concern is that p -values do not convey information about the size of the effect or the precision with which that effect is estimated (e.g., Cumming, 2014).

As one prominent alternative to p -value NHST, recent years have seen an increased interest in Bayesian inference (Vandekerckhove et al., 2017; Wagenmakers et al., 2016b), a paradigm in which prior uncertainty about parameters and models is updated by means of observed data to yield posterior uncertainty. Specifically, the posterior distribution quantifies the information about the effect size under the alternative hypothesis, whereas the Bayes factor quantifies the predictive adequacy of the null hypothesis as compared to the predictive adequacy of an exactly-specified alternative hypothesis (e.g., Etz and Wagenmakers, 2017; Jeffreys, 1961; Kass and Raftery, 1995; Myung and Pitt, 1997).

A discussion on the merits and demerits of the different statistical paradigms is beyond the scope of this paper. We agree with the ASA’s recommendation to go beyond p , and that it is prudent to adopt an inclusive statistical approach. For when the results of different statistical paradigms point in the same direction, this bolsters one’s confidence in the conclusions, but when the results are in blatant contradiction, this will weaken one’s confidence.

In the spirit of promoting a more inclusive statistical approach, our primary goal is to demonstrate the ease with which published classical results can be subjected to a Bayesian reanalysis using the recently developed “Summary Stats” module in JASP (JASP Team, 2017). Depending on the analysis at hand, this module takes as input commonly-reported statistics such as t , r , and R^2 together with sample size N , and returns a comprehensive Bayesian assessment.¹ Importantly, this Bayesian assessment can be executed in the absence of the raw data.

¹The website <http://pcl.missouri.edu/bayesfactor>, designed and maintained by Jeff Rouder, exploits the same idea, but focuses exclusively on the Bayes factor.

This is essential when the data are no longer available or when they cannot be shared; but even when the raw data are publicly available, the analysis presented here is much more efficient – reviewers, readers, and reporters can obtain a comprehensive Bayesian assessment almost instantaneously. We believe that the richness of the Bayesian report contrasts favourably with a report of just the summary statistics themselves. We illustrate this claim using a seminal study published more than half a century ago.

8.2 The Festinger & Carlsmith (1959) cognitive dissonance study

In a landmark publication,² Festinger and Carlsmith (1959, hereafter FC) outlined a theory to account for *cognitive dissonance*, a phenomenon they described as follows: “If a person is induced to do or say something which is contrary to his private opinion, there will be a tendency for him to change his opinion so as to bring it into correspondence with what he has done or said” (p. 209). Early experiments on cognitive dissonance (e.g., Kelman, 1953) induced participants to make a statement contrary to their personal opinion for the chance to gain a reward. It was hypothesised that for greater rewards there would be a greater change to the opinion, but the data showed the reverse: the smaller the reward, the greater change in opinion. FC proposed a theory that could account for this behavioural pattern, which they subsequently put to the test in an ingenious experiment.

FC’s experiment included control, high reward, and low reward conditions, each with twenty participants. All participants performed a boring task for one hour, after which they were asked to take a survey and answer questions about, among other things, their enjoyment of the study. Where the conditions differ is what happens after completing the boring task, but before completing the survey. In the reward conditions, participants were asked to interact with a confederate by telling them that the experiment was interesting and fun; for this they received either twenty dollars (high reward) or one dollar (low reward). In the control condition participants went straight to the post-interview and did not interact with the confederate. According to FC, the crucial test of their theory lies in comparing the post-interview enjoyment ratings from the low versus high reward conditions, where the low reward condition is predicted to have higher enjoyment ratings. In line with their theory’s prediction, FC found a higher mean enjoyment rating in the low reward group than in the high reward group, $t(38) = 2.22$, $p = .032$, and this was taken as support for their theoretical position. No effect size is reported in the original paper but this can be easily computed from the t -value and degrees of freedom, giving $d = 0.720$.

²Cited over 3,300 times according to Google Scholar, May 19, 2017.

8. BAYESIAN REANALYSES FROM SUMMARY STATISTICS: A GUIDE FOR ACADEMIC CONSUMERS

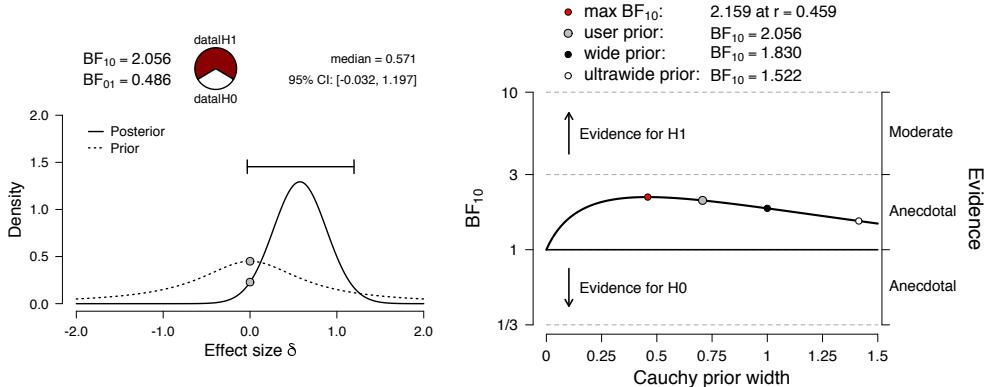


Figure 8.1: A comprehensive Bayesian reanalysis of the seminal study by Festinger and Carlsmith (1959), obtained by entering $t = 2.22$ and $N_1 = N_2 = 20$ into the JASP Summary Stats module. See text for details.

8.3 Bayesian reanalysis

We wish to conduct a Bayesian reanalysis of the FC result, but the raw data from this study are no longer available. However, the Summary Stats module in JASP affords a comprehensive Bayesian reanalysis using only the test statistic reported in the original paper.³ Inputting the reported t -value and sample sizes for the two groups yields the results shown in Fig. 8.1.

In the left panel, the dotted line represents the prior distribution for effect size under \mathcal{H}_1 : a zero-centred *Cauchy* distribution (i.e., a t -distribution with one degree of freedom; Jeffreys, 1948; Ly et al., 2016a, 2016b), here with interquartile range set to a default value of $r = 0.707$ (e.g., Morey and Rouder, 2015; for a larger family of informed prior distributions see Gronau et al., 2017a). Thus, under \mathcal{H}_1 –that is, assuming the effect is present– the expectation is that the effect is most likely to be small, although the possibility that it is large is not ruled out.

In the left panel, the solid line is the posterior distribution for effect size, that is, the knowledge about effect size obtained after updating the prior distribution using the observed data, and assuming that \mathcal{H}_1 holds. This posterior distribution has a median of 0.571,⁴ and a relatively wide 95% central credible interval that ranges from -0.032 to 1.197 –in other words, 95% of the posterior mass lies in the interval from -0.032 to 1.197 ; clearly, the effect has not been estimated with much precision. More generally, by computing the area under the posterior distribution between $\delta = a$ and $\delta = b$, one can assess how plausible it is that δ falls in the interval from a to b (e.g., Wagenmakers et al., 2016b; Wagenmakers et al., 2017a). For instance, by comparing the area under the posterior distribution to the right of zero against that to the left of zero, we quantify how much more likely it is

³The Summary Stats module is activated via the options menu located in the top right corner of the JASP window.

⁴Note that the prior distribution has shrunk the sample value of $d = 0.720$ toward zero.

that the effect is positive rather than negative, under \mathcal{H}_1 – that is, under the presumption that the effect is present.

In general, the posterior distribution quantifies all that we know about effect size δ , given that \mathcal{H}_1 holds and the effect exists. The latter point is worth emphasising since it has been argued that one may perform a Bayesian null hypothesis test by judging whether the 95% credible interval overlaps with zero. Despite its beguiling simplicity, such a procedure is incorrect (Berger, 2006; Jeffreys, 1961), because it begs the question – the extent to which a null hypothesis is plausible cannot be assessed when this hypothesis has been ruled out in advance (i.e., under the continuous prior distribution assumed by \mathcal{H}_1 , the probability of any single point such as $p(\delta = 0)$ equals zero).

In order to perform a Bayesian hypothesis test, one needs to compare the predictive performance of the null hypothesis \mathcal{H}_0 against that of the alternative hypothesis \mathcal{H}_1 . The result of this comparison is known as the Bayes factor, and the left panel of Fig. 8.1 reveals that it equals 2.056 – that is, the observed FC data are only about twice as likely under \mathcal{H}_1 than under \mathcal{H}_0 . Arch-Bayesian Harold Jeffreys deemed this level of evidence “not worth more than a bare mention” (Jeffreys, 1961, p. 432). The proportion wheel on top visualises the strength of the evidence.⁵

The Bayes factor quantifies relative predictive performance, and the predictive performance from \mathcal{H}_1 is determined in part by the prior distribution. Under a default prior specification, it is natural to wonder how robust the conclusions are to plausible changes in the prior distribution. To address this issue, the Summary Stats module allows one to select the option “Bayes factor robustness check”. The right panel of Fig. 8.1 shows the result: the Bayes factor as a function of the interquartile range r of the Cauchy prior distribution. The values range from $r = 0$ (when \mathcal{H}_1 reduces to \mathcal{H}_0 and the Bayes factor is 1 regardless of the data) to $r = 1.5$. Across this entire range, the Bayes factor never exceeds 3; in fact, the maximum Bayes factor in favour of \mathcal{H}_1 equals 2.159, obtained when the width r is set to 0.459.

In this particular scenario we find that a seminal result, significant with a p -value of .032, does not yield compelling evidence against \mathcal{H}_0 when assessed from a default Bayesian perspective.⁶ Even though the evidence against \mathcal{H}_0 is relatively inconclusive, the posterior distribution can nevertheless be used as a prior in further studies, and allows one to compute the so-called replication Bayes factor (Ly et al., 2017b; Verhagen and Wagenmakers, 2014).

In sum, the Bayesian reanalyses shown in Fig. 8.1 are easily obtained in JASP and paint an inferential picture more complete than the one provided by the statement “ $t(38) = 2.22, p = .032$ ”.

⁵See also <https://osf.io/3acm7/>.

⁶For a further discussion of the FC results, see <https://mattiheino.com/2016/11/13/legacy-of-psychology/>.

8.4 Concluding comments

The Summary Stats module in JASP unlocks a comprehensive Bayesian experience from a few commonly-reported summary statistics. Here we illustrated the module for the case of an independent-samples t -test, but the Summary Stats module can also be used for inference concerning paired-samples t -tests, correlation coefficients, binomial proportions, and linear regression models. An entire literature filled with classical statistics is now open for a straightforward Bayesian reanalysis.

Two remarks are in order. First, even when the summary statistics are “sufficient” (i.e., they capture all relevant information) on general grounds it is still beneficial to have access to the raw data. The raw data can be used to confirm that the statistical model is appropriate, the desirability of which is vividly displayed by Anscombe’s quartet (e.g., Anscombe, 1973; Matejka and Fitzmaurice, 2017).⁷

Second, the Bayesian analyses discussed above are “objective” or “uninformative” in the sense that under \mathcal{H}_1 , the prior distributions for effect size are centred around zero, the value specified by \mathcal{H}_0 . However, the Bayesian framework can be extended to include *informed* prior distributions – these distributions incorporate context-specific expectations and need not be centred around zero (Gronau et al., 2017a). We plan to add the extensions to informed priors to JASP in the near future. Just like the reanalysis with objective priors, the reanalysis with informed priors is a function solely of the summary statistics.

In closing, the Bayesian reanalyses outlined here provide an opportunity to expand summary statistics to statements about posterior distributions and Bayes factors. This expansion affords (1) an additional inferential perspective that supplements the classical perspective; (2) a reanalysis of published findings without requiring the raw data, and (3) a highly efficient method for editors, reviewers, readers, and reporters to gauge whether the conclusions from a different statistical paradigm contradict or confirm the classical conclusions. We hope that this reanalysis will spur a more graded assessment of statistical evidence and a reporting of statistical outcome measures that is both comprehensive and inclusive.

⁷See also Alberto Cairo’s Anscombusaurus at <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>.

Chapter 9

Replication Bayes Factors from Evidence Updating

Abstract

We describe a general method that allows experimenters to quantify the evidence from the data of a direct replication attempt given data already acquired from an original study. These so-called replication Bayes factors are a reconceptualisation of those introduced by Verhagen and Wagenmakers (2014) for the common t -test. This reconceptualisation is computationally simpler and generalises easily to most common experimental designs for which Bayes factors are available.

Keywords: Evidence synthesis, hypothesis testing, meta-analysis, replication.

9.1 Introduction

The past five years have witnessed a dramatic increase in interest for replication studies, largely in response to psychology’s “crisis of confidence” (e.g., Pashler and Wagenmakers, 2012). While this crisis is not unique to the field of psychology by any means, psychologists have been at the forefront of efforts to assess and improve reproducibility in science by way of large-scale replication initiatives, such as the Reproducibility Project: Psychology (Open Science Collaboration, 2015), the *Social Psychology* special issue on replication (Nosek and Lakens, 2014), and the various ManyLabs efforts (Ebersole et al., 2016; Klein et al., 2014). Although the importance of direct replication has been contested by some (for an overview of the most common arguments see Zwaan et al., 2017), the increasing prominence

This chapter is submitted for publication and also available as PsyArXiv preprint: osf.io/preprints/psyarxiv/u8m2s as: Ly, A., Etz, A., Marsman, M. and Wagenmakers, E.-J. (2017). Replication Bayes Factors from Evidence Updating.

of replication studies has prompted researchers to examine the question of how to assess, statistically, the degree to which a replication study succeeds or fails.

A number of complementary questions may arise when evaluating replication studies:

1. Completely ignoring the data of the original study, what is the evidence that the effect is present or absent in the replication attempt? (e.g., Marsman et al., 2017).
2. Taking the data of the original study fully into account, what is the evidence that the effect is present or absent in the replication attempt? (e.g., Verhagen and Wagenmakers, 2014).
3. Pooling the data from the original study and the replication attempt, what is the evidence that the effect is present or absent? (e.g., Scheibehenne et al., 2016).
4. Comparing the data from the original study and the replication attempt, what is the evidence that the effect sizes are similar or dissimilar? (e.g., Bayarri and Mayoral, 2002).

Here we focus on answering the second question using the “replication Bayes factor”, which can be conceptualised as contrasting the position of a hypothetical skeptic and proponent:

“The 1st hypothesis is that of the skeptic and holds that the effect is spurious; this is the null hypothesis that postulates a zero effect size, $\mathcal{H}_0 : \delta = 0$. The 2nd hypothesis is that of the proponent and holds that the effect is consistent with the one found in the original study, an effect that can be quantified by a posterior distribution. Hence, the 2nd hypothesis –the replication hypothesis– is given by $\mathcal{H}_r : \delta \sim \text{‘posterior distribution from original study.’}$ The weighted-likelihood ratio [i.e., the replication Bayes factor] between \mathcal{H}_0 and \mathcal{H}_r quantifies the evidence that the data provide for replication success and failure.” (Verhagen and Wagenmakers, 2014, p. 1457)

Verhagen and Wagenmakers (2014) proposed this replication Bayes factor in the context of the *t*-test, and Wagenmakers et al. (2016c) extended it to the correlation test. The main idea is intuitive: first the original result is summarised by its posterior distribution, and, subsequently, this posterior is used as a prior for the replication attempt. Despite its intuitive appeal in terms of the coherent updating of information, the replication Bayes factor comes with at least three challenges: (1) the procedure is not exact, as the posterior distribution from the original study often needs to be approximated by a convenient function; (2) the procedure requires technicalities and is not easy to apply; (3) the procedure does not generalise well to more complicated designs such as ANOVA (but see Harms, 2016; Wagenmakers et al., 2016c).

Here we outline an alternative procedure that solves these challenges. Specifically, the rules of Bayesian updating reveal that the replication Bayes factor

quantifies the change in evidence provided by the replication experiment, given that the evidence provided by the original study is already available. This means that any software package that is able to output ordinary Bayes factors can also be used to provide replication Bayes factors, by simply feeding it the combined data sets.

Below we first describe the Bayes factor in general terms; subsequently we outline the new conceptualisation of the replication Bayes factor and then apply it to a number of concrete examples. We end by discussing the method's limitations and future challenges.

9.2 The Bayes factor

The Bayes factor is “fundamental to the Bayesian comparison of alternative statistical models” (O’Hagan and Forster, 2004, p. 55) and it represents “the standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis and Raftery, 1997, p. 648) and “the primary tool used in Bayesian inference for hypothesis testing and model selection” (Berger, 2006, p. 378).

Developed and promoted by Jeffreys (1961), the Bayes factor contrasts the predictive performance of two competing models (e.g., Etz and Wagenmakers, 2017; Kass and Raftery, 1995; Ly et al., 2016a, 2016b). Here we focus on the standard scenario that features a null hypothesis, \mathcal{H}_0 , which stipulates the absence of an effect, and an alternative hypothesis, \mathcal{H}_1 , which stipulates the presence of an effect. Both hypotheses are falsifiable in the sense that they make specific predictions about the to-be-observed data. This is accomplished by assigning the model parameters specific values, or –in case the values are unknown and require estimation from the data– entire distributions. For instance, in the case of the t -test, \mathcal{H}_0 assigns effect size δ a single specific value, namely $\delta = 0$ (i.e., the effect is absent); in contrast, \mathcal{H}_1 assigns effect size δ a distribution that reflects the uncertainty about the true effect (e.g., $\delta \sim \mathcal{N}(0, 1)$; i.e., the effect is present but likely to be small).

When the competing hypotheses have been adorned with prior distributions so as to allow concrete predictions about to-be-observed data, the evidence provided by the actually observed data d is given by the hypotheses' relative predictive adequacy for those data (Wagenmakers et al., 2016b):

$$\underbrace{\frac{P(\mathcal{H}_1 | d)}{P(\mathcal{H}_0 | d)}}_{\text{Posterior model odds}} = \underbrace{\frac{p(d | \mathcal{H}_1)}{p(d | \mathcal{H}_0)}}_{\text{Predictive updating factor}} \times \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior model odds}} \quad (9.2.1)$$

The predictive updating factor –henceforth: the Bayes factor– quantifies the change in beliefs about the relative plausibility of the competing hypotheses brought about by the observed data. The predictions that a hypothesis makes for the observed data is obtained by averaging the predictions across the parameter space, weighted by the prior plausibility of the parameter values. For a single hypothesis, this average predictive adequacy is also known as the marginal likelihood or the prior predictive likelihood:

$$\underbrace{p(d)}_{\substack{\text{Average} \\ \text{predictive adequacy}}} = \underbrace{\int_{\Theta} d\theta}_{\substack{\text{Summed across} \\ \text{all values of } \theta}} \underbrace{f(d | \theta)}_{\substack{\text{likelihood} \\ \text{for a specific } \theta,}} \times \underbrace{\pi(\theta)}_{\substack{\text{prior plausibility of that } \theta.}} \quad (9.2.2)$$

The Bayes factor is the ratio of the average predictive adequacies for the two competing models:

$$BF_{10}(d) = \frac{p(d | \mathcal{H}_1)}{p(d | \mathcal{H}_0)} = \frac{\int_{\Theta_1} f(d | \theta_1, \mathcal{H}_1) \pi(\theta_1 | \mathcal{H}_1) d\theta_1}{\int_{\Theta_0} p(d | \theta_0, \mathcal{H}_0) \pi(\theta_0 | \mathcal{H}_0) d\theta_0}, \quad (9.2.3)$$

where θ_1 is the parameter vector under \mathcal{H}_1 , and θ_0 is the (typically shorter) parameter vector under \mathcal{H}_0 . Thus, when $BF_{10}(d) = 3$ the data are three times more likely under \mathcal{H}_1 than under \mathcal{H}_0 , and when $BF_{10}(d) = 1/7$ (or equivalently, $BF_{01}(d) = 7$), the data are seven times more likely under \mathcal{H}_0 than under \mathcal{H}_1 .

The Bayes factor offers several advantages for the analysis of empirical data (e.g., Dienes, 2014; Rouder, 2014; Schönbrodt and Wagenmakers, 2017; Wagenmakers et al., 2017b). Specifically, the Bayes factor allows the researcher to quantify evidence, to discriminate between absence of evidence (i.e., $BF_{01}(d) \approx 1$) versus evidence of absence (i.e., $BF_{01}(d) \gg 1$). The Bayes factor also allows one to monitor the evidence as the data come in (Gronau and Wagenmakers, 2017) and to design experiments in order to ensure compelling evidence. Finally, the Bayes factor can also be used to quantify replication success, a topic to which we turn next. For a more detailed introduction to the various fundamental Bayesian concepts see Wagenmakers et al. (2017a, 2017b), and Etz and Vandekerckhove (2017).

9.3 Bayesian updating in action

For concreteness, consider the article by Krupenye et al. (2016) titled “Great apes anticipate that other individuals will act according to false beliefs”. In two experiments, the authors used

“(...) an anticipatory looking test (originally developed for human infants) to show that three species of great apes reliably look in anticipation of an agent acting on a location where he falsely believes an object to be, even though the apes themselves know that the object is no longer there. Our results suggest that great apes also operate, at least on an implicit level, with an understanding of false beliefs.”
(Krupenye et al., 2016, p. 110)

The Krupenye et al. (2016) article presents two experiments. In each experiment, the apes could either look at the target or at the distractor. Here we start by presenting a Bayesian reanalysis of the first experiment. In this experiment. “(...) we tested 40 apes [19 chimpanzees, 14 bonobos, and 7 orangutans (...)]. Thirty subjects looked to either the target or the distractor during the central-approach

period. Of these 30, 20 looked first at the target ($P = 0.098$, two-tailed binomial test)" (Krupenye et al., 2016, p. 113).

Now we will reanalyse these results from a Bayesian perspective using the Summary Stats module in JASP (<https://jasp-stats.org>; Ly et al., 2017e). In our reanalysis, we assume that the data we observe are binomial and governed by a population parameter θ , the unknown proportion of apes in the population who first look at the target. The hypothesis that the apes are performing at chance level is specified as $\mathcal{H}_0 : \theta = 0.5$. This hypothesis is contrasted with \mathcal{H}_1 , the hypothesis that θ can take on values other than 0.5. For illustrative purposes, under \mathcal{H}_1 we assign θ a default prior distribution of Beta(1, 1) that is uniform across the interval from 0 to 1. With the model in place, our uncertainty about the unknown parameter θ is then updated by the data (i.e., 20 out of 30 looks at the target), and this yields the results shown in Fig. 9.1.

In Fig. 9.1, consider the two grey dots that mark the height of the prior and posterior distribution at $\theta = 0.5$, the null hypothesis of chance performance. These heights can be used to obtain the Savage-Dickey representation of the Bayes factor, an intuitive depiction of its strength and direction: If the dot at $\theta = 0.5$ gets higher from prior to posterior, the Bayes factor will provide evidence in favour of the null hypothesis (and vice-versa); moreover, the ratio of the heights of the dots exactly equals the Bayes factor (Dickey and Lientz, 1970; Wagenmakers et al., 2010). In this analysis the two dots are almost at an equal height, and the Bayes factor obtained is $BF_{10}(d) = 1.153$, which indicates that the data are non-diagnostic in choosing between the two hypotheses under scrutiny.

We may have gained hardly any evidence for the one hypothesis over the other. However, assume we know that the null hypothesis is false, uninteresting, or generally unworthy of attention. Then we are left with \mathcal{H}_1 , and the corresponding posterior information about θ is shown as the full curve in Fig. 9.1. The area under this curve to the right of $\theta = 0.5$ is much larger than the area to the left of $\theta = 0.5$; consequently, if we only take \mathcal{H}_1 into consideration, the previously non-diagnostic data inform us that θ is likely to be higher than 0.5 (see also Etz and Vandekerckhove, 2017, Example 5); indeed, the 95% credible interval ranges from 0.486 to 0.808.

The idea of Verhagen and Wagenmakers was to use this posterior from the first experiment as an informed prior for a second experiment. This is in accordance with Bayesian parameter updating and the adage "today's posterior is tomorrow's prior" (Lindley, 1972, p. 2). The resulting "replication Bayes factor" quantifies the relative predictive adequacy of the null hypothesis versus an alternative hypothesis that is completely informed by the knowledge of the parameter obtained from the first study.

To demonstrate the procedure, consider the second experiment conducted by Krupenye et al. (2016): "In experiment two, we tested 30 subjects (29 from experiment one, plus one additional bonobo). Twenty-two apes made explicit looks to the target or the distractor during this period. Of these 22, 17 looked first at the target ($P = 0.016$, two-tailed binomial test)" (Krupenye et al., 2016, p. 113).

In order to compute the replication Bayes factor, we take the posterior distribution from Experiment 1 (i.e., the solid line in Fig. 9.1), and use it as a prior distribution for the analysis of the second experiment. Recall that the original

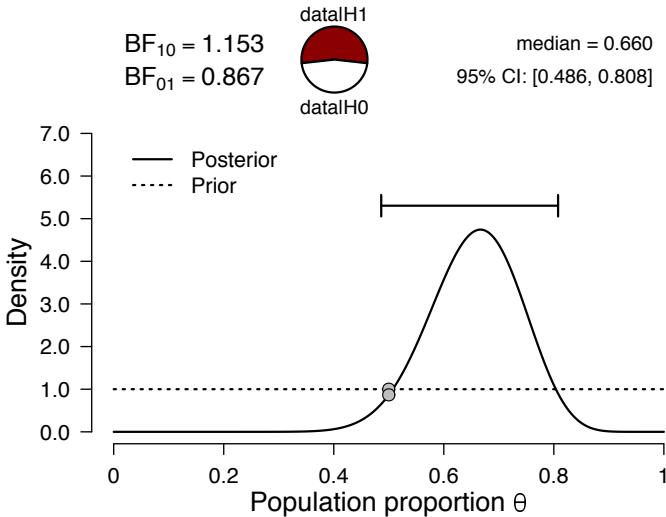


Figure 9.1: Bayesian reanalysis of the results from the first experiment in Krupenye et al. (2016), where 20 out of 30 apes ($\approx 67\%$) first looked at the target. Figure from JASP.

uniform prior was a Beta(1, 1) distribution; after incorporating the twenty successes and ten failures from the first experiment, the posterior remains a beta distribution, namely, Beta($1 + 20, 1 + 10$). This distribution can be specified in the Summary Stats module of JASP.

The result is displayed in Fig. 9.2. The dotted line quantifies the knowledge of an idealised proponent, who believes the effect is present and has access to the data from Experiment 1. The solid line is the posterior distribution when this knowledge has been updated using the data from Experiment 2. This posterior distribution does not assign much mass to values of θ near 0.5, and consequently the replication Bayes factor is relatively strong: the data are about 16 times more likely under the proponent's \mathcal{H}_r than under the skeptic's \mathcal{H}_0 .

This process of updating to a posterior and then using it as a prior for the analysis of the next experiment is relatively straightforward for this simple example. For more complex models, however, the process can be burdensome, approximate, and intricate. In the remainder of this paper we will propose an easier, more exact way forward that focuses on updating the evidence rather than updating the parameter priors.

9.4 The replication Bayes factor reconceptualised

The example above demonstrated how the replication Bayes factor can be obtained by a standard Bayesian parameter updating process, that is, by using the posterior distribution from the first experiment as a prior distribution for the replication test of the second experiment.

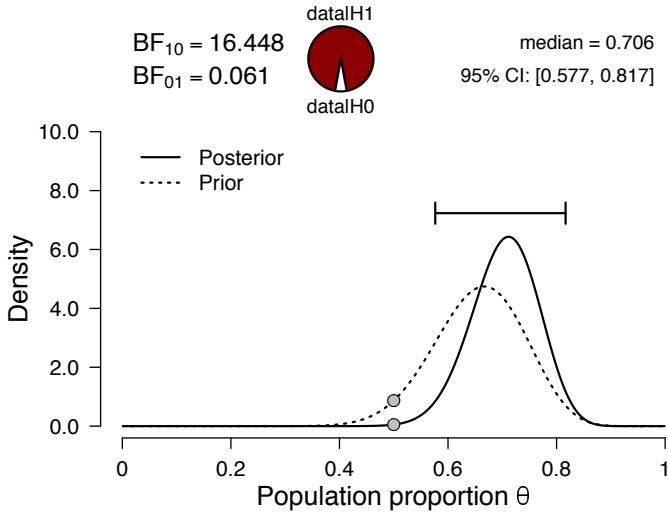


Figure 9.2: Bayesian reanalysis of the results from the second experiment in Krupenye et al. (2016) –where 17 out of 22 apes ($\approx 77\%$) first looked at the target– after having updated θ using the data from the first experiment. Figure from JASP.

However, there exists a simpler way to obtain the replication Bayes factor, one that does not explicitly require the parameter updating process. To explain this alternative method we revisit Krupenye et al. (2016) and analyse the data from both experiments together (i.e., $20 + 17 = 37$ first looks at the target out of $30 + 22 = 52$ trials). Fig. 9.3 shows the results. The posterior distribution equals the one shown in Fig. 9.2; in other words, it does not matter whether the original prior distribution is updated in two steps –first the data from Experiment 1, then the data from Experiment 2– or all at once. Crucially, this property also holds for the Bayes factor (e.g., Jeffreys, 1938, pp. 190–192). The Bayes factor for the combined result, shown in Fig. 9.3, equals 18.961. The Bayes factor for the first experiment equals 1.153 (see Fig. 9.1), and the Bayes factor for the second experiment –after updating based on the knowledge obtained in the first experiment– equals 16.448 (see Fig. 9.2).¹ Multiplying these two Bayes factors yields $1.153 \times 16.448 = 18.965$, the same result as is obtained when all data are analysed at once.²

In other words, the multiplication of component Bayes factors, when properly

¹For a warning concerning the multiplication of Bayes factors that have not been properly updated see Jeffreys, 1938, pp. 190–192; Jeffreys, 1961, section 6.0; and Wagenmakers et al., 2015b).

²The difference between 18.965 and 18.961 is due to the fact that the JASP output is accurate to three decimal places.

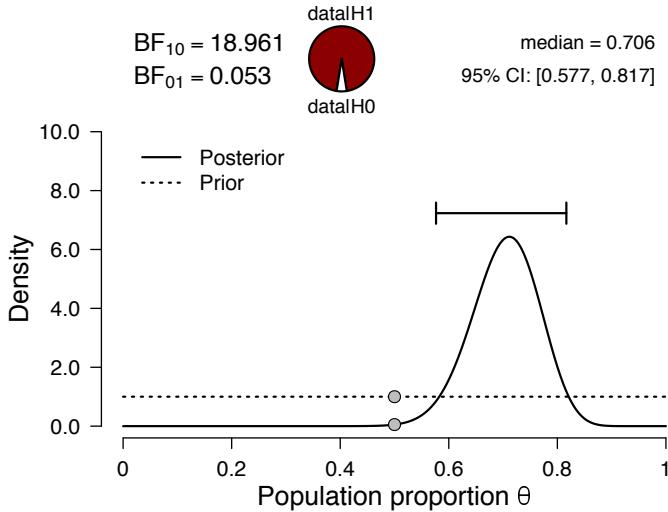


Figure 9.3: Bayesian reanalysis of the results from the first and second experiment in Krupenye et al. (2016) combined, where 37 out of 52 apes ($\approx 71\%$) first looked at the target. Figure from JASP.

updated, yields the complete Bayes factor:

$$\underbrace{\text{BF}_{10}(d_{\text{orig}}, d_{\text{rep}})}_{\text{Complete BF}} = \underbrace{\text{BF}_{10}(d_{\text{orig}})}_{\text{BF original experiment}} \times \underbrace{\text{BF}_{10}(d_{\text{rep}} | d_{\text{orig}})}_{\text{Replication BF}}, \quad (9.4.1)$$

where d_{orig} denotes the original data and d_{rep} the data from the replication attempt. Note that the replication Bayes factor is the change in the Bayes factor due to the observation of the replication data and quantifies the additional evidence for the alternative hypothesis given what was already observed in the original study.

Rearranging Eq. (9.4.1) then yields the crucial identity

$$\text{BF}_{10}(d_{\text{rep}} | d_{\text{orig}}) = \frac{\text{BF}_{10}(d_{\text{orig}}, d_{\text{rep}})}{\text{BF}_{10}(d_{\text{orig}})}, \quad (9.4.2)$$

which shows that the replication Bayes factor may be obtained by dividing the complete Bayes factor by the Bayes factor from the original experiment. Importantly, the replication Bayes factor is obtained much easier by updating the evidence than by updating the parameters, as the evidence updating procedure does not require the researcher to approximate the posterior from the original study and specify it in a software program. For complex models, this requirement is prohibitive. We now turn to additional examples that demonstrate the ease with which the evidence-updating (henceforth “EU”) replication Bayes factor can be obtained.

9.5 Example 1: A *t*-test to assess whether superstition improves performance

Consider perhaps the most routine replication scenario, one where a researcher conducts a replication of a study whose analysis featured a *t*-test. For a common *t*-test, JASP allows the specification of a Cauchy, *t*, or normal prior for the effect size δ and the user is free to specify the centre and scale of this prior (for technical details see Gronau et al., 2017a). However, in contrast to parameter θ from the binomial test, the posterior for δ in a *t*-test has no known distributional form. The applied scientist is therefore unable to use the posterior as a prior to calculate a replication Bayes factor in JASP.

To overcome this hurdle, Verhagen and Wagenmakers (2014) proposed to approximate the posterior on effect size obtained from the *t*-test with a normal distribution; this normal distribution is then used as a prior for the analysis of the replication experiment. Unfortunately, this approximation in the intermediate step between the original and the replication study makes this method computationally involved and hard to generalise to other designs.

To illustrate the simplicity of the EU replication Bayes factor, we revisit a recently published replication study by Calin-Jageman and Caldwell (2014) on the effect of superstition and performance in golf players (Damisch et al., 2010). The authors summarised the background as follows:

“Can superstitions actually improve performance? Damisch, Sto-berock, and Mussweiler (2010) reported a striking experiment in which manipulating superstitious feelings markedly increased golfing ability. Participants attempted 10 putts, each from a distance of 100 cm. Some participants were primed for superstition prior to the task by being told ‘Here is the ball. So far it has turned out to be a lucky ball.’ Controls were simply told ‘This is the ball everyone has used so far.’ Remarkably, this manipulation produced a substantial increase in golf performance: Controls made 48% of putts while superstition-primed participants made 65% of putts ($d = 0.83$, 95% CI [0.05, 1.60]).” (Calin-Jageman and Caldwell, 2014, p. 239)

A classical *t*-test³ of the original data resulted in a statistically significant result, $t(26) = 2.14, p = .042, d = .83$. As shown in Fig. 9.4, a Bayesian independent samples *t*-test using the JASP Summary Stats module returns $BF_{10}(d_{\text{orig}}) = 1.820$, a level of evidence that is not compelling. Calin-Jageman and Caldwell (2014) performed a direct replication of this work. Their Experiment 1 featured 58 control participants and 66 “superstition-activated” participants. The latter group outperformed the controls by only 2%, a result that is not statistically significant (i.e., $t(122) = .29, p = .77, d = .05$).

To compute the EU replication Bayes factor, we first need to compute the complete Bayes factor for these two datasets. Since both the original and replication papers report the raw means and standard deviations for each of the two

³This analysis is consistent with the one used in the original experiment and the replication attempt. A more appropriate statistical analysis arguably uses a hierarchical binomial model.

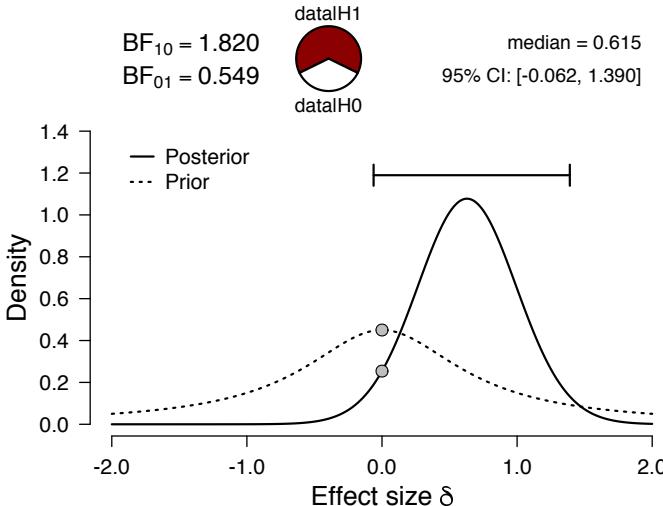


Figure 9.4: Bayesian reanalysis of the original results from Experiment 1 of Damisch et al. (2010), where golfers who played with a “lucky” ball made more putts ($t(26) = 2.14, p = .042, d = .83$). Figure from JASP.

groups (which are sufficient statistics for the t -test, see Ly et al., 2017c), we can straightforwardly compute the overall t -value for the combined data (see Appendix 9.A for a description of the algebra involved); this yields an overall $t = 1.14$, which corresponds to a complete Bayes factor of $\text{BF}_{10}(d_{\text{orig}}, d_{\text{rep}}) = 0.318$. The replication Bayes factor can now be obtained by simply dividing the complete Bayes factor by the Bayes factor from the original data alone and leads to $\text{BF}_{10}(d_{\text{rep}} | d_{\text{orig}}) = 0.175$. In other words, the skeptic’s null hypothesis predicted the data from the replication attempt $1/.175 = 5.72$ times better than the proponent’s alternative hypothesis informed by the original dataset.

9.6 Example 2: A contingency table analysis to test whether more valuable stimuli are judged to be relatively rare

The previous example featured a t -test and therefore the replication Bayes factor could also have been approximated using the parameter-updating procedure outlined in Verhagen and Wagenmakers (2014). We now turn to an example for which this parameter-updating procedure is problematic: the default Bayesian test for independence in a contingency table (Gunel and Dickey, 1974; Jamil et al., 2017).

The test for independence involves the construction of a model that is more complex than the models used for the t -test. Consequently, in JASP, the researcher can only input a parameter that governs the relative concentration of the joint prior distribution, and –for the special case of a 2×2 table– receive a posterior

9.6. Example 2: A contingency table analysis to test whether more valuable stimuli are judged to be relatively rare

Table 9.1: Data from Dai et al. (2008), who concluded that endowing a category may lead participants to judge that category to be relatively rare.

Endowed	Estimates		
	Fewer flowers	Fewer birds	Total
Flowers	15	12	27
Birds	8	21	29
Total	23	33	56

distribution for the log-odds ratio, a derived summary measure that quantifies the degree of association. This generic setup does not allow researchers to obtain a joint parameter posterior from past studies and use it as a prior for current studies, frustrating the parameter-updating version of the replication Bayes factor.

However, a contingency table replication test is straightforwardly implemented by using the EU replication Bayes factor, as we now demonstrate by an example taken from the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015). As part of the RP:P, Fuchs, Estel, and Göllner performed a replication of a study by Dai et al. (2008), who

“(...) tested a novel heuristic for making judgements of relative frequency. According to this so-called value heuristic, ‘people judge the frequency of a class of objects on the basis of the subjective value of the objects’ (p. 18). Based on the principle that scarcity increases an object’s value, the authors [Dai et al.] formulate the hypothesis that individuals will assess more valuable stimulus classes to be less frequent even when value is not diagnostic of frequency.”

The data from Dai and colleagues’ original study are presented in Table 9.1. The raw data suggest that endowing a category leads participants to judge that category as having fewer occurrences, in line with their original hypothesis. Subjecting this original finding to a classical contingency table test results in $\chi^2(1, 56) = 4.51$, $p = .037$, and a default Bayesian reanalysis (Gunel and Dickey, 1974) using JASP yields $BF_{10}(d_{\text{orig}}) = 2.880$.

The data from Fuchs and colleagues’ replication attempt are shown in Table 9.2. A classical contingency table test applied to these data returns $\chi^2(1, 51) = 1.57$, $p = .21$, which is not statistically significant. To reanalyse this data using our EU replication Bayes factor, we first combine the data into a single sample (see Table 9.3) and compute the complete Bayes factor, $BF_{10}(d_{\text{orig}}, d_{\text{rep}}) = 0.298$. To obtain the replication Bayes factor we simply divide $BF_{10}(d_{\text{orig}}, d_{\text{rep}})$ by $BF_{10}(d_{\text{orig}})$, which yields $BF_{10}(d_{\text{rep}} | d_{\text{orig}}) = 0.103$. This means that the replication data are predicted $1/0.103 = 9.71$ times better by the null hypothesis than by the alternative hypothesis informed by the original data set.

Table 9.2: Data from the replication experiment by Fuchs and colleagues. The data do not support the original finding of Dai et al. (2008).

Endowed	Estimates		
	Fewer flowers	Fewer birds	Total
Flowers	11	16	27
Birds	14	10	24
Total	25	26	51

Table 9.3: Data from the original and replication experiment combined. Note that this pooling procedure assumes that the data are exchangeable, that is, it presumes that the replication study is direct and close.

Endowed	Estimates		
	Fewer flowers	Fewer birds	Total
Flowers	26	28	54
Birds	22	31	53
Total	48	59	107

9.7 Concluding comments

The replication Bayes factor (Verhagen and Wagenmakers, 2014) provides an intuitive measure of replication success: rather than ignoring the original study, the replication Bayes factor uses the posterior distribution obtained from the original study as a prior distribution for the test of the data from the replication study.

Here we provided an additional perspective on the replication Bayes factor, namely as the change in evidence brought about by observing the results from the replication study. The advantage of this “evidence-updating” or EU perspective on the replication Bayes factor is that it does not require approximations and that it can be easily applied to complex models.

Both the original parameter-updating version and the current EU version of the replication Bayes factor are based on the idea of evidence synthesis and scientific learning (e.g., Marsman et al., 2016a; Scheibehenne et al., 2016). With more than two studies, the proposed method is similar to a fixed-effects meta-analysis that assumes the data to be exchangeable.⁴

As with any statistical method, it can become vulnerable when its core assumptions are violated. For the EU replication Bayes factor, the most serious threat to its validity arises when the replication is not close, and aspects differ that the model assumes to be the same. Consider the t -test. The parameter-updating ver-

⁴For an extension of the methodology to random-effects models and model-averaging see Gronau et al., 2017c; Scheibehenne et al., 2017.

sion updates only the test-relevant parameter δ , but the nuisance parameters (e.g., the grand mean, which is common to \mathcal{H}_0 and \mathcal{H}_1) were not updated. This small omission is rectified by the EU version that automatically and implicitly updates the joint prior for all model parameters. But this updating of nuisance parameters also creates a lack of robustness: when the nuisance parameters do undergo a large change from original to replication study, the results can be misleading. For instance, assume that a replication attempt successfully reproduces the main effect of condition, but all participants are 150 ms slower. When the raw data from the two studies are combined, this artificially inflates the variance and may make it appear as if the replication failed.

A similar warning applies for a correlation test, where the parameter of interest –the correlation coefficient ρ – may be of similar magnitude in the original and the replication study, but global changes in the location parameters of the bivariate normal can skew the outcome of the EU replication Bayes factor. For instance, suppose one studies the relation between income and body weight. The replication attempt finds the same correlation but on average participants earn \$10,000 more and weigh 15 pounds less. Visually this yields two clouds of points; each may have the same shape and orientation, but pooling the raw data may create a misleading impression.

The solution to this lack of robustness is two-fold. First, users must be aware that this is a potential problem. Second, the data may be transformed to absorb any changes in nuisance parameters. For instance, correlational data may be mean-centred before being combined.

Another vulnerability of the replication Bayes factor (regardless of whether it is the parameter-updating version or the EU version) is that, in rare case, it brings about a replication paradox. The paradox is that when a replication attempt strongly suggests that the results go in the direction opposite to the one found in the original study, the replication Bayes factor may yield compelling evidence in favour of the alternative hypothesis that the effect has successfully replicated. As with all uses of probability theory, such paradoxes reveal a lack of proper understanding. Appendix 9.C illustrates the paradox and explains that it can be resolved by imposing an order-restriction.

No single measure of replication success suffices to address all questions that surround the interpretation of a replication attempt. We advocate an inclusive approach to the statistical assessment of replication success, and we hope that the EU replication Bayes factor can be one of many tools that are at researchers' disposal, to be applied not just across laboratories but also within laboratories.

9.A Deriving the t -value across all data sets

The two-sample t -statistic over the combined data $d_{\text{all}} = (d_{\text{orig}}, d_{\text{rep}})$ can be computed from the sample means and variances of the two data sets

$$d_{\text{orig}} = (n_{\text{orig},x}, \bar{x}_{\text{orig}}, s_{\text{orig},x}^2, n_{\text{orig},y}, \bar{y}_{\text{orig}}, s_{\text{orig},y}^2), \quad (9.\text{A}.1)$$

$$d_{\text{rep}} = (n_{\text{rep},x}, \bar{x}_{\text{rep}}, s_{\text{rep},x}^2, n_{\text{rep},y}, \bar{y}_{\text{rep}}, s_{\text{rep},y}^2), \quad (9.\text{A}.2)$$

where $n_{\text{orig},x}, n_{\text{orig},y}$ are the sample sizes, $\bar{x}_{\text{orig}}, \bar{y}_{\text{orig}}$ the sample means and $\bar{s}_{\text{orig},x}^2, \bar{s}_{\text{orig},y}^2$ the (unbiased) sample variance of the first and second group from the original data set. The same symbols with orig replaced by rep have an analogous meaning. The combined two-sample t -statistic under the assumption of equal variance is then given by

$$t_{\text{all}} = \frac{\bar{x}_{\text{all}} - \bar{y}_{\text{all}}}{\sqrt{s^2 \left[\frac{1}{n_{\text{all},x}} + \frac{1}{n_{\text{all},y}} \right]}}, \quad (9.\text{A}.3)$$

where $n_{\text{all},x} = n_{\text{orig},x} + n_{\text{rep},x}$, $n_{\text{all},y} = n_{\text{orig},y} + n_{\text{rep},y}$ are the combined sample sizes of the first and second group respectively, and where

$$\bar{x}_{\text{all}} = \frac{n_{\text{orig},x}\bar{x}_{\text{orig}} + n_{\text{rep},x}\bar{x}_{\text{rep}}}{n_{\text{all},x}}, \quad (9.\text{A}.4)$$

$$\bar{y}_{\text{all}} = \frac{n_{\text{orig},y}\bar{y}_{\text{orig}} + n_{\text{rep},y}\bar{y}_{\text{rep}}}{n_{\text{all},y}}, \quad (9.\text{A}.5)$$

are the combined means of the two groups and

$$s^2 = \frac{1}{n_{\text{all},x} + n_{\text{all},y} - 2} \left[\sum_{i=1}^{n_{\text{all},x}} (x_i - \bar{x}_{\text{all}})^2 + \sum_{i=1}^{n_{\text{all},y}} (y_i - \bar{y}_{\text{all}})^2 \right], \quad (9.\text{A}.6)$$

the combined sample variance, where

$$\sum_{i=1}^{n_{\text{all},x}} (x_i - \bar{x}_{\text{all}})^2 = \nu_{\text{orig},x} s_{\text{orig},x}^2 + n_{\text{orig},x} \bar{x}_{\text{orig}}^2 + \nu_{\text{rep},x} s_{\text{rep},x}^2 + n_{\text{rep},x} \bar{x}_{\text{rep}}^2 - n_{\text{all},x} \bar{x}_{\text{all}}^2 \quad (9.\text{A}.7)$$

$$\sum_{i=1}^{n_{\text{all},y}} (y_i - \bar{y}_{\text{all}})^2 = \nu_{\text{orig},y} s_{\text{orig},y}^2 + n_{\text{orig},y} \bar{y}_{\text{orig}}^2 + \nu_{\text{rep},y} s_{\text{rep},y}^2 + n_{\text{rep},y} \bar{y}_{\text{rep}}^2 - n_{\text{all},y} \bar{y}_{\text{all}}^2 \quad (9.\text{A}.8)$$

are the combined sums of squares of the first and second group, respectively, with $\nu_{\text{orig},x} = n_{\text{orig},x} - 1$, $\nu_{\text{orig},y} = n_{\text{orig},y} - 1$ denoting the degrees of freedom.

Proof. The combined mean of the first group follows from the the equality

$$n_{\text{all}} \bar{x}_{\text{all}} = \sum_{i=1}^{n_{\text{all},x}} x_i = n_{\text{orig},x} \bar{x}_{\text{orig}} + n_{\text{rep},x} \bar{x}_{\text{rep}}, \quad (9.\text{A}.9)$$

and the combined mean of the second group can be derived analogously. Recall that the sums of squares $\sum(x_i - \bar{x})^2$ is defined as the the sum of the squares centred at zero minus n times the square of the mean, that is,

$$\nu_{\text{orig},x} s_{\text{orig},x}^2 = \sum_{i=1}^{n_{\text{orig},x}} (x_{\text{orig},i} - \bar{x}_{\text{orig}})^2 = -n_{\text{orig},x} \bar{x}_{\text{orig}}^2 + \sum_{i=1}^{n_{\text{orig},x}} x_{\text{orig},i}^2. \quad (9.\text{A}.10)$$

The same holds for the the sums of squares of the replication data and the combined data d_{all} . As such, we can write the first sums of squares in the numerator of s^2 as

$$\sum_{i=1}^{n_{\text{all},x}} (x_i - \bar{x}_{\text{all}})^2 = \nu_{\text{orig},x} s_{\text{orig},x}^2 + n_{\text{orig},x} \bar{x}_{\text{orig}}^2 + \nu_{\text{rep},x} s_{\text{rep},x}^2 + n_{\text{rep},x} \bar{x}_{\text{rep}}^2 - n_{\text{all},x} \bar{x}_{\text{all}}^2 \quad (9.A.11)$$

and the derivation is similar for y . \square

9.B Replication Bayes factors as conditional Bayes factors

Let $d_{\text{orig}}, d_{\text{rep}}$ be exchangeable and write $\pi(\theta_0 | d_{\text{orig}})$ and $\pi(\theta_1 | d_{\text{orig}})$ for the posterior for the parameters of the null model and alternative model respectively. Thus,

$$\pi(\theta_j | d_{\text{orig}}) = \frac{f(d_{\text{orig}} | \theta_j)\pi(\theta_j)}{\int f(d_{\text{orig}} | \theta_j)\pi(\theta_j)d\theta_j} = \frac{f(d_{\text{orig}} | \theta_j)\pi(\theta_j)}{p(d_{\text{orig}} | \mathcal{M}_j)} \quad (9.B.1)$$

where $p(d_{\text{orig}} | \mathcal{M}_j)$ is the marginal likelihood for \mathcal{M}_j . The procedure that uses the posterior based on the original data set d_{orig} as a prior for the replication data set can now be rewritten as

$$\text{BF}_{r0}(d_{\text{rep}}) = \frac{\int f(d_{\text{rep}} | \theta_1)\pi(\theta_1 | d_{\text{orig}})d\theta_1}{\int f(d_{\text{rep}} | \theta_0)\pi(\theta_0 | d_{\text{orig}})d\theta_0}, \quad (9.B.2)$$

$$= \frac{p(d_{\text{orig}} | \mathcal{M}_0)}{p(d_{\text{orig}} | \mathcal{M}_1)} \frac{\int f(d_{\text{rep}} | \theta_1)f(d_{\text{orig}} | \theta_1)\pi(\theta_1)d\theta_1}{\int f(d_{\text{rep}} | \theta_0)f(d_{\text{orig}} | \theta_0)\pi(\theta_0)d\theta_0}, \quad (9.B.3)$$

$$= \text{BF}_{01}(d_{\text{orig}})\text{BF}_{10}(d_{\text{orig}}, d_{\text{rep}}) = \frac{\text{BF}_{10}(d_{\text{orig}}, d_{\text{rep}})}{\text{BF}_{10}(d_{\text{orig}})}, \quad (9.B.4)$$

$$= \text{BF}_{10}(d_{\text{orig}} | d_{\text{rep}}). \quad (9.B.5)$$

Hence, the parameter-updating and evidence-updating replication Bayes factor are equivalent to each other under the assumption that d_{orig} and d_{rep} are exchangeable and the fixed effect assumption.

9.C Replication paradox and solution

Regardless of whether it is calculated from parameter-updating or evidence-updating, the replication Bayes factor can produce a paradoxical result whenever the data from a replication attempt strongly indicate that the result is in the direction opposite of the one obtained in the original experiment. Here we illustrate the paradox and explain its resolution.

For concreteness, assume that the original experiment is the first study of Kru-penye et al. (2016), where 20 out of 30 apes first looked at the target (see Fig. 9.1). Now imagine a hypothetical replication in which only 5 out of 50 apes look at the target, contradicting the direction of the original effect. One may intuit that

this disappointing result indicates compelling evidence against the proponent's alternative hypothesis as given by the posterior distribution from Fig. 9.1. Surprisingly, however, Fig. 9.5 indicates that the Bayes factor is 35.6 in favour of the proponent's alternative hypothesis.

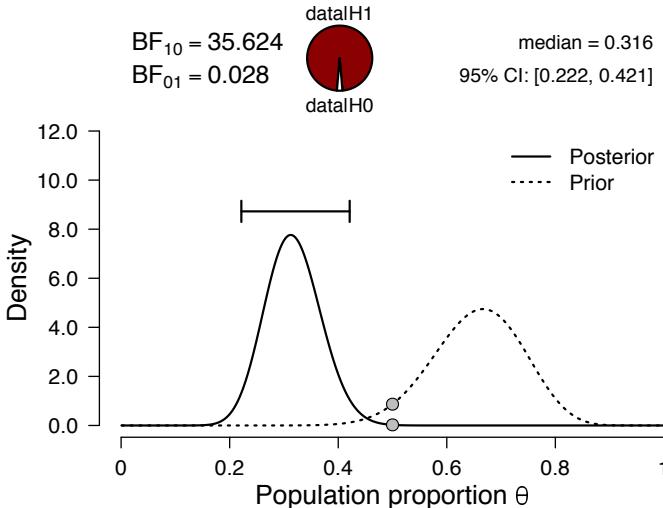


Figure 9.5: A replication paradox. In the first experiment by Krupenye et al. (2016), 20 out of 30 apes (i.e., $\approx 67\%$) had looked at the target first; in a hypothetical replication experiment, only 5 out of 50 apes did so (i.e., 10%). The effect in the hypothetical replication attempt goes in the direction opposite to that of the original study, and yet the replication Bayes factor indicates strong support in favour of the proponent's alternative hypothesis. Figure from JASP.

The key insight is to realise that the replication Bayes factor –just as other Bayes factors– quantifies *relative* evidence. With only 5 out of 50 looks at the target, the null hypothesis utterly fails to account for the data. The proponent's \mathcal{H}_r as specified by the dotted line in Fig. 9.5 also predicts these data poorly but not across all of its parameter space; indeed, \mathcal{H}_r has some prior mass on values of θ below 0.5. This resolves the paradox. The surprise at the support for the proponent's hypothesis (when the replication results contra-indicate the direction found in the original study) reflects the implicit notion that the proponent's hypothesis ought to have a direction. Specifically, in the Krupenye et al. (2016) example the authors clearly had a direction in mind when they discussed their findings. Consider the same test but now impose the restriction that $\theta \geq .5$. The result is shown in Fig. 9.6; now the Bayes factor is 72 in favour of the null hypothesis.

Generally we advocate the use of order-restrictions to create more informative tests of the underlying theory (e.g., Matzke et al., 2015b). However, it should be kept in mind that such order restrictions blind the researcher to the possibility that the effect might actually go in the direction opposite to that postulated by theory. When the data suggest that this may indeed be the case, follow-

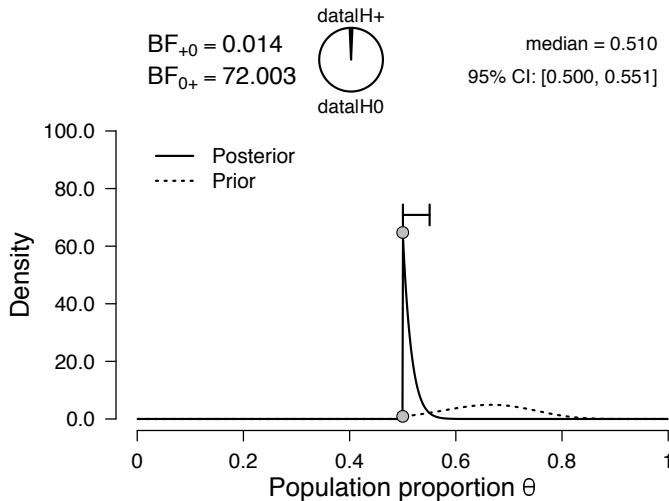


Figure 9.6: A replication paradox resolved. In the first experiment by Krupenye et al. (2016), 20 out of 30 apes (i.e., $\approx 67\%$) had looked at the target first; in a hypothetical replication experiment, only 5 out of 50 apes did so (i.e., 10%). The effect in the hypothetical replication attempt goes in the direction opposite to that of the original study. By imposing an order-restriction and allowing θ to take on only values larger than 0.5, the replication Bayes factor now indicates strong support in favour of the skeptic's null hypothesis. Figure from JASP.

up experiments may instantiate this novel prediction as a new hypothesis and examine its adequacy.

Part IV

Analytic Results

Chapter 10

Analytic Posteriors for Pearson's Correlation Coefficient

Abstract

Pearson's correlation is one of the most common measures of linear dependence. Recently, Bernardo (2015) introduced a flexible class of priors to study this measure in a Bayesian setting. For this large class of priors we show that the (marginal) posterior for Pearson's correlation coefficient and all of the posterior moments are analytic. Our results are available in the open-source software package JASP.

Keywords: Bivariate normal distribution, hypergeometric functions, reference priors.

10.1 Introduction

Pearson's product-moment correlation coefficient ρ is a measure of the linear dependency between two random variables. Its sampled version, commonly denoted by r , has been well-studied by the founders of modern statistics such as Galton, Pearson, and Fisher. Based on geometrical insights Fisher (1915, 1921) was able to derive the exact sampling distribution of r , and established that this sampling distribution converges to a normal distribution as the sample size increases. Fisher's study of the correlation has led to the discovery of variance-stabilising transformations, sufficiency (Fisher, 1920), and, arguably, the maximum likelihood estimator (Fisher, 1922; Stigler, 2007). Similar efforts were made in Bayesian statistics which focus on inferring the unknown ρ from the data that were actually observed. This type of analysis requires the statistician to (i) choose a prior on the parameters,

This chapter is published online as: Ly, A., Marsman, M., and Wagenmakers, E.-J. (2017). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*. doi: <http://dx.doi.org/10.1111/stan.12111>.

thus, also on ρ , and to (ii) calculate the posterior. Here we derive analytic posteriors for ρ given a large class of priors that include the recommendations of Jeffreys (1961), Lindley (1965), Bayarri (1981), and, more recently, Berger and Sun (2008) and Berger et al. (2015). Jeffreys's work on the correlation coefficient can also be found in the second edition of his book (Jeffreys, 1961), originally published in 1948; see Robert et al. (2009) for a modern re-read of Jeffreys's work. An earlier attempt at a Bayesian analysis of the correlation coefficient can be found in Jeffreys (1935). Before presenting the results, we first discuss some notations and recall the likelihood for the problem at hand.

10.2 Notation and result

Let $(X_1, X_2)'$ have a bivariate normal distribution with mean $\mu = (\mu_1, \mu_2)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where σ_1^2, σ_2^2 are the population variances of X_1 and X_2 , and where ρ is

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_1\sigma_2} = \frac{E(X_1X_2) - \mu_1\mu_2}{\sigma_1\sigma_2}. \quad (10.2.1)$$

Pearson's correlation coefficient ρ measures the linear association between X_1 and X_2 . In brief, the model is parameterised by the five unknowns $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

Bivariate normal data consisting of n pairs of observations can be sufficiently summarised as $d = (n, \bar{x}_1, \bar{x}_2, s_1, s_2, r)$, where

$$r = \frac{1}{n} \sum_{j=1}^n \left(\frac{x_{1j} - \bar{x}_1}{s_1} \right) \left(\frac{x_{2j} - \bar{x}_2}{s_2} \right)$$

is the sample correlation coefficient, $\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}$ the sample mean and $s_i^2 = \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$ the average sums of squares. The bivariate normal model implies that the observations d are functionally related to the parameters by the following likelihood function

$$\begin{aligned} f(d|\theta) &= \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \right)^{-n} \\ &\times \exp \left(-\frac{n}{2(1-\rho^2)} \left[\left(\frac{\bar{x}_1 - \mu_1}{\sigma_1^2} \right)^2 - 2\rho \frac{(\bar{x}_1 - \mu_1)(\bar{x}_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{\bar{x}_2 - \mu_2}{\sigma_2^2} \right)^2 \right] \right) \\ &\times \exp \left(-\frac{n}{2(1-\rho^2)} \left[\left(\frac{s_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{rs_1s_2}{\sigma_1\sigma_2} \right) + \left(\frac{s_2}{\sigma_2} \right)^2 \right] \right). \end{aligned} \quad (10.2.2)$$

For inference we use the following class of priors

$$\pi_\eta(\theta) \propto \underbrace{(1-\rho^2)^{\alpha-1} (1+\rho^2)^{\frac{\beta}{2}}}_{\pi_{\alpha,\beta}(\rho)} \underbrace{\sigma_1^{\gamma-1}}_{\pi_\gamma(\sigma_1)} \underbrace{\sigma_2^{\delta-1}}_{\pi_\delta(\sigma_2)}, \quad (10.2.3)$$

where η denotes the hyperparameters, that is, $\eta = (\alpha, \beta, \gamma, \delta)$. This class of priors is inspired by the one José Bernardo used in his talk on reference priors for the bivariate normal distribution at the “11th International Workshop on Objective Bayes Methodology in honor of Susie Bayarri”. This class of priors contains certain recommended priors as special cases.

If we set $\alpha = 1, \beta = \gamma = \delta = 0$ in Eq. (10.2.3), we retrieve the prior that Jeffreys recommended for both estimation and testing (Jeffreys, 1961, pp. 174–179 and 289–292). This recommendation is *not* the prior derived from Jeffreys’s rule based on the Fisher information (e.g., Ly et al., 2017c), as discussed in Berger and Sun (2008). With $\alpha = 1, \beta = \gamma = \delta = 0$, thus, a uniform prior on ρ , Jeffreys showed that the marginal posterior for ρ is approximately proportional to $h_a(n, r | \rho)$, where

$$h_a(n, r | \rho) = (1 - \rho^2)^{\frac{n-1}{2}} (1 - \rho r)^{\frac{3-2n}{2}},$$

represents the ρ -dependent part of the likelihood Eq. (10.2.2) with $\theta_0 = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ integrated out. For n large enough, the function h_a is a good approximation to the true reduced likelihood $h_{\gamma, \delta}$ given below.¹

If we set $\alpha = \beta = \gamma = \delta = 0$ in Eq. (10.2.3), we retrieve Lindley’s reference prior for ρ . Lindley (1965, pp. 214–221) established that the posterior of $\tanh^{-1}(\rho)$ is asymptotically normal with mean $\tanh^{-1}(r)$ and variance n^{-1} , which relates the Bayesian method of inference for ρ to that of Fisher. In Lindley’s (1965, p. 216) derivation it is explicitly stated that the likelihood with θ_0 integrated out cannot be expressed in terms of elementary functions. In his analysis, Lindley approximates the true reduced likelihood $h_{\gamma, \delta}$ with the same h_a that Jeffreys used before. Bayarri (1981) furthermore showed that with the choice $\gamma = \delta = 0$ the marginalisation paradox (Dawid et al., 1973) is avoided.

In their overview, Berger and Sun (2008) showed that for certain a, b with $\alpha = b/2 - 1, \beta = 0, \gamma = a - 2$ and $\delta = b - 1$ the priors in Eq. (10.2.3) correspond to a subclass of the generalised Wishart distribution. Furthermore, a right-Haar prior (e.g., Sun and Berger, 2007) is retrieved when we set $\alpha = \beta = 0, \gamma = -1, \delta = 1$ in Eq. (10.2.3). This right-Haar prior then has a posterior that can be constructed through simulations. That is, by simulating from a standard normal distribution and two chi-squared distributions (Berger and Sun, 2008, Table 1). This constructive posterior also corresponds to the fiducial distribution for ρ (e.g., Fraser, 1961, Hannig et al., 2006). Another interesting case is given by $\alpha = 0, \beta = 1, \gamma = \delta = 0$, which corresponds to the one-at-a-time reference prior for σ_1 and σ_2 , see also Jeffreys (1961, p. 187).

The analytic posteriors for ρ follow directly from exact knowledge of the reduced likelihood $h_{\gamma, \delta}(n, r | \rho)$, rather than its approximation used in previous work. We give full details, because we did not encounter this derivation in earlier work.

Theorem 10.2.1 (The reduced likelihood $h_{\gamma, \delta}(n, r | \rho)$). *If $|r| < 1$, $n > \gamma + 1$ and $n > \delta + 1$, then the likelihood $f(d | \theta)$ times the prior Eq. (10.2.3) with the common parameters $\theta_0 = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ integrated out is a function $f_{\gamma, \delta}$ that factors as*

$$f_{\gamma, \delta}(d | \rho) = p_{\gamma, \delta}(d_0) h_{\gamma, \delta}(n, r | \rho). \quad (10.2.4)$$

¹We thank an anonymous reviewer for clarifying how Jeffreys derived this approximation.

The first factor is the marginal likelihood with ρ fixed at zero, which does not depend on r nor on ρ , that is,

$$\begin{aligned} p_{\gamma,\delta}(d_0) &= \int \int \int \int f(d | \theta_0, \rho = 0) \pi_\gamma(\sigma_1) \pi_\delta(\sigma_2) d\mu_1 d\mu_2 d\sigma_1 d\sigma_2 \\ &= 2^{\frac{-\gamma-\delta-4}{2}} \frac{\pi^{1-n}}{n} (ns_1^2)^{\frac{1+\gamma-n}{2}} (ns_2^2)^{\frac{1+\delta-n}{2}} \Gamma\left(\frac{n-\gamma-1}{2}\right) \Gamma\left(\frac{n-\delta-1}{2}\right), \end{aligned} \quad (10.2.5)$$

where $d_0 = (n, \bar{x}_1, \bar{x}_2, s_1, s_2)$. We refer to the second factor as the reduced likelihood, a function of ρ which is given by a sum of an even and an odd function, that is, $h_{\gamma,\delta} = A_{\gamma,\delta} + B_{\gamma,\delta}$ where

$$A_{\gamma,\delta}(n, r | \rho) = (1 - \rho^2)^{\frac{n-\gamma-\delta-1}{2}} {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}; r^2 \rho^2\right), \quad (10.2.6)$$

$$B_{\gamma,\delta}(n, r | \rho) = 2r\rho(1 - \rho^2)^{\frac{n-\gamma-\delta-1}{2}} W_{\gamma,\delta}(n) {}_2F_1\left(\frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}; r^2 \rho^2\right), \quad (10.2.7)$$

where $W_{\gamma,\delta}(n) = \left[\Gamma\left(\frac{n-\gamma}{2}\right) \Gamma\left(\frac{n-\delta}{2}\right) \right] / \left[\Gamma\left(\frac{n-\gamma-1}{2}\right) \Gamma\left(\frac{n-\delta-1}{2}\right) \right]$ and where ${}_2F_1$ denotes Gauss' hypergeometric function. \diamond

Proof. To derive $f_{\gamma,\delta}(d | \rho)$ we have to perform three integrals: (i) with respect to $\pi(\mu_1, \mu_2) \propto 1$, (ii) $\pi_\gamma(\sigma_1) \propto \sigma_1^{\gamma-1}$, and (iii) $\pi_\delta(\sigma_2) \propto \sigma_2^{\delta-1}$.

(i) The integral with respect to $\pi(\mu_1, \mu_2) \propto 1$ yields

$$\begin{aligned} f(d | \sigma_1, \sigma_2, \rho) &= \frac{\left(2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2\right)^{1-n}}{n} \\ &\times \exp\left(\frac{-n}{2(1-\rho^2)}\left[\frac{s_1^2}{\sigma_1^2} - 2\rho\frac{rs_1s_2}{\sigma_1\sigma_2} + \frac{s_2^2}{\sigma_2^2}\right]\right), \end{aligned} \quad (10.2.8)$$

where we abbreviated $f(d | \sigma_1, \sigma_2, \rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(d | \theta_0, \rho) d\mu_1 d\mu_2$. The factor $p_{\gamma,\delta}(d_0)$ follows directly by setting ρ to zero in Eq. (10.2.8) and two independent gamma integrals with respect to σ_1 and σ_2 resulting in Eq. (10.2.5). These gamma integrals cannot be used when ρ is not zero. For $f_{\gamma,\delta}(d | \rho)$ which is a function of ρ , we use results from special functions theory.

(ii) For the second integral, we collect only that part of Eq. (10.2.8) that involves σ_1 into a function g , that is,

$$\int_0^\infty g(d | \sigma_1) \pi_\gamma(\sigma_1) d\sigma_1 = \int_0^\infty \sigma_1^{\gamma-n} \exp\left(-\frac{ns_1^2}{2(1-\rho^2)} \frac{1}{\sigma_1^2} + \frac{ns_1s_2}{\sigma_2(1-\rho^2)} r\rho \frac{1}{\sigma_1}\right) d\sigma_1.$$

The assumption $n > \gamma + 1$ and the substitution $u = \sigma_1^{-1}$ allow us to solve this integral using Lemma 10.A.1, which we distilled from the Bateman manuscript project (Bateman et al., 1954) with $a = \frac{ns_1^2}{2(1-\rho^2)}$, $b = -\frac{ns_1s_2}{(1-\rho^2)\sigma_2} r\rho$ and $c = n - \gamma - 1$. This yields

$$\int_0^\infty g(d | \sigma_1) \pi_\gamma(\sigma_1) d\sigma_1 = 2^{\frac{n-\gamma-3}{2}} \left(\frac{1-\rho^2}{ns_1^2}\right)^{\frac{n-\gamma-1}{2}} \left[\mathring{A}_\gamma + \mathring{B}_\gamma\right], \quad (10.2.9)$$

where

$$\dot{A}_\gamma = \Gamma\left(\frac{n-\gamma-1}{2}\right) {}_1F_1\left(\frac{n-\gamma-1}{2}; \frac{1}{2}; \frac{ns_2^2(r\rho)^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}\right), \quad (10.2.10)$$

$$\dot{B}_\gamma = \sqrt{\frac{2ns_2^2(r\rho)^2}{(1-\rho^2)}} \sigma_2^{-1} \Gamma\left(\frac{n-\gamma}{2}\right) {}_1F_1\left(\frac{n-\gamma}{2}; \frac{3}{2}; \frac{ns_2^2(r\rho)^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}\right), \quad (10.2.11)$$

and where ${}_1F_1$ denotes the confluent hypergeometric function. The functions \dot{A}_γ and \dot{B}_γ are the even and odd solution of Weber's differential equation in the variable $z = (r\rho)^2 \frac{ns_2^2}{2(1-\rho^2)\sigma_2^2}$, respectively.

(iii) With $f_\gamma(d|\sigma_2, \rho) = \int_0^\infty f(d|\sigma_1, \sigma_2, \rho) \pi_\gamma(\sigma_1) d\sigma_1$, we see that $f_{\gamma,\delta}(d|\rho)$ follows from integrating σ_2 out of the following expression

$$f_\gamma(d|\sigma_2, \rho) \pi_\delta(\sigma_2) = 2^{\frac{-n-\gamma-1}{2}} \frac{\pi^{1-n}}{n} (ns_1^2)^{\frac{1+\gamma-n}{2}} (1-\rho^2)^{\frac{-\gamma}{2}} \times \left[\check{A}_\gamma(d|\sigma_2, \rho) + \check{B}_\gamma(d|\sigma_2, \rho) \right], \quad (10.2.12)$$

where

$$\begin{aligned} \check{A}_\gamma &= \Gamma\left(\frac{n-\gamma-1}{2}\right) k(d|\rho, \sigma_2) \\ \check{B}_\gamma &= \left(\frac{2ns_2^2}{1-\rho^2}\right)^{\frac{1}{2}} r\rho \Gamma\left(\frac{n-\gamma}{2}\right) l(d|\rho, \sigma_2), \end{aligned} \quad (10.2.13)$$

and where

$$k(d|\rho, \sigma_2) = \sigma_2^{\delta-n} e^{-\frac{ns_2^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}} {}_1F_1\left(\frac{n-\gamma-1}{2}; \frac{1}{2}; (r\rho)^2 \frac{ns_2^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}\right) \quad (10.2.14)$$

$$l(d|\rho, \sigma_2) = \sigma_2^{\delta-n-1} e^{-\frac{ns_2^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}} {}_1F_1\left(\frac{n-\gamma}{2}; \frac{3}{2}; (r\rho)^2 \frac{ns_2^2}{2(1-\rho^2)} \frac{1}{\sigma_2^2}\right) \quad (10.2.15)$$

Hence, the last integral with respect to σ_2 only involves the functions k and l . The assumption $n > \delta + 1$ and the substitution $t = \frac{ns_2^2}{2(1-\rho^2)} \sigma_2^{-2}$, thus, $\int d\sigma_2 = \int -\frac{1}{2} \sqrt{\frac{ns_2^2}{2(1-\rho^2)}} t^{-\frac{3}{2}} dt$ allows us to solve this integral using Eq. (7.621.4) from Gradshteyn and Ryzhik (2007, p. 822) with $s = 1$, $\tilde{k} = (r\rho)^2$. This yields

$$\begin{aligned} \int_0^\infty k(d|\rho, \sigma_2) d\sigma_2 &= 2^{\frac{n-\delta-3}{2}} \left(\frac{1-\rho^2}{ns_2^2}\right)^{\frac{n-\delta-1}{2}} \Gamma\left(\frac{n-\delta-1}{2}\right) \\ &\quad \times {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}; r^2 \rho^2\right), \end{aligned} \quad (10.2.16)$$

$$\begin{aligned} \int_0^\infty l(d|\rho, \sigma_2) d\sigma_2 &= 2^{\frac{n-\delta-2}{2}} \left(\frac{1-\rho^2}{ns_2^2}\right)^{\frac{n-\delta}{2}} \Gamma\left(\frac{n-\delta}{2}\right) \\ &\quad \times {}_2F_1\left(\frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}; r^2 \rho^2\right). \end{aligned} \quad (10.2.17)$$

After we combine the results we see that $f_{\gamma,\delta}(d | \rho) = \tilde{A}_{\gamma,\delta}(d | \rho) + \tilde{B}_{\gamma,\delta}(d | \rho)$, where

$$\frac{\tilde{A}_{\gamma,\delta}(d | \rho)}{p_{\gamma,\delta}(d_0)} = (1 - \rho^2)^{\frac{n-\gamma-\delta-1}{2}} {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}; r^2\rho^2\right),$$

$$\frac{\tilde{B}_{\gamma,\delta}(d | \rho)}{p_{\gamma,\delta}(d_0)} = 2r\rho(1 - \rho^2)^{\frac{n-\gamma-\delta-1}{2}} W_{\gamma,\delta}(n) {}_2F_1\left(\frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}; r^2\rho^2\right).$$

Hence, $f_{\gamma,\delta}(d | \rho)$ is of the asserted form. Note that $A_{\gamma,\delta} = \frac{\tilde{A}_{\gamma,\delta}(d | \rho)}{p_{\gamma,\delta}(d_0)}$ is even, while $\frac{\tilde{B}_{\gamma,\delta}(d | \rho)}{p_{\gamma,\delta}(d_0)}$ is an odd function of ρ . \square

This main theorem confirms Lindley's insights; $h_{\gamma,\delta}(n, r | \rho)$ is indeed not expressible in terms of elementary functions and the prior on ρ is updated by the data only through its sampled version r and the sample size n . As a result, the marginal likelihood for data d then factors into $p_\eta(d) = p_{\gamma,\delta}(d_0)p_{\alpha,\beta}(n, r; \gamma, \delta)$, where $p_{\alpha,\beta}(n, r; \gamma, \delta) = \int h_{\gamma,\delta}(n, r | \rho)\pi_{\alpha,\beta}(\rho)d\rho$ is the normalising constant of the marginal posterior of ρ . More importantly, the fact that the reduced likelihood is the sum of an even and an odd function allows us to fully characterise the posterior distribution of ρ for the priors Eq. (10.2.3) in terms of its moments. These moments are easily computed, as the prior $\pi_{\alpha,\beta}(\rho)$ itself is symmetric around zero. Furthermore, the prior $\pi_{\alpha,\beta}(\rho)$ can be normalised as

$$\pi_{\alpha,\beta}(\rho) = \frac{(1 - \rho^2)^{\alpha-1}(1 + \rho^2)^{\frac{\beta}{2}}}{\mathcal{B}(\frac{1}{2}, \alpha) {}_2F_1(-\frac{\beta}{2}, \frac{1}{2}; \frac{1}{2} + \alpha; -1)}, \quad (10.2.18)$$

where $\mathcal{B}(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$ denotes the beta function. The case with $\beta = 0$ is also known as the (symmetric) stretched beta distribution on $(-1, 1)$ and leads to Lindley's reference prior when we ignore the normalisation constant, i.e., $\mathcal{B}(\frac{1}{2}, \alpha)$, and, subsequently, let $\alpha \rightarrow 0$.

Corollary 10.2.1 (Characterisation of the marginal posteriors of ρ). *If $n > \gamma + \delta - 2\alpha + 1$, then the main theorem implies that whenever the marginal likelihood with all the parameters integrated out factors as $p_\eta(d) = p_{\gamma,\delta}(d_0)p_{\alpha,\beta}(n, r; \gamma, \delta)$, where*

$$p_{\alpha,\beta}(n, r; \gamma, \delta) = \int_{-1}^1 h_{\gamma,\delta}(n, r | \rho)\pi_{\alpha,\beta}(\rho)d\rho = \int_{-1}^1 A_{\gamma,\delta}(n, r | \rho)\pi_{\alpha,\beta}(\rho)d\rho, \quad (10.2.19)$$

defines the normalising constant of the marginal posterior for ρ . Observe that the integral involving $B_{\gamma,\delta}$ is zero, because $B_{\gamma,\delta}$ is odd on $(-1, 1)$. More generally, the k th posterior moment of ρ is

$$E_{\alpha,\beta}(\rho^k | n, r; \gamma, \delta) = \begin{cases} \frac{1}{p_{\alpha,\beta}(n, r; \gamma, \delta)} \int_{-1}^1 \rho^k A_{\gamma,\delta}(n, r | \rho)\pi_{\alpha,\beta}(\rho)d\rho & \text{if } k \text{ is even,} \\ \frac{1}{p_{\alpha,\beta}(n, r; \gamma, \delta)} \int_{-1}^1 \rho^k B_{\gamma,\delta}(n, r | \rho)\pi_{\alpha,\beta}(\rho)d\rho & \text{if } k \text{ is odd.} \end{cases} \quad (10.2.20)$$

These posterior moments define the series

$$E_{\alpha,\beta}(\rho^k | n, r; \gamma, \delta) = \begin{cases} \frac{1}{C_{\alpha,\beta}} \sum_{m=0}^{\infty} \frac{(\frac{n-\gamma-1}{2})_m (\frac{n-\delta-1}{2})_m}{(\frac{1}{2})_m m!} a_{k,m} r^{2m} & \text{if } k \text{ is even,} \\ \frac{2W_{\gamma,\delta}(n)}{C_{\alpha,\beta}} \sum_{m=0}^{\infty} \frac{(\frac{n-\gamma}{2})_m (\frac{n-\delta}{2})_m}{(\frac{3}{2})_m m!} b_{k,m} r^{2m+1} & \text{if } k \text{ is odd,} \end{cases} \quad (10.2.21)$$

where $C_{\alpha,\beta} = \mathcal{B}(\frac{1}{2}, \alpha) {}_2F_1(-\frac{\beta}{2}, \frac{1}{2}; \alpha + \frac{1}{2}; -1)$ is the normalisation constant of the prior Eq. (10.2.18), $W_{\gamma,\delta}(n)$ is the ratios of gamma functions as defined under Eq. (10.2.7) and $(x)_m = \frac{\Gamma(x+m)}{\Gamma(x)} = x(x+1)(x+2)\dots(x+m-1)$ refers to the Pochhammer symbol for rising factorials. The terms $a_{k,m}$ and $b_{k,m}$ are

$$a_{k,m} = \mathcal{B}\left(\frac{1}{2} + \frac{k+2m}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) \quad (10.2.22)$$

$$\times {}_2F_1\left(\frac{-\beta}{2}, \frac{k+2m+1}{2}; \frac{k+2m+2\alpha+n-\gamma-\delta}{2}; -1\right),$$

$$b_{k,m} = \mathcal{B}\left(\frac{1}{2} + \frac{k+2m+1}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) \quad (10.2.23)$$

$$\times {}_2F_1\left(\frac{-\beta}{2}, \frac{k+2m+2}{2}; \frac{k+2m+2\alpha+n-\gamma-\delta+1}{2}; -1\right).$$

The series defined in Eq. (10.2.21) are hypergeometric when β is a non-negative integer. \diamond

Proof. The series $E_{\alpha,\beta}(\rho^k | n, r; \gamma, \delta)$ result from term-wise integration of the hypergeometric functions in $A_{\gamma,\delta}$ and $B_{\gamma,\delta}$. The assumption $n > \gamma + \delta - 2\alpha + 1$ and the substitution $x = \rho^2$ allows us to solve these integrals using Eq. (3.197.8) in Gradshteyn and Ryzhik (2007, p. 317) with their $\tilde{\alpha} = 1$, $u = 1$, $\lambda = \frac{\beta}{2}$, $\mu = \alpha + \frac{n-\gamma-\delta-1}{2}$ and $\nu = \frac{1}{2} + \frac{k+2m}{2}$ when k is even, while we use $\nu = \frac{1}{2} + \frac{k+2m+1}{2}$ when k is odd. A direct application of the ratio test shows that the series converge when $|r| < 1$. \square

10.3 Analytic posteriors for the case $\beta = 0$

For most of the priors discussed above we have $\beta = 0$, which leads to the following simplification of the posterior.

Corollary 10.3.1 (Characterisation of the marginal posteriors of ρ , when $\beta = 0$). *If $n > \gamma + \delta - 2\alpha + 1$ and $|r| < 1$, then the marginal posterior for ρ is*

$$\begin{aligned} \pi_{\alpha}(\rho | n, r; \gamma, \delta) &= \frac{(1 - \rho^2)^{\frac{2\alpha+n-\gamma-\delta-3}{2}}}{p_{\alpha}(n, r; \gamma, \delta) \mathcal{B}(\frac{1}{2}, \alpha)} \\ &\times \left[{}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}; r^2 \rho^2\right) \right. \\ &\quad \left. + 2r\rho W_{\gamma,\delta}(n) {}_2F_1\left(\frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}; r^2 \rho^2\right) \right], \end{aligned} \quad (10.3.1)$$

where $p_\alpha(n, r; \gamma, \delta)$ refers to the normalising constant of the (marginal) posterior of ρ , which is given by

$$p_\alpha(n, r; \gamma, \delta) = \frac{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha\right)} {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \alpha + \frac{n-\gamma-\delta}{2}; r^2\right). \quad (10.3.2)$$

More generally, when $\beta = 0$, the k th posterior moment is

$$\frac{\mathcal{B}\left(\frac{1}{2} + \frac{k}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) {}_3F_2\left(\frac{k+1}{2}, \frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}, \frac{k+2\alpha+n-\gamma-\delta}{2}; r^2\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{2\alpha+n-\gamma-\delta}{2}; r^2\right)},$$

when k is even, and

$$2rW_{\gamma, \delta}(n) \frac{\mathcal{B}\left(\frac{1}{2} + \frac{k+1}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) {}_3F_2\left(\frac{k+2}{2}, \frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}, \frac{k+2\alpha+n-\gamma-\delta+1}{2}; r^2\right)}{\mathcal{B}\left(\frac{1}{2}, \alpha + \frac{n-\gamma-\delta-1}{2}\right) {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{2\alpha+n-\gamma-\delta}{2}; r^2\right)},$$

when k is odd. \diamond

Proof. The assumption $n > \gamma + \delta - 2\alpha + 1$ and the substitution $x = \rho^2$ allows us to use Eq. (7.513.12) in Gradshteyn and Ryzhik (2007, p. 814) with $\mu = \alpha + \frac{n-\gamma-\delta-1}{2}$ and $\nu = \frac{1}{2} + \frac{k}{2}$ when k is even, while we use $\nu = \frac{1}{2} + \frac{k+1}{2}$ when k is odd. The normalising constant of the posterior $p_\alpha(n, r; \gamma, \delta)$ is a special case with $k = 0$. \square

The marginal posterior for ρ updated from the generalised Wishart prior, the right-Haar prior and Jeffreys's recommendation then follow from a direct substitution of the values for α, γ and δ as discussed under Eq. (10.2.3). Lindley's reference posterior for ρ is given by

$$\frac{{}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{1}{2}; r^2\rho^2\right) + 2r\rho W_{0,0}(n) {}_2F_1\left(\frac{n}{2}, \frac{n}{2}; \frac{3}{2}; r^2\rho^2\right)}{\mathcal{B}\left(\frac{1}{2}, \frac{n-1}{2}\right) {}_2F_1\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{n}{2}; r^2\right)} (1 - \rho^2)^{\frac{n-3}{2}},$$

which follows from Eq. (10.3.1) by setting $\gamma = \delta = 0$ and, subsequently, letting $\alpha \rightarrow 0$.

Lastly, for those who wish to sample from the posterior distribution, we suggest the use of an independence-chain Metropolis algorithm (IMH; Tierney, 1994) with Lindley's normal approximation of the posterior of $\tanh^{-1}(\rho)$ as the proposal. This method could be used when Pearson's correlation is embedded within a hierarchical model, as the posterior for ρ will then be a full conditional distribution within a Gibbs sampler. For $\alpha = 1$, $\beta = \gamma = \delta = 0$, $n = 10$ observations and $r = 0.6$, the acceptance rate of the IMH algorithm was already well above 75%, suggesting a fast convergence of the Markov chain. For n larger, the acceptance rate further increases. The R code for the independence-chain Metropolis algorithm can be found on the first author's home page. In addition, this analysis is also implemented in the open-source software package JASP.

10.A A lemma distilled from the Bateman Project

Lemma 10.A.1. *For $a, c > 0$ the following equality holds*

$$\int_0^\infty u^{c-1} \exp\left(-au^2 - bu\right) du = 2^{-1} a^{-\frac{c}{2}} \left[\mathring{A}(a, b, c) + \mathring{B}(a, b, c) \right], \quad (10.A.1)$$

that is, the integral is solved by the functions

$$\begin{aligned} \mathring{A}(a, b, c) &= \Gamma\left(\frac{c}{2}\right) {}_1F_1\left(\frac{c}{2}; \frac{1}{2}; \frac{b^2}{4a}\right), \\ \mathring{B}(a, b, c) &= -\frac{b}{\sqrt{a}} \Gamma\left(\frac{c+1}{2}\right) {}_1F_1\left(\frac{c+1}{2}; \frac{3}{2}; \frac{b^2}{4a}\right), \end{aligned} \quad (10.A.2)$$

which define the even and odd solutions to Weber's differential equation in the variable $z = \frac{b}{\sqrt{2a}}$ respectively. \diamond

Proof. By Bateman et al. (1954, p. 313, Eq. 13) we note that,

$$\int_0^\infty u^{c-1} \exp\left(-au^2 - bu\right) du = (2a)^{\frac{-c}{2}} \Gamma(c) \exp\left(\frac{b^2}{8a}\right) D_{-c}\left(\frac{b}{\sqrt{2a}}\right), \quad (10.A.3)$$

where $D_\lambda(z)$ is Whittaker's (1902) parabolic cylinder function (Abramowitz and Stegun, 1964). By virtue of Eq. (4) on p. 117 of Bateman et al. (1953), we can decompose $D_\lambda(z)$ into a sum of an even and odd function. Replacing this decomposition for $D_\lambda(z)$ in Eq. (10.A.3) and an application of the duplication formula of the gamma function yields the statement. \square

Chapter 11

Analytic Posteriors for the Binomial Rate Parameters, and the Odds Ratio

Abstract

We present analytic posteriors for a binomial rate parameter and the odds ratio. Both expressions involve hypergeometric functions and can be used to derive Bayes factors for these scenarios.

Keywords: Bayesian inference, hypergeometric functions.

11.1 Introduction

This chapter contains derivations of analytic posteriors for the rate of a binomial distribution and the odds ratio.

11.2 Binomial distribution

11.2.1 A localised prior for the binomial rate parameter

Definition 11.2.1 (Localised beta prior). We say that θ has a beta distribution localised at θ_0 if its density is given by

$$\pi_\eta(\theta) = \underbrace{\frac{1}{\text{Beta}(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \left(\frac{1-\theta_0}{\theta_0}\right)^\alpha}_{\text{Beta}(\theta; \alpha, \beta)} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-(\alpha+\beta)}, \quad (11.2.1)$$

where η is shorthand for the parameter vector $\eta = (\alpha, \beta, \theta_0)$ and where $\text{Beta}(\theta; \alpha, \beta)$ refers to the (standard) two-parameter beta distribution. \diamond

With $\theta_0 = 1/2$ we retrieve the (standard) beta $\text{Beta}(\theta; \alpha, \beta)$. We choose to write the last term as $[1 - \theta + \theta(\frac{1-\theta_0}{\theta_0})] = (1 - [2 - \frac{1}{\theta_0}] \theta)$ due to its relation with the hypergeometric function.

11. ANALYTIC POSTERIORS FOR THE BINOMIAL RATE PARAMETERS, AND THE ODDS RATIO

Theorem 11.2.1 (Marginal likelihood of a binomially distributed random variable with the beta prior localised at θ_0). *The localised beta prior has the following marginal likelihood*

$$p_\eta(d) = \binom{n}{y} \frac{\mathcal{B}(\alpha + y, \beta + n - y)}{\mathcal{B}(\alpha, \beta)} \times \left(\frac{1-\theta_0}{\theta_0}\right)^\alpha {}_2F_1(\alpha + \beta, \alpha + y; \alpha + \beta + n; 2 - \frac{1}{\theta_0}), \quad (11.2.2)$$

where d refers to the data y and n and where

$${}_2F_1(u, v; w; z) = \sum_{k=0}^{\infty} \frac{(u)_k (v)_k}{(w)_k k!} z^k, \quad (11.2.3)$$

is Gauss' hypergeometric function (Oberhettinger, 1972, Section 15), where $(u)_k = \frac{\Gamma(u+k)}{\Gamma(u)}$ denotes Pochhammer's raising factorial. \diamond

Proof. Writing $p_\eta(\emptyset) = \mathcal{B}(\alpha, \beta) \left(\frac{1-\theta_0}{\theta_0}\right)^{-\alpha}$ for the normalisation constant of the prior combined with $u_1 = y + \alpha$, $u_2 = n - y + \beta$, $v = \alpha + \beta$, and by definition of the marginal likelihood, we have

$$p_\eta(\emptyset)p_\eta(d) = \binom{n}{y} \int_0^1 \theta^{u_1-1} (1-\theta)^{u_2-1} (1-\theta[2 - \frac{1}{\theta_0}])^{-v} d\theta, \quad (11.2.4)$$

$$= \binom{n}{y} \mathcal{B}(u_1, u_2) {}_2F_1(v, u_1; u_1 + u_2; 2 - \frac{1}{\theta_0}). \quad (11.2.5)$$

The last equality follows from Euler's integral representation of the hypergeometric function (Abramowitz and Stegun, 1964, p. 558). \square

Corollary 11.2.1 (Localised beta posterior and its characterisation). *The posterior is*

$$\pi_\eta(\theta | d) = \pi_{\alpha, \beta}(\theta | d) \frac{(1 - [2 - \frac{1}{\theta_0}] \theta)^{-(\alpha+\beta)}}{2F_1(\alpha + \beta, \alpha + y; \alpha + \beta + n; 2 - \frac{1}{\theta_0})}, \quad (11.2.6)$$

where

$$\pi_{\alpha, \beta}(\theta | d) = \frac{\theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1}}{\mathcal{B}(\alpha + y, \beta + n - y)}, \quad (11.2.7)$$

is the posterior based on the standard beta prior. The last term of the localised posterior $\pi_\eta(\theta | d)$ can be thought of as a "skewness" term due to the localisation. The k th posterior moment is

$$\begin{aligned} E_\eta(\theta^k | d) &= \frac{\mathcal{B}(\alpha + y + k, \beta + n - y)}{\mathcal{B}(\alpha + y, \beta + n - y)} \\ &\times \frac{{}_2F_1(\alpha + \beta, \alpha + y + k; \alpha + \beta + n + k; 2 - \frac{1}{\theta_0})}{2F_1(\alpha + \beta, \alpha + y; \alpha + \beta + n; 2 - \frac{1}{\theta_0})}, \\ &= \frac{(\alpha + y)_k}{(\alpha + \beta + n)_k} \frac{{}_2F_1(\alpha + \beta, \alpha + y + k; \alpha + \beta + n + k; 2 - \frac{1}{\theta_0})}{2F_1(\alpha + \beta, \alpha + y; \alpha + \beta + n; 2 - \frac{1}{\theta_0})}, \end{aligned} \quad (11.2.8)$$

where $E_\eta(\cdot | d)$ refers to the expectation with respect to the posterior $\pi_\eta(\theta | d)$. \diamond

Proof. The statements follow directly from the proof given above with $u_1 = y + \alpha + k$. \square

Remark 11.2.1. *The normalisation constant $p_\eta(\emptyset)$ can be retrieved from $p_\eta(d)$ by taking $y = n = 0$, that is, the normalisation constant can also be expressed as a hypergeometric function. Consequently, the localised prior can be viewed as a partially conjugate prior for the binomial distribution by which we mean that the prior and posterior are of the same form, but that only some of its parameters need updating. More specifically, to update the prior to a posterior, only the exponents of θ and $(1 - \theta)$ need to be changed, and, subsequently, one of the upper and the lower terms of the hypergeometric function ${}_2F_1$.* \diamond

Proof. The definition of the normalisation constant of the prior and Eq. (11.2.5) implies that

$$p_\eta(\emptyset) = \int \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta[2-\frac{1}{\theta_0}])^{-(\alpha+\beta)} d\theta, \quad (11.2.9)$$

$$= \mathcal{B}(\alpha, \beta) {}_2F_1(\alpha, \alpha + \beta; \alpha + \beta; 2 - \frac{1}{\theta_0}), \quad (11.2.10)$$

which should suffice for the proof. Note that one of the upper and the lower term of the hypergeometric function are the same, which implies that it can be simplified. Indeed, by Eq. (15.4.6) of Olver et al. (2010, p. 386), or just the definition of Gauss' hypergeometric function, we have

$${}_2F_1(\alpha, \alpha + \beta; \alpha + \beta; 2 - \frac{1}{\theta_0}) = (1 - [2 - \frac{1}{\theta_0}])^{-\alpha} = (\frac{1-\theta_0}{\theta_0})^{-\alpha}, \quad (11.2.11)$$

which completes the proof. \square

Corollary 11.2.2 (Min-sided prior, marginal likelihood, posterior and its characterisation). *By construction, the prior associated to the min-sided hypothesis $\mathcal{H}_- : \theta \in (0, \theta_0)$ is*

$$\pi_\eta^{(-)}(\theta) = \frac{(\frac{1-\theta_0}{\theta_0})^\alpha}{\mathcal{B}(\frac{1}{2}; \alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta[2-\frac{1}{\theta_0}])^{-(\alpha+\beta)} \mathbf{1}_{(0, \theta_0]}(\theta), \quad (11.2.12)$$

where $\mathcal{B}(\frac{1}{2}; \alpha, \beta)$ is the incomplete beta integral evaluated at a half. The min-sided marginal likelihood is

$$p_\eta^{(-)}(d) = \frac{\theta_0^y (1-\theta_0)^\alpha}{\mathcal{B}(\frac{1}{2}; \alpha, \beta)(y+\alpha)} \binom{n}{y} \quad (11.2.13)$$

$$\times AF_1(y+\alpha; 1-\beta-n+y, \alpha+\beta; y+\alpha+1; \theta_0, 2\theta_0-1),$$

where

$$AF_1(u; v_1, v_2; w; x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(u)_{m+n} (v_1)_m (v_2)_n}{m! n! (w)_{m+n}} x^m y^n, \quad (11.2.14)$$

11. ANALYTIC POSTERIORS FOR THE BINOMIAL RATE PARAMETERS, AND THE ODDS RATIO

is known as Appell's hypergeometric function of the first kind. As such, the min-sided posterior is

$$\pi_\eta^{(-)}(\theta | d) = \frac{y + \alpha}{\theta_0^{\alpha+y}} \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-(\alpha+\beta)} \mathbf{1}_{(0,\theta_0]}(\theta) \\ / AF_1(y + \alpha; 1 - \beta - n + y, \alpha + \beta; y + \alpha + 1; \theta_0, 2\theta_0 - 1). \quad (11.2.15)$$

Lastly, the k th posterior moment is

$$E_\eta^{(-)}(\theta^k | d) = \frac{y + \alpha}{y + \alpha + k} \theta_0^k \\ \times \frac{AF_1(y + \alpha + k; 1 - \beta - n + y, \alpha + \beta; y + \alpha + 1 + k; \theta_0, 2\theta_0 - 1)}{AF_1(y + \alpha; 1 - \beta - n + y, \alpha + \beta; y + \alpha + 1; \theta_0, 2\theta_0 - 1)} \quad (11.2.16)$$

where $E_\eta^{(-)}(\cdot | d)$ is the expectation with respect to the min-sided posterior. \diamond

Proof. To simplify matters we write $p_\eta^{(-)}(\emptyset) = \mathcal{B}(\frac{1}{2}; \alpha, \beta)(\frac{1-\theta_0}{\theta_0})^{-\alpha}$ for the normalisation constant of the prior. The integral of interest is then of the form

$$p_\eta^{(-)}(\emptyset)p_\eta^{(-)}(d) = \binom{n}{y} \int_0^{\theta_0} \theta^{u-1} (1 - \theta)^{-v_1} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-v_2} d\theta, \quad (11.2.17)$$

where $u = y + \alpha + k$, $v_1 = 1 - n + y - \beta$, $v_2 = \alpha + \beta$. Using the change of variable $t = \theta/\theta_0$, thus, $\int d\theta = \int \theta_0 dt$ we can rewrite the integral as

$$p_\eta^{(-)}(\emptyset)p_\eta^{(-)}(d) = \binom{n}{y} \theta_0^u \int_0^1 t^{u-1} (1 - \theta_0 t)^{-v_1} (1 - [2\theta_0 - 1]t)^{-v_2} dt, \quad (11.2.18)$$

$$= \binom{n}{y} \frac{\theta_0^u}{u} AF_1(u; v_1, v_2; u + 1; \theta_0, 2\theta_0 - 1), \quad (11.2.19)$$

where the latter equality is an (Euler) integral representation of Appell's hypergeometric function due to $u = y + \alpha + k$ being positive, see Eq. (3.211) of Gradshteyn and Ryzhik (2007, p. 318) and Bailey (1964, p. 77). Entering the terms u, v_1, v_2 with $k = 0$ yields the marginal likelihood, and, subsequently, the posterior and the posterior moments. \square

Remark 11.2.2. The normalisation constant $p_\eta(\emptyset)$ can be written as an Appell function AF_1 , which implies that the min-sided localised prior can be thought of as a partially conjugate prior for the binomial distribution by which we mean that the prior and posterior are of the same form, but that only some of its parameters are updated. More specifically, to update the prior, only the exponents of the θ and $(1 - \theta)$ terms need updating. Similarly, only the Pochhammer coefficients in the Appell series need to be updated for the normalisation constant of the posterior. \diamond

Proof. By definition of the normalisation constant of the min-sided prior, the transformation $t = \frac{\theta}{\theta_0}$ and Eq. (11.2.19) we have

$$p_\eta^{(-)}(\emptyset) = \int_0^{\theta_0} \theta^{\alpha-1} (1 - \theta)^{\beta-1} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-(\alpha+\beta)} d\theta \quad (11.2.20)$$

$$= \frac{\theta_0^\alpha}{\alpha} AF_1(\alpha; 1 - \beta, \alpha + \beta; \alpha + 1; \theta_0, 2\theta_0 - 1). \quad (11.2.21)$$

This should suffice for the statement. As a sanity check we note that the lower term of this Appell function is the sum of two of its upper terms, that is, $\alpha + 1 = \alpha + \beta + 1 - \beta$, which allows us to use Eq. (16.16.1) of Olver et al. (2010, p. 414) resulting in

$$p_\eta^{(-)}(\emptyset) = \frac{\theta_0^\alpha}{\alpha} [1 - (2\theta_0 - 1)]^{-\alpha} {}_2F_1(\alpha, 1 - \beta; \alpha + 1; \frac{\theta_0 - 2\theta_0 + 1}{1 - 2\theta_0 + 1}), \quad (11.2.22)$$

$$= (\frac{1 - \theta_0}{\theta_0})^\alpha \underbrace{{}_2F_1(\alpha, 1 - \beta; \alpha + 1; \frac{1}{2})}_{\mathcal{B}(\frac{1}{2}; \alpha, \beta)}, \quad (11.2.23)$$

due to the relation between the incomplete beta function and Gauss' hypergeometric function, that is, Eq. (8.17.7) of Olver et al. (2010, p. 183). \square

Corollary 11.2.3 (Plus-sided prior, marginal likelihood and posterior). *By construction the prior associated to the plus-sided hypothesis $\mathcal{H}_+ : \theta \in (\theta_0, 1)$ is*

$$\pi_\eta^{(+)}(\theta) = \frac{1}{p_\eta^{(+)}(\emptyset)} \theta^{\alpha-1} (1-\theta)^{\beta-1} (1 - \theta[2 - \frac{1}{\theta_0}])^{-(\alpha+\beta)} \mathbf{1}_{(\theta_0, 1]}(\theta), \quad (11.2.24)$$

where $p_\eta^{(+)}(\emptyset) = [\mathcal{B}(\alpha, \beta) - \mathcal{B}(\frac{1}{2}; \alpha, \beta)](\frac{1 - \theta_0}{\theta_0})^{-\alpha}$ denotes the normalisation constant of the prior. The plus-sided marginal likelihood is

$$\begin{aligned} p_\eta^{(+)}(d) &= \binom{n}{y} \frac{\theta_0^{y-1} (1 - \theta_0)^{n-y}}{2^{\alpha+\beta} [\mathcal{B}(\alpha, \beta) - \mathcal{B}(\frac{1}{2}; \alpha, \beta)]} (n - y + \beta)^{-1} \\ &\times AF_1(1; 1 - \alpha - y, \alpha + \beta; \beta + n - y + 1; \frac{\theta_0 - 1}{\theta_0}, \frac{2\theta_0 - 1}{2\theta_0}). \end{aligned} \quad (11.2.25)$$

As such, the plus-sided posterior is

$$\begin{aligned} \pi_\eta^{(+)}(\theta | d) &= \frac{2^{\alpha+\beta} (n - y + \beta)}{\theta_0^{y+\alpha-1} (1 - \theta_0)^{n-y-\alpha}} \\ &\times \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1} (1 - \theta[2 - \frac{1}{\theta_0}])^{-(\alpha+\beta)} \mathbf{1}_{(\theta_0, 1]}(\theta) \\ &\Big/ AF_1(1; 1 - y - \alpha, n - y + \beta - 1; \frac{\theta_0 - 1}{\theta_0}, \frac{2\theta_0 - 1}{2\theta_0} 2\theta_0). \end{aligned} \quad (11.2.26)$$

Lastly, the k th posterior moment is

$$E_\eta^{(+)}(\theta^k | d) = \theta_0^k \frac{AF_1(1; 1 - y - k - \alpha, \alpha + \beta; n - y + \beta + 1; \frac{\theta_0 - 1}{\theta_0}, \frac{2\theta_0 - 1}{2\theta_0})}{AF_1(1; 1 - y - \alpha, \alpha + \beta; n - y + \beta + 1; \frac{\theta_0 - 1}{\theta_0}, \frac{2\theta_0 - 1}{2\theta_0})}, \quad (11.2.27)$$

where $E_\eta^{(+)}(\cdot | d)$ is the expectation with respect to the plus-sided posterior. \diamond

Proof. To simplify matters we write $p_\eta^{(+)}(\emptyset) = [\mathcal{B}(\alpha, \beta) - \mathcal{B}(\frac{1}{2}; \alpha, \beta)](\frac{1 - \theta_0}{\theta_0})^{-\alpha}$ for the normalisation constant of the prior. With $v_1 = 1 - y - k - \alpha$, $w_1 = n - y + \beta$, $v_2 = \alpha + \beta$ the integral of interest is then

$$p_\eta^{(+)}(\emptyset) p_\eta^{(+)}(d) = \binom{n}{y} \int_{\theta_0}^1 \theta^{-v_1} (1 - \theta)^{w_1 - 1} (1 - [2 - \frac{1}{\theta_0}] \theta)^{-v_2}. \quad (11.2.28)$$

11. ANALYTIC POSTERIORS FOR THE BINOMIAL RATE PARAMETERS, AND THE ODDS RATIO

Using the change of variable $x = (\theta - \theta_0)/(1 - \theta_0)$, thus, $\int d\theta = \int (1 - \theta_0)dx$ this integral is then

$$p_\eta^{(+)}(\emptyset)p_\eta^{(+)}(d) = \binom{n}{y} 2^{-v_2}\theta_0^{-v_1}(1 - \theta_0)^{w_1 - v_2} \quad (11.2.29)$$

$$\begin{aligned} & \times \int_0^1 (1 - x)^{w_1 - 1} (1 - [\frac{\theta_0 - 1}{\theta_0}]x)^{-v_1} (1 - [\frac{2\theta_0 - 1}{2\theta_0}]x)^{-v_2} dx, \\ & = \binom{n}{y} 2^{-v_2}\theta_0^{-v_1}(1 - \theta_0)^{w_1 - v_2} w_1^{-1} \end{aligned} \quad (11.2.30)$$

$$\times AF_1(1; v_1, v_2; w_1 + 1; \frac{\theta_0 - 1}{\theta_0}, \frac{2\theta_0 - 1}{2\theta_0}),$$

where the latter equality is an (Euler) integral representation of Appell's hypergeometric function due to $w_1 = n - y + \beta > 0$. The normalisation constant $p_\eta^{(+)}(\emptyset)$ follows from setting $n = y = k = 0$, the marginal likelihood follows from setting $k = 0$. \square

Corollary 11.2.4 (The two-sided Bayes factor and its relationship to the one-sided Bayes factors). *Let $f(d | \theta_0) = \binom{n}{y} \theta_0^y (1 - \theta_0)^{n-y}$ and define the two-sided Bayes factor as $BF_{10;\eta}(d) = \frac{p_\eta(d)}{f(d|\theta_0)}$, then this two-sided Bayes factor is a convex combination of the one-sided Bayes factors $BF_{-0;\eta}(d) = \frac{p_\eta^{(-)}(d)}{f(d|\theta_0)}$ and $BF_{+0;\eta}(d) = \frac{p_\eta^{(+)}(d)}{f(d|\theta_0)}$, that is,*

$$BF_{10;\eta}(d) = \frac{p_\eta^{(-)}(\emptyset)}{p_\eta(\emptyset)} BF_{-0;\eta}(d) + \frac{p_\eta^{(+)}(\emptyset)}{p_\eta(\emptyset)} BF_{+0;\eta}(d) \quad (11.2.31)$$

where $p_\eta(\emptyset), p_\eta^{(-)}(\emptyset), p_\eta^{(+)}(\emptyset)$ are the normalisation constants of the two-sided, min-sided and plus-sided priors. For the localised prior this implies that

$$BF_{10;\eta}(d) = \frac{\mathcal{B}(\frac{1}{2}; \alpha, \beta)}{\mathcal{B}(\alpha, \beta)} BF_{-0;\eta}(d) + \left[1 - \frac{\mathcal{B}(\frac{1}{2}; \alpha, \beta)}{\mathcal{B}(\alpha, \beta)}\right] BF_{+0;\eta}(d). \quad (11.2.32)$$

For symmetric priors, thus, $\alpha = \beta = a$, the relationship simplifies to

$$BF_{10;a,\theta_0}(d) = \frac{1}{2} BF_{-0;a,\theta_0}(d) + \frac{1}{2} BF_{+0;a,\theta_0}(d), \quad (11.2.33)$$

and this holds for any θ_0 . \diamond

Proof. Writing $\pi_u(\theta)$ for the unnormalised (two-sided) prior, that is, $\pi_u(\theta) = p_\eta(\emptyset)\pi_\eta(\theta)$ we now have

$$p_\eta(d) = \frac{1}{p_\eta(\emptyset)} \int_0^1 f(d | \theta) \pi_u(\theta) d\theta \quad (11.2.34)$$

$$= \frac{1}{p_\eta(\emptyset)} \left[\int_0^{\theta_0} f(d | \theta) \pi_u(\theta) d\theta + \int_{\theta_0}^1 f(d | \theta) \pi_u(\theta) d\theta \right] \quad (11.2.35)$$

$$= \frac{p_\eta^{(-)}(\emptyset)}{p_\eta(\emptyset)} p_\eta^{(-)}(d) + \frac{p_\eta^{(+)}(\emptyset)}{p_\eta(\emptyset)} p_\eta^{(+)}(d) \quad (11.2.36)$$

dividing both sides by $f(d | \theta_0)$ now yields the result. \square

Corollary 11.2.5 (Sums of Appell functions of the first kind). *Note that we have shown that certain Gauss' hypergeometric functions can be written as a sum of two Appell functions of the first kind, that is,*

$$g(u_1, u_2, v, \theta_0) = \mathcal{B}(u_1, u_2) {}_2F_1(v, u_1; u_1 + u_2; 2 - \frac{1}{\theta_0}) \quad (11.2.37)$$

where

$$\begin{aligned} g(u_1, u_2, v, \theta_0) &= \frac{\theta_0^{u_1}}{u_1} {}_1F_1(u_1; 1 - u_2, v; u_1 + 1; \theta_0, 2\theta_0 - 1) \\ &\quad + 2^{-v} \theta_0^{u_1-1} (1 - \theta_0)^{u_2-v} u_2^{-1} \\ &\quad \times {}_1F_1(1; 1 - u_1, v; u_2 + 1; \frac{\theta_0-1}{\theta_0}, \frac{2\theta_0-1}{2\theta_0}) \end{aligned} \quad (11.2.38)$$

is the sum of Appell functions of the first kind. \diamond

Proof. The assertion follows from the three calculations given above with $u = u_1, v_1 = 1 - u_2, v_2 = v$ in Eq. (11.2.21) and $v_1 = 1 - u_1, w_1 = u_2, v_2 = v$ in Eq. (11.2.30). \square

We have shown how to obtain analytic results by using a localised beta prior on θ . This prior is related to the so-called generalised beta prime distribution.

Definition 11.2.2 (Generalised beta prime distribution). We say that a random variable ζ has a generalised beta prime distribution and write

$$\zeta \sim \text{genBetaPrime}(\alpha, \beta, u, v), \quad (11.2.39)$$

if the density of ζ at the outcome z is given by

$$f_\zeta(z) = \frac{v^{-\alpha u}}{\mathcal{B}(\alpha, \beta)} \frac{|u| z^{\alpha u - 1}}{(1 + (\frac{z}{v})^u)^{\alpha + \beta}}, \quad (11.2.40)$$

where $0 < z$ and the parameters α, β, v are positive. \diamond

Example 11.2.1 (Localised beta prior and the generalised beta prime distribution). Let θ be distributed as a beta distribution $\mathcal{B}(\alpha, \beta)$ localised at θ_0 , then the odds form of θ , that is, $\zeta = \frac{\theta}{1-\theta}$ is distributed as a generalised beta prime distribution $\zeta \sim \text{genBetaPrime}(\alpha, \beta, 1, \frac{\theta_0}{1-\theta_0})$. \diamond

Proof. The result follows from first rewriting the localised beta as

$$\int \pi_\eta(\theta) d\theta = \frac{1}{\mathcal{B}(\alpha, \beta)} \int (\frac{\theta}{1-\theta})^\alpha (\frac{1-\theta_0}{\theta_0})^\alpha [1 + (\frac{\theta}{1-\theta})(\frac{1-\theta_0}{\theta_0})]^{-\alpha-\beta} \theta^{-1} (1-\theta)^{-1} d\theta,$$

and the replacement $\theta = \frac{z}{1+z}$, $z_0 = \frac{\theta_0}{1-\theta_0}$, thus, $\int d\theta = \int (1+z)^{-2} dz$. \square

11.3 Products of generalised beta prime distributions and the odds ratio

Theorem 11.3.1 (Products of generalised beta prime distributions). *Let $\zeta_1 \sim \text{genBetaPrime}(\alpha_1, \beta_1, u, v_1)$ and $\zeta_2 \sim \text{genBetaPrime}(\alpha_2, \beta_2, u, v_2)$ be independent generalised beta prime distributions with common shape u . The density of the product $\Omega = \zeta_1 \zeta_2$ is then equivalently given by*

$$f_\Omega(\omega) = \begin{cases} C\left(\frac{\omega}{v_1 v_2}\right)^{\alpha_2 u - 1} {}_2F_1(\alpha_2 + \beta_2, \alpha_2 + \beta_1; \alpha_+ + \beta_-; 1 - (\frac{\omega}{v_1 v_2})^u), \\ C\left(\frac{v_1 v_2}{\omega}\right)^{\beta_1 u + 1} {}_2F_1(\alpha_1 + \beta_1, \alpha_2 + \beta_1; \alpha_+ + \beta_-; 1 - (\frac{v_1 v_2}{\omega})^u), \\ C\left(\frac{\omega}{v_1 v_2}\right)^{\alpha_1 u - 1} {}_2F_1(\alpha_1 + \beta_1, \alpha_1 + \beta_2; \alpha_+ + \beta_-; 1 - (\frac{\omega}{v_1 v_2})^u), \\ C\left(\frac{v_1 v_2}{\omega}\right)^{\beta_2 u + 1} {}_2F_1(\alpha_2 + \beta_2, \alpha_1 + \beta_2; \alpha_+ + \beta_-; 1 - (\frac{v_1 v_2}{\omega})^u), \end{cases} \quad (11.3.1)$$

where $\alpha_+ = \alpha_1 + \alpha_2$, $\beta_- = \beta_1 + \beta_2$, and $C = \frac{|u|}{v_1 v_2} \frac{\mathcal{B}(\alpha_2 + \beta_1, \alpha_1 + \beta_2)}{\mathcal{B}(\alpha_1, \beta_1) \mathcal{B}(\alpha_2, \beta_2)}$. The four equivalent results can also be derived using Kummer's 24 solutions, thus, Klein's 4-group. \diamond

Proof. By the convolution theorem for products of independent random variables we have $f_\Omega(\omega) = \int_{Z_2} \frac{1}{z_2} f_{\zeta_1}(\frac{\omega}{z_2}) f_{\zeta_2}(z_2) dz_2$, where we have written Z_i for the domain of ζ_i and z_i for a specific outcome of ζ_i . For $f_{\zeta_i}(\cdot)$ the generalised beta prime density this yields

$$f_\Omega(\omega) = \tilde{C} \int z_2^{(\alpha_2 - \alpha_1)u - 1} [1 + (\frac{\omega}{z_2 v_1})^u]^{-(\alpha_1 + \beta_1)} [1 + (\frac{z_2}{v_2})^u]^{-(\alpha_2 + \beta_2)} dz_2, \quad (11.3.2)$$

where $\tilde{C} = \frac{v_1^{1-2\alpha_1 u} v_2^{-\alpha_2 u}}{\mathcal{B}(\alpha_1, \beta_1) \mathcal{B}(\alpha_2, \beta_2)} |u|^2 \omega^{\alpha_1 u - 1}$. Writing the linear terms as $(z_2^u + b)^{-c}$ for some $b, c > 0$ we then get

$$f_\Omega(\omega) = \check{C} \int z_2^{(\alpha_2 + \beta_1)u - 1} [z_2^u + (\frac{w}{v_1})^u]^{-(\alpha_1 + \beta_1)} [v_2^u + z_2^u]^{-(\alpha_2 + \beta_2)} dz_2, \quad (11.3.3)$$

where $\check{C} = \frac{v_1^{-\alpha_1 u} v_2^{\beta_2 u}}{\mathcal{B}(\alpha_1, \beta_1) \mathcal{B}(\alpha_2, \beta_2)} |u|^2 \omega^{\alpha_1 u - 1}$. The change of variable $x = z_2^u$, thus, $\int dz_2 = \int u^{-1} x^{\frac{1}{u}-1} dx$ then leads to

$$f_\Omega(\omega) = \hat{C} \int x^{\alpha_2 + \beta_1 - 1} [x + (\frac{\omega}{v_1})^u]^{-(\alpha_1 + \beta_1)} [x + v_2^u]^{-(\alpha_2 + \beta_2)} dx, \quad (11.3.4)$$

where $\hat{C} = \frac{v_1^{-\alpha_1 u} v_2^{\beta_2 u}}{\mathcal{B}(\alpha_1, \beta_1) \mathcal{B}(\alpha_2, \beta_2)} |u| \omega^{\alpha_1 u - 1}$. To solve this integral we use Eq. (3.197.1) of Gradshteyn and Ryzhik (2007, p. 317) with in their notation $\nu = \alpha_2 + \beta_1$, $\beta = (\frac{\omega}{v_1})^u$, $\mu = \alpha_1 + \beta_1$, $\gamma = v_2^u$, $\varrho = \alpha_2 + \beta_2$ resulting in

$$\begin{aligned} f_\Omega(\omega) &= \hat{C} (\frac{\omega}{v_1})^{(\alpha_2 - \alpha_1)u} v_2^{-(\alpha_2 + \beta_2)u} \\ &\quad \times \mathcal{B}(\alpha_2 + \beta_1, \alpha_1 + \beta_2) {}_2F_1(\alpha_2 + \beta_2, \alpha_2 + \beta_1; \alpha_+ + \beta_-; 1 - (\frac{\omega}{v_1 v_2})^u), \end{aligned} \quad (11.3.5)$$

and the first equation of the assertion follows after rearranging the terms in \hat{C} .

		A		
		A	not A	Row total
B		Y_{11}	Y_{12}	$Y_{1.}$
	not B	Y_{21}	Y_{22}	$Y_{2.}$
Column total		$Y_{.1}$	$Y_{.2}$	$Y_{..}$

 Table 11.1: The focus is on the (in)dependence between A and B .

On the other hand, using Eq. (3.197.1) of Gradshteyn and Ryzhik (2007, p. 317) with in their notation $\nu = \alpha_2 + \beta_1, \beta = (\frac{\omega}{v_1})^u, \mu = \alpha_1 + \beta_1, \gamma = v_2^u, \varrho = \alpha_2 + \beta_2$, instead, yields

$$\begin{aligned} f_\Omega(\omega) &= \hat{C}(\frac{\omega}{v_1})^{-(\alpha_1 + \beta_1)u} v_2^{(\beta_1 - \beta_2)u} \\ &\times \mathcal{B}(\alpha_2 + \beta_1, \alpha_1 + \beta_2) {}_2F_1(\alpha_1 + \beta_1, \alpha_2 + \beta_1; \alpha_+ + \beta_-; 1 - (\frac{v_1 v_2}{\omega})^u), \end{aligned} \quad (11.3.6)$$

and the second equation of the assertion follows after rearranging the terms in \hat{C} . The third and the fourth equation can be derived analogously by considering the product convolution in terms of $f_\Omega(\omega) = \int_{Z_1} \frac{1}{z_1} f_{\zeta_2}(\frac{\omega}{z_1}) f_{\zeta_1}(z_1) dz_1$ instead. \square

Example 11.3.1 (Odds ratio). *Let the data Y be arranged in a 2×2 contingency table, see Table 11.1. When the A assignment is independent on the B assignment, we have $P(A, B) = P(A)P(B)$. If this independence relationship is perfectly mimicked in the data we have $P(Y_{11}) = P(Y_{.1})P(Y_{1.})$. The (sample) odds ratio in a 2-by-2 contingency table is a measure of deviation of independence and defined as $O = \frac{Y_{11}Y_{22}}{Y_{12}Y_{21}}$.*

In a Bayesian setting we are also interested in the implied deviation of independence on the population level. For this type of inference we have to (1) assume a model that specifies how the observed data are related to the unobserved parameters, and (2) a prior on the parameter, which allows for probabilistic statements about the parameter given the data. A general model for the data is $Y_{ij} \sim \text{Pois}(\lambda_{ij})$ with $Y_{11}, Y_{12}, Y_{21}, Y_{22}$ all independent of each other. A computationally convenient choice would be to take $\lambda_{ij} \sim \text{Gam}(\alpha_{ij}, \beta_{ij})$.

Analogous to the sample we define the (population) odds ratio as $\Omega = \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}$. Note that the two ratios are distributed as generalised beta prime distributions

$$\frac{\lambda_{11}}{\lambda_{12}} \sim \text{genBetaPrime}(\alpha_{11}, \alpha_{12}, 1, \frac{\beta_{12}}{\beta_{11}}), \quad (11.3.7)$$

$$\frac{\lambda_{22}}{\lambda_{21}} \sim \text{genBetaPrime}(\alpha_{22}, \alpha_{21}, 1, \frac{\beta_{21}}{\beta_{22}}), \quad (11.3.8)$$

respectively. As such, Ω is distributed according to Eq. (11.3.1). Thus,

$$f_\Omega(\omega) = \begin{cases} C(\frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}}\omega)^{\alpha_{22}-1} {}_2F_1(\alpha_{2.}, \alpha_{.2}; \alpha_{..}; 1 - \frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}}\omega), \\ C(\frac{\beta_{12}\beta_{21}}{\beta_{11}\beta_{22}}\frac{1}{\omega})^{\alpha_{12}+1} {}_2F_1(\alpha_{1.}, \alpha_{.2}; \alpha_{..}; 1 - \frac{\beta_{12}\beta_{21}}{\beta_{11}\beta_{22}}\frac{1}{\omega}), \\ C(\frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}}\omega)^{\alpha_{11}-1} {}_2F_1(\alpha_{1.}, \alpha_{.1}; \alpha_{..}; 1 - \frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}}\omega), \\ C(\frac{\beta_{12}\beta_{21}}{\beta_{11}\beta_{22}}\frac{1}{\omega})^{\alpha_{21}+1} {}_2F_1(\alpha_{2.}, \alpha_{.1}; \alpha_{..}; 1 - \frac{\beta_{12}\beta_{21}}{\beta_{11}\beta_{22}}\frac{1}{\omega}), \end{cases} \quad (11.3.9)$$

11. ANALYTIC POSTERIORS FOR THE BINOMIAL RATE PARAMETERS, AND THE ODDS RATIO

where $\alpha_{i\cdot}$ denotes the i th row sum, $\alpha_{\cdot j}$ the j th column sum and $\alpha_{\cdot\cdot}$ the total sum of the α_{ij} parameters and where $C = \frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}} \frac{\mathcal{B}(\alpha_{\cdot 2}, \alpha_{\cdot 1})}{\mathcal{B}(\alpha_{11}, \alpha_{12})\mathcal{B}(\alpha_{22}, \alpha_{21})}$. \diamond

Proof. Theorem 11.3.2 below implies that $\lambda_{i1}/\lambda_{i2}$ is indeed a generalised beta prime distribution and the result follows from Theorem 11.3.1. \square

Theorem 11.3.2 (Ratios of gammas). *Let $X \sim \text{Gam}(\alpha_x, \beta_x)$ and $Y \sim \text{Gam}(\alpha_y, \beta_y)$ be independent, then $Z = X/Y \sim \text{genBetaPrime}(\alpha_x, \alpha_y, 1, \frac{\beta_y}{\beta_x})$.* \diamond

Proof. By the convolution theorem for independent ratios we have $f_Z(z) = \int y f_X(zy) f_Y(y) dy$, thus, with $C = \frac{\beta_x^{\alpha_x}}{\Gamma(\alpha_x)} \frac{\beta_y^{\alpha_y}}{\Gamma(\alpha_y)}$ and $\alpha_{\cdot\cdot} = \alpha_x + \alpha_y$ we have

$$f_Z(z) = C \int y(zy)^{\alpha_x-1} e^{-\beta_x zy} y^{\alpha_y-1} e^{-\beta_y y} dy, \quad (11.3.10)$$

$$= Cz^{\alpha_x-1} \int y^{\alpha_{\cdot\cdot}-1} e^{-(\beta_x z + \beta_y)y} dy, \quad (11.3.11)$$

$$= Cz^{\alpha_x-1} \Gamma(\alpha_{\cdot\cdot}) (\beta_x z + \beta_y)^{-(\alpha_{\cdot\cdot})}, \quad (11.3.12)$$

$$= \frac{(\frac{\beta_y}{\beta_x})^{-\alpha_x}}{\mathcal{B}(\alpha_x, \alpha_y)} z^{\alpha_x-1} \left[1 + \frac{z}{\beta_y/\beta_x} \right]^{-(\alpha_x+\alpha_y)}, \quad (11.3.13)$$

which is exactly what we wanted to show. \square

Corollary 11.3.1 (Analytic posterior for the odds ratio). *The analytic posterior for the odds ratio is Eq. (11.3.9) with α_{ij} replaced by $\alpha_{ij} + y_{ij}$.* \diamond

11.4 Concluding remarks

We hope that these analytic results provide further insights to posteriors and Bayes factors for the test of two proportions, and the 2-by-2 odds ratio.

Part V

Two Tutorials

Chapter 12

A Tutorial on Bridge Sampling

Abstract

The marginal likelihood plays an important role in many areas of Bayesian statistics such as parameter estimation, model comparison, and model averaging. In most applications, however, the marginal likelihood is not analytically tractable and must be approximated using numerical methods. Here we provide a tutorial on bridge sampling (Bennett, 1976; Meng and Wong, 1996), a reliable and relatively straightforward sampling method that allows researchers to estimate the marginal likelihood for models of varying complexity. First, we introduce bridge sampling and three related sampling methods using the beta-binomial model as a running example. We then apply bridge sampling to estimate the marginal likelihood for the Expectancy Valence (EV) model, a popular model for reinforcement learning. Our results indicate that bridge sampling provides accurate estimates for both a single participant and a hierarchical version of the EV model. We conclude that bridge sampling is an attractive method for mathematical psychologists who typically aim to approximate the marginal likelihood for a limited set of possibly high-dimensional models.

Keywords: Bayes factor, hierarchical model, marginal likelihood, normalising constant, predictive accuracy, reinforcement learning.

12.1 Introduction

Bayesian statistics has become increasingly popular in mathematical psychology (Andrews and Baguley, 2013; Bayarri et al., 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen et al., 2015; Wetzels et al., 2016). The Bayesian approach is conceptually

This chapter is published as: Gronau, Q. F., Sarafoglou, A., Matzke, D. M., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.J., & Steingroever, H. (2017) A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–91. doi: <https://doi.org/10.1016/j.jmp.2017.09.005> Also available as *arXiv preprint*, arXiv:1705.01064.

simple, theoretically coherent, and easily applied to relatively complex problems. These problems include hierarchical modelling (Matzke et al., 2015a; Matzke and Wagenmakers, 2009; Rouder and Lu, 2005; Rouder et al., 2005, 2007) or the comparison of non-nested models (Lee, 2008; Pitt et al., 2002; Shiffrin et al., 2008). Three major applications of Bayesian statistics concern parameter estimation, model comparison, and Bayesian model averaging. In all three areas, the marginal likelihood –that is, the probability of the observed data given the model of interest– plays a central role (see also Gelman and Meng, 1998).

First, in parameter estimation, we consider a single model and aim to quantify the uncertainty for a parameter of interest θ after having observed the data d . This is realised by means of a posterior distribution that can be obtained using Bayes theorem, as

$$\pi(\theta | d) = \frac{f(d | \theta) \pi(\theta)}{\int f(d | \theta) \pi(\theta) d\theta} = \frac{\underbrace{f(d | \theta)}_{\text{marginal likelihood}} \underbrace{\pi(\theta)}_{\text{prior}}}{\underbrace{p(d)}_{\text{likelihood}}} \quad (12.1.1)$$

Here, the marginal likelihood of the data $p(d)$ ensures that the posterior distribution $\pi(\theta | d)$ is a proper probability density function (pdf) in the sense that it integrates to one. This illustrates why in parameter estimation the marginal likelihood is referred to as a normalising constant.

Second, in model comparison, we consider $m \in \mathbb{N}$ number of competing models, and are interested in the relative plausibility of a particular model \mathcal{M}_i , where $i = 1, 2, \dots, m$, given the prior model probability and the evidence from the data d (see three special issues on this topic in the *Journal of Mathematical Psychology*: Mulder and Wagenmakers, 2016; Myung et al., 2000b; Wagenmakers and Waldorp, 2006a). This relative plausibility is quantified by the so-called posterior model probability $P(\mathcal{M}_i | d)$ of model \mathcal{M}_i given the data d (Berger and Molina, 2005)

$$P(\mathcal{M}_i | d) = \frac{p(d | \mathcal{M}_i) P(\mathcal{M}_i)}{\sum_{j=1}^m p(d | \mathcal{M}_j) P(\mathcal{M}_j)}, \quad (12.1.2)$$

where $p(d | \mathcal{M}_j)$ denotes the marginal likelihood of model \mathcal{M}_j , and where the denominator is the sum of the marginal likelihood times the prior model probability of all m models. In model comparison, the marginal likelihood for a specific model is also referred to as the model evidence (Didelot et al., 2011), the integrated likelihood (Kass and Raftery, 1995), and the predictive likelihood of the model (Gemanian and Lopes, 2006, Chapter 7). As a function of the data it is also known as the predictive probability of the data (Kass and Raftery, 1995), or the prior predictive density (Ntzoufras, 2009). Note that computationally the marginal likelihood of Eq. (12.1.2) is the same as the marginal likelihood of Eq. (12.1.1). However, for the latter equation we dropped the model index because in parameter estimation we consider only one model.

If only two models \mathcal{M}_1 and \mathcal{M}_2 are considered, Eq. (12.1.2) can be used to quantify the relative posterior model plausibility of model \mathcal{M}_1 compared to model \mathcal{M}_2 . This relative plausibility is given by the ratio of the posterior probabilities of both models, and is referred to as the posterior model odds

$$\underbrace{\frac{P(\mathcal{M}_1 | d)}{P(\mathcal{M}_2 | d)}}_{\text{Posterior model odds}} = \underbrace{\frac{p(d | \mathcal{M}_1)}{p(d | \mathcal{M}_2)}}_{\text{BF}_{12}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}}_{\text{Prior model odds}}. \quad (12.1.3)$$

Eq. (12.1.3) illustrates that the posterior model odds are the product of two factors: The right most factor is the ratio of the prior probabilities of the models also known as the prior model odds. The factor in the middle is the ratio of the marginal likelihoods —the so-called Bayes factor (Etz and Wagenmakers, 2017; Jeffreys, 1961; Ly et al., 2016a, 2016b; Robert, 2016). The Bayes factor plays an important role in model comparison and is referred to as the “standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis and Raftery, 1997, p. 648) and “the primary tool used in Bayesian inference for hypothesis testing and model selection” (Berger, 2006, p. 378).

Third, the marginal likelihood plays an important role in Bayesian model averaging (BMA; Hoeting et al., 1999) where aspects of parameter estimation and model comparison are combined. As in model comparison, BMA considers several models; however, it does not aim to identify a single best model. Instead, it fully acknowledges model uncertainty. Model averaged parameter inference can be obtained by combining, across all models, the posterior distribution of the parameter of interest weighted by each model’s posterior model probability, and as such depends on the marginal likelihood of the models. This procedure assumes that the parameter of interest has identical interpretation across the different models. Model averaged predictions can be obtained in a similar manner.

A problem that arises in all three areas –parameter estimation, model comparison, and BMA– is that an analytical expression of the marginal likelihood can be obtained only for certain restricted examples. This is a pressing problem in Bayesian modelling, and in particular in mathematical psychology where models can be non-linear and equipped with a large number of parameters, especially when the models are implemented in a hierarchical framework. Such a framework incorporates both commonalities and differences between participants of one group by assuming that the model parameters of each participant are drawn from a group-level distribution (for advantages of the Bayesian hierarchical framework see Ahn et al., 2011; Navarro et al., 2006; Rouder and Lu, 2005; Rouder et al., 2005, 2008; Scheibehenne and Pachur, 2015; Shiffrin et al., 2008; Wetzels et al., 2010b). For instance, consider a four-parameter Bayesian hierarchical model with four group-level distributions each characterised by two parameters and a group size of 30 participants; this then results in 30×4 individual-level parameters and 2×4 group-level parameters for a total of 128 parameters. In sum, even simple models quickly become complex once hierarchical aspects are introduced and this frustrates the derivation of the marginal likelihood.

To overcome this problem, several Monte Carlo sampling methods have been proposed to approximate the marginal likelihood. In this tutorial we focus on four such methods: the bridge sampling estimator (Bennett, 1976; Chapter 5 of Chen et al., 2012; Meng and Wong, 1996), and its three commonly used special cases—the naive Monte Carlo estimator, the importance sampling estimator, and

the generalised harmonic mean estimator (for alternative methods see Gamerman and Lopes, 2006, Chapter 7; and for alternative approximation methods relevant to model comparison and BMA, see Carlin and Chib (1995) and Green (1995).¹ As we will illustrate throughout this tutorial, the bridge sampler is accurate, efficient, and relatively straightforward to implement (e.g., DiCicco et al., 1997; Frühwirth-Schnatter, 2004; Meng and Wong, 1996).

The goal of this tutorial is to bring the bridge sampling estimator to the attention of mathematical psychologists. We aim to explain this estimator and facilitate its application by suggesting a step-by-step implementation scheme. To this end, we first show how bridge sampling and the three special cases can be used to approximate the marginal likelihood in a simple beta-binomial model. We begin with the naive Monte Carlo estimator and progressively work our way up –via the importance sampling estimator and the generalised harmonic mean estimator– to the most general case considered: the bridge sampling estimator. This order was chosen such that key concepts are introduced gradually and estimators are of increasing complexity and sophistication. The first three estimators are included in this tutorial with the sole purpose of facilitating the reader’s understanding of bridge sampling. In the second part of this tutorial, we outline how the bridge sampling estimator can be used to derive the marginal likelihood for the Expectancy Valence (EV; Busemeyer and Stout, 2002) model, a popular, yet relatively complex reinforcement-learning model for the Iowa gambling task (Bechara et al., 1994). We apply bridge sampling to both an individual-level and a hierarchical implementation of the EV model.

Throughout the chapter, we use the software package R to implement the bridge sampling estimator for the various models. The interested reader is invited to reproduce our results by downloading the code and all relevant materials from our Open Science Framework folder at osf.io/f9cq4.

12.2 Four sampling methods to approximate the marginal likelihood

In this section we outline four standard methods to approximate the marginal likelihood. For more detailed explanations and derivations, we recommend Ntzoufras (2009, Chapter 11) and Gamerman and Lopes (2006, Chapter 7); a comparative review of the different sampling methods is presented in DiCicco et al. (1997).

For concreteness let Y represent the number of correct responses given by a participant in n test items. We assume that Y follows a binomial distribution given by

$$f(d | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \quad (12.2.1)$$

where d refers to the number of successes y and n the number of trials, thus, $y = 0, 1, \dots, n$ and $n \in \mathbb{N}$. Data are assumed to be known. For instance, suppose

¹The appendix gives a derivation showing that the first three estimators are indeed special cases of the bridge sampler.

our participant answered $y = 2$ items correctly in $n = 10$ trials and plugging these observations into Eq. (12.2.1) yields a function of θ , i.e,

$$f(d | \theta) = \binom{10}{2} \theta^2 (1 - \theta)^8. \quad (12.2.2)$$

Such a function is in general known as a likelihood function. The parameter $\theta \in (0, 1)$ can be thought of as the participant's latent ability, which is unknown. To learn θ we assign it a so-called prior distribution $\pi(\theta)$. The prior can be thought of as our knowledge about the participant's ability before we observe the data. For the running example it is computationally convenient to choose a so-called beta distribution for θ with $\alpha, \beta > 0$, that is,

$$\pi(\theta) = \text{Beta}(\theta; \alpha, \beta) = \frac{1}{\mathcal{B}(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad (12.2.3)$$

where $\mathcal{B}(\alpha, \beta)$ is the beta function defined as $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and where $\Gamma(n) = (n-1)!$ whenever $n \in \mathbb{N}$. To ease the exposition, we set $\alpha = \beta = 1$. This choice corresponds to the uniform prior on θ , which is depicted as the dotted line in Fig. 12.1. The uniform prior on θ is interpreted as each value of θ being equally probable.²

Using Bayes rule we can update our prior knowledge about the participant's latent ability θ into a posterior as

$$\pi(\theta | d) = \frac{f(d | \theta) \pi(\theta)}{p(d)}, \quad (12.2.4)$$

where the marginal likelihood $p(d)$ is defined as

$$\underbrace{p(d)}_{\text{Marginal likelihood}} = \int \underbrace{f(d | \theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}} d\theta. \quad (12.2.5)$$

The marginal likelihood makes the posterior for θ a proper probability function so that it integrates to one, which is why $p(d)$ is also referred to as the normalising constant of the posterior. In general, we cannot perform this integral analytically and have to resort to numerical methods such as the bridge sampler, instead.

The running example, however, is chosen in such a way that both the posterior and the target of estimation $p(d) = \int f(d | \theta) \pi(\theta) d\theta$ can be calculated explicitly. Filling in the binomial likelihood and the beta prior, we see that for the running example the posterior is proportional to

$$\pi(\theta | d) \propto \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1}. \quad (12.2.6)$$

Note that this expression is of the same form as the beta distribution given in Eq. (12.2.3). Consequently, the posterior for θ is also a beta distribution, namely

$$\pi(\theta | d) = \text{Beta}(\theta | y + \alpha, n - y + \beta) = \frac{\binom{n}{y} \theta^{y+\alpha-1} (1 - \theta)^{n-y+\beta-1}}{p(d)}, \quad (12.2.7)$$

²But see Ly et al. (2017c).

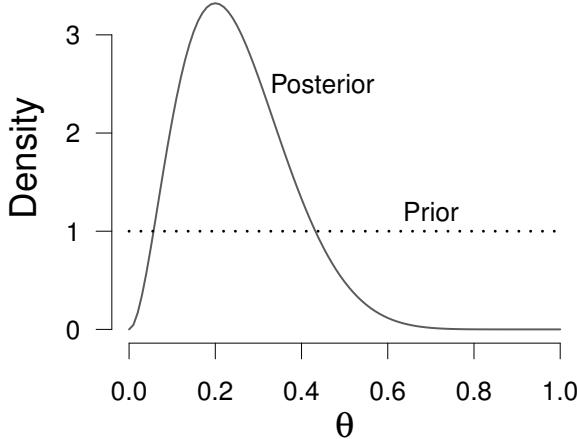


Figure 12.1: Prior and posterior distribution for the rate parameter θ from the beta-binomial model. The Beta(θ ; 1, 1) prior on the parameter θ is represented by the dotted line; the Beta(θ ; 3, 9) posterior distribution is represented by the solid line and was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials.

see the full curve in Fig. 12.1. The denominator is the marginal likelihood and given by

$$p(d) = \int_0^1 f(d | \theta) \pi(\theta) d\theta = \int_0^1 \binom{n}{y} \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1} d\theta \quad (12.2.8)$$

$$= \binom{n}{y} \mathcal{B}(y + \alpha, n - y + \beta). \quad (12.2.9)$$

For $\alpha = \beta = 1$ and the observations $y = 2$ out of $n = 10$, we get

$$p(d) = \binom{10}{2} \mathcal{B}(3, 9) = \frac{1}{11} \approx 0.0909, \quad (12.2.10)$$

the target of the four estimation methods.

12.2.1 Method 1: The naive Monte Carlo estimator of the marginal likelihood

The simplest method to approximate the marginal likelihood is provided by the naive Monte Carlo estimator (Hammersley and Handscomb, 1964; Raftery and

Banfield, 1991). This method uses the standard definition of the marginal likelihood, Eq. (12.2.5), and relies on viewing integrals as sums. The integral implies that the marginal likelihood $p(d)$ is a weighted average of the likelihood where the weights correspond to the prior distribution for the parameters. In other words, the marginal likelihood is the expected value of the likelihood with respect to the prior, that is,

$$p(d) = \mathbb{E}_{\text{prior}}[f(d | \theta)]. \quad (12.2.11)$$

To estimate this population mean, we use a sample mean by sampling, say, K samples from the prior and subsequently average the values of the integrand, the likelihood, at these K samples. This yields the naive Monte Carlo estimator $\hat{p}_1(d)$

$$\hat{p}_1(d) = \underbrace{\frac{1}{K} \sum_{i=1}^K f(d | \tilde{\theta}_i)}_{\text{Average likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim \pi(\theta)}_{\text{samples from the prior distribution}}. \quad (12.2.12)$$

for the target $p(d)$.

12.2.1.1 Running example

To obtain the naive Monte Carlo estimate of the marginal likelihood in our running example, we need K samples from the $\text{Beta}(\theta; 1, 1)$ prior distribution for θ . For illustrative purposes, we limit the number of samples to $K = 12$ whereas in practice one should take K to be very large. To do so in R, we use the command

```
priorSamples <- rbeta(n=12, shape=1, shape=1)
```

and we obtained the following samples

$$\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\} = \{0.58, 0.76, 0.03, 0.93, 0.27, 0.97, 0.45, 0.46, 0.18, 0.64, 0.06, 0.15\}.$$

We use the tilde to make explicit that these values for θ are sampled. All sampled values are represented by the grey dots in Fig. 12.2. Following Eq. (12.2.12), the next step is to evaluate the likelihood, Eq. (12.2.1), at each sampled value $\tilde{\theta}_i$, weight these values by $1/K$, and sum them to obtain the average likelihood $\hat{p}_1(d)$, thus,

$$\hat{p}_1(d) = \frac{1}{12} \sum_{i=1}^{12} f(d | \tilde{\theta}_i) = \frac{1}{12} \sum_{i=1}^{12} \binom{n}{y} \tilde{\theta}_i^y (1 - \tilde{\theta}_i)^{n-y}, \quad (12.2.13)$$

$$= \frac{1}{12} \binom{10}{2} (0.58^2 (1 - 0.58)^8 + \dots + 0.15^2 (1 - 0.15)^8), \quad (12.2.14)$$

$$= 0.0945, \quad (12.2.15)$$

where in the second line we filled in $y = 2$ and $n = 10$. To evaluate the likelihood for the first posterior sample $\tilde{\theta}_1 = 0.58$ we use the command `dbinom(x=2, size=10, prob=0.58)`, while the estimate $\hat{p}_1(d)$ is obtained from the command `1/12*sum(dbinom(x=2, size=10, prob=priorSamples))` in R.

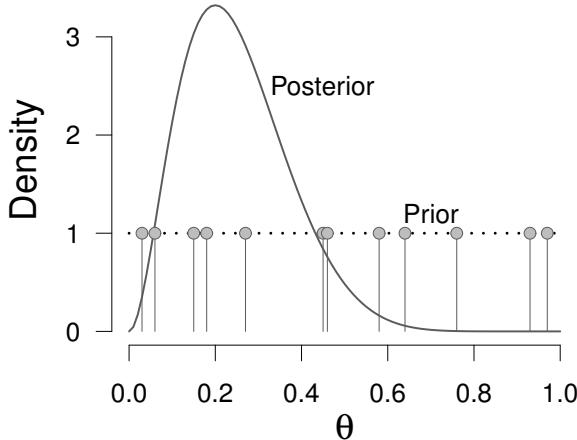


Figure 12.2: Illustration of the naive Monte Carlo estimator for the beta-binomial example. The dotted line represents the prior distribution and the solid line represents the posterior distribution that was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials. The grey dots represent the $K = 12$ samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\}$ randomly drawn from the $\text{Beta}(\theta; 1, 1)$ prior.

12.2.2 Method 2: The importance sampling estimator of the marginal likelihood

The naive Monte Carlo estimator introduced in the last section performs well if the prior and posterior distribution have a similar shape and strong overlap. However, the estimator is unstable if the posterior distribution is peaked relative to the prior (Gamerman and Lopes, 2006; Ntzoufras, 2009). In such a situation, most of the sampled values for θ , say, 98 out of $K = 100$, result in likelihood values close to zero and contribute only minimally to the estimate. This means that those few samples that result in high likelihood values, say, 2 out of $K = 100$, dominate the estimate of the marginal likelihood, which in effect results in high variance of the estimator (Newton and Raftery, 1994; Pajor, 2016).³ The importance sampling estimator, on the other hand, overcomes this shortcoming by boosting sampled values in regions of the parameter space where the integrand of Eq. (12.2.5) is large. This

³The interested reader is referred to Pajor (2016) for a recent improvement on the calculation of the naive Monte Carlo estimator. The proposed improvement involves trimming the prior distribution in such a way that regions with low likelihood values are eliminated, thereby increasing the accuracy and efficiency of the estimator.

is realised by using samples from a so-called importance density $g_{IS}(\theta)$ instead of the prior distribution. The advantage of sampling from an importance density is that values for θ that result in high likelihood values are sampled most frequently, whereas values for θ with low likelihood values are sampled only rarely.

To derive the importance sampling estimator, the definition of the marginal likelihood, Eq. (12.2.5), is once again used as the starting point. The trick is to multiply and divide by the importance density $g_{IS}(\theta)$ as follows

$$\begin{aligned} p(d) &= \int f(d|\theta) \pi(\theta) d\theta = \int f(d|\theta) \pi(\theta) \frac{g_{IS}(\theta)}{g_{IS}(\theta)} d\theta = \int \frac{f(d|\theta) \pi(\theta)}{g_{IS}(\theta)} g_{IS}(\theta) d\theta, \\ &= \mathbb{E}_{g_{IS}(\theta)} \left[\frac{f(d|\theta) \pi(\theta)}{g_{IS}(\theta)} \right]. \end{aligned} \quad (12.2.16)$$

In other words, the marginal likelihood is the expected value of the ratio $\frac{f(d|\theta) \pi(\theta)}{g_{IS}(\theta)}$ with respect to the importance density $g_{IS}(\theta)$. To estimate the population mean $\mathbb{E}_{g_{IS}(\theta)} \left[\frac{f(d|\theta) \pi(\theta)}{g_{IS}(\theta)} \right]$, we use a sample mean by sampling K samples from $g_{IS}(\theta)$ and subsequently average the values of the integrand at these K samples. This yields the importance estimator $\hat{p}_2(d)$

$$\hat{p}_2(d) = \underbrace{\frac{1}{K} \sum_{i=1}^K \frac{f(d|\tilde{\theta}_i) \pi(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}}_{\text{average adjusted likelihood}}, \quad \underbrace{\tilde{\theta}_i \sim g_{IS}(\theta)}_{\substack{\text{samples from the} \\ \text{importance density}}}. \quad (12.2.17)$$

Choosing a suitable importance density is crucial and should (1) be easy to evaluate; (2) have the same domain as the posterior distribution; (3) closely resemble the posterior distribution, and (4) have fatter tails than the posterior distribution (Neal, 2001; Vandekerckhove et al., 2015). The latter criterion ensures that values in the tails of the distribution cannot misleadingly dominate the estimate (Neal, 2001).⁴

12.2.2.1 Running example

To obtain the importance sampling estimate of the marginal likelihood for our running example, we first need to choose an importance density $g_{IS}(\theta)$. An importance density that fulfils the four above mentioned desiderata is a mixture between (i) a first rough approximation of the posterior based on the posterior samples, and (ii) a uniform density across the range of θ (Vandekerckhove et al., 2015).

⁴To illustrate the need for an importance density with fatter tails than the posterior distribution, imagine you sample from the tail region of an importance density with thinner tails. In this case, the numerator in Eq. (12.2.17) would be substantially larger than the denominator resulting in a very large ratio. Since this specific ratio is only one component of the sum displayed in Eq. (12.2.17), this component would dominate the importance sampling estimate. Hence, thinner tails of the importance density run the risk of producing unstable estimates across repeated computations. In fact, the estimator may have infinite variance (e.g., Ionides, 2008; Owen and Zhou, 2000).

The rough approximation of the posterior leads to an importance density that closely resemble the posterior distribution, while the uniform distribution secures that the importance density has the same range as posterior. By mixing these two distributions we get an importance sampler that has thick enough tails. For the rough approximation we use a beta distribution, because we can sample from it easily.

The relative impact of the uniform density is quantified by a mixture weight γ that ranges between 0 and 1. The larger γ , the higher the influence of the uniform density resulting in a less peaked distribution with thick tails. If $\gamma = 1$, the importance density simplifies to the uniform distribution on $[0, 1]$,⁵ and if $\gamma = 0$ the importance density simplifies to the rough first approximation to the posterior distribution.

First, we have to sample $K = 12$ samples from the posterior distribution. Posterior samples can be obtained *without* knowing the normalising constant $p(d)$ using so-called Markov chain Monte Carlo (MCMC) methods that are made accessible through software packages such as WinBUGS, JAGS and Stan. These MCMC methods exploit the fact that the posterior is known up to a constant whenever the observations are given, the likelihood is chosen, and a prior is specified, see for instance Ntzoufras (2009) and Robert (2015) for an introduction.

For the running example, however, we do not need WinBUGS, JAGS or Stan, as we can generate posterior samples directly in R. Recall that for the data at hand, the posterior distribution is proportional to the beta distribution $\text{Beta}(\theta; 3, 9)$ and to obtain, say, $K = 12$ posterior samples we use the R command `rbeta(n=12, shape1=3, shape2=9)`, which for us resulted in

$$\{\ddot{\theta}_1, \ddot{\theta}_2, \dots, \ddot{\theta}_{12}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.$$

We use $\ddot{\theta}_i$ to refer to the i th sample from the posterior distribution to distinguish it from the previously used $\tilde{\theta}_i$ —the i th sample from a distribution other than the posterior distribution, such as a prior distribution or an importance density.

Second, as a first approximation to the posterior we use a beta distribution fitted to these posterior samples using the methods of moments. Recall that a beta distributed random variable $X \sim \text{Beta}(\alpha, \beta)$ has a mean of $\mathbb{E}(X) = \alpha/(\alpha + \beta)$ and a variance of $V(X) = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Filling in the mean $\bar{\ddot{\theta}} = 0.232$ and variance of $s_{\ddot{\theta}}^2 = 0.014$ of our posterior sample $\{\ddot{\theta}_1, \dots, \ddot{\theta}_{12}\}$ and solving for α and β , we retrieve the parameters

$$\hat{\alpha} = \bar{\ddot{\theta}} \left(\frac{\bar{\ddot{\theta}}(1 - \bar{\ddot{\theta}})}{s_{\ddot{\theta}}^2} - 1 \right) = 0.232 \left(\frac{0.232(1 - 0.232)}{0.0142} - 1 \right) = 2.673,$$

$$\hat{\beta} = (1 - \bar{\ddot{\theta}}) \left(\frac{\bar{\ddot{\theta}}(1 - \bar{\ddot{\theta}})}{s_{\ddot{\theta}}^2} - 1 \right) = (1 - 0.232) \left(\frac{0.232(1 - 0.232)}{0.0142} - 1 \right) = 8.865.$$

⁵In our running example, the importance sampling estimator then reduces to the naive Monte Carlo estimator.

As such, we use the beta distribution $\mathcal{B}(\theta; 2.673, 8.865)$ as the first component in the importance sampler.

Third, we choose a mixture weight. With a mixture weight of $\gamma = 0.30$ on the uniform component –a choice that was made to ensure that, visually, the tails of the importance density are clearly thicker than the tails of the posterior distribution– we obtain the following importance density

$$\gamma \text{Beta}(\theta; 1, 1) + (1 - \gamma) \text{Beta}(\theta; \hat{\alpha}, \hat{\beta}) = .3 + .7 \text{ Beta}(\theta; 2.673, 8.865). \quad (12.2.18)$$

Note that the importance distribution is a mixture of the prior and the fitted beta distribution, both from which we can easily draw samples from. This importance density is represented by the dashed line in Fig. 12.3. The figure also shows the posterior distribution (solid line). As is evident from the figure, the beta mixture importance density resembles the posterior distribution, but has fatter tails.

In general, it is advised to choose the mixture weight on the uniform component γ small enough to make the estimator efficient, yet large enough to produce fat tails to stabilise the estimator. A suitable mixture weight can be realised by gradually decreasing the mixture weight and investigating whether stability is still guaranteed (i.e., robustness analysis).

Fourth, to draw one sample from the importance density we first draw a dummy variable Z that takes on the value 1 with 30% chance and 0 otherwise. If $Z = 1$ we draw from the uniform distribution, otherwise we draw from the fitted beta distribution. For $K = 12$ the R code simplifies to

```
K <- 12
numFittedBeta <- rbinom(n=1, size=K, prob=0.3)
postSamples <- c(rbeta(n=numFittedBeta, shape1=2.673, shape2=8.865),
                  rbeta(n=K-numFittedBeta, shape1=1, shape2=1))
```

which for us resulted in

$$\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\} = \{0.11, 0.07, 0.33, 0.25, 0.41, 0.39, 0.25, 0.13, 0.64, 0.26, 0.74, 0.92\}.$$

These samples are represented by the grey dots in Fig. 12.3. The final step is to compute the average adjusted likelihood, i.e., $\frac{f(d|\theta)\pi(\theta)}{g_{IS}(\theta)}$, at the $K = 12$ samples. This yields the following importance sampling estimate for the marginal likelihood

$$\begin{aligned}
 \hat{p}_2(d) &= \frac{1}{12} \sum_{i=1}^{12} \frac{f(d|\tilde{\theta}_i) \pi(\tilde{\theta}_i)}{.3 + .7 \text{ Beta}(\tilde{\theta}_i; 2.673, 8.865)} \\
 &= \frac{1}{12} \left(\frac{\binom{10}{2} 0.11^2 (1 - 0.11)^8 \times 1}{.3 + .7 \text{ Beta}(0.11; 2.673, 8.865)} + \dots + \frac{\binom{10}{2} 0.92^2 (1 - 0.92)^8 \times 1}{.3 + .7 \text{ Beta}(0.92; 2.673, 8.865)} \right) \\
 &= \frac{1}{12} \binom{10}{2} (0.0021 + \dots + 4.7 \times 10^{-9}) \\
 &= 0.0829. \tag{12.2.19}
 \end{aligned}$$

where the .3 in the numerator is multiplied with the density of the prior, which is one for every θ . To evaluate the density of the beta density at the first value

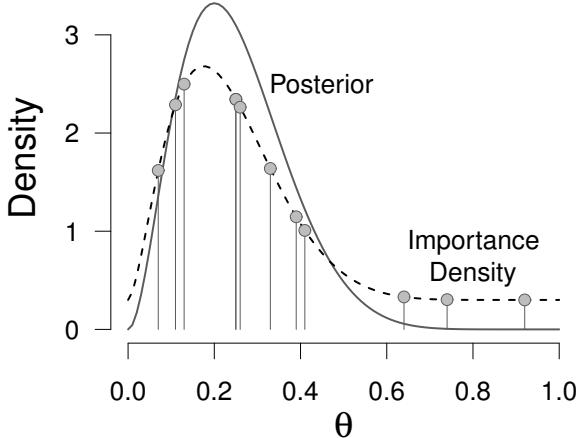


Figure 12.3: Illustration of the importance sampling estimator for the beta-binomial model. The dashed line represents our beta mixture importance density and the solid grey line represents the posterior distribution that was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials. The grey dots represent the $K = 12$ samples $\{\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{12}\}$ randomly drawn from our beta mixture importance density.

of the importance density, say, $\tilde{\theta}_i = 0.11$ we use the command `dbeta(x=0.11, shape1=2.673, shape2=8.865)` in R.

12.2.3 Method 3: The generalised harmonic mean estimator of the marginal likelihood

Just as the importance sampling estimator, the generalised harmonic mean estimator focuses on regions of the parameter space where the integrand of Eq. (12.2.5) is large by using an importance density $g_{IS}(\theta)$ (Gelfand and Dey, 1994).⁶ However, in contrast to the importance sampling estimator, the generalised harmonic mean estimator requires an importance density with thinner tails for an analogous reason as in importance sampling.

To derive the generalised harmonic mean estimator, also known as reciprocal importance sampling estimator (Frühwirth-Schnatter, 2004), we integrate

⁶Note that the generalised harmonic mean estimator is a more stable version of the harmonic mean estimator (Newton and Raftery, 1994). A problem of the harmonic mean estimator is that it is dominated by the samples that have small likelihood values.

$1/p(d) = \frac{\pi(\theta | d)}{f(d | \theta)\pi(\theta)}$ with respect to a proposal density $g_{IS}(\theta)$ that integrates to one, that is,

$$\begin{aligned}\frac{1}{p(d)} &= \int \frac{1}{p(d)} g_{IS}(\theta) d\theta = \int \frac{\pi(\theta | d)}{f(d | \theta)\pi(\theta)} g_{IS}(\theta) d\theta = \int \frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)} \pi(\theta | d) d\theta \\ &= \mathbb{E}_{\text{post}} \left[\frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)} \right],\end{aligned}\quad (12.2.20)$$

which in turn leads to

$$p(d) = \left(\mathbb{E}_{\text{post}} \left[\frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)} \right] \right)^{-1}. \quad (12.2.21)$$

In other words, the reciprocal of the marginal likelihood is the expected value of the ratio $\frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)}$ with respect to the posterior. To estimate the population mean $\mathbb{E}_{\text{post}} \left[\frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)} \right]$, we use a sample mean by sampling K samples from the posterior and subsequently average the values of the integrand at these K samples. This yields the generalised harmonic mean estimator $\hat{p}_3(d)$ (Gelfand and Dey, 1994), where

$$\hat{p}_3(d) = \left(\frac{1}{K} \sum_{j=1}^K \underbrace{\frac{g_{IS}(\hat{\theta}_j)}{f(d | \hat{\theta}_j)\pi(\hat{\theta}_j)}}_{\substack{\text{importance density} \\ \text{likelihood prior}}} \right)^{-1}, \quad \underbrace{\hat{\theta}_j \sim \pi(\theta | d)}_{\substack{\text{samples from the} \\ \text{posterior distribution}}}. \quad (12.2.22)$$

Note that the generalised harmonic mean estimator –in contrast to the importance sampling estimator– evaluates samples from the posterior distribution. Consequently, the sum in Eq. (12.2.22) will contain relatively few terms with $\hat{\theta}_j$ coming from the tail of the posterior. To avoid having the estimator $\hat{p}_3(d)$ miss out on the contribution of the integrand for θ from the tail, we require that for these values of θ that the ratio $\frac{g_{IS}(\theta)}{f(d | \theta)\pi(\theta)}$ itself is already small. This condition implies that $g_{IS}(\theta) < f(d | \theta)\pi(\theta) \propto \pi(\theta | d)$ for θ in the tail of the posterior.

Thus, an importance density for the generalised harmonic mean estimator should (1) have thinner tails than the posterior distribution (Newton and Raftery, 1994; DiCiccio et al., 1997), (2) be easy to evaluate, (3) have the same domain as the posterior distribution, and (4) closely resemble the posterior distribution.

12.2.3.1 Running example

To obtain a generalised harmonic mean estimate of $p(d)$, we need to choose a suitable importance density. The trick is to transform the parameters onto the real line and use a normal distribution fitted to the posterior samples as the importance density. First, we draw $K = 12$ samples from the posterior distribution. Reusing the samples from the last section, we obtain

$$\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{12}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, 0.21, 0.12\}.$$

Second, to fit a normal distribution to the posterior samples, we probit transform the posterior samples $\hat{\xi}_j = \Phi^{-1}(\hat{\theta}_j)$ that range over the entire real line.⁷ For the first sample, we use the command `qnorm(0.22)` in R. Applying this to our particular posterior samples $\hat{\theta}_1, \dots, \hat{\theta}_{12}$ yields

$$\{\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_{12}\} = \{-0.77, -0.99, -1.34, -0.39, -1.55, -0.61, -0.64, -0.23, \\ -0.84, -0.18, -0.81, -1.17\}.$$

These probit-transformed samples are represented by the grey dots in Fig. 12.4.

Third, we search for the normal distribution that provides the best fit to the probit-transformed posterior samples $\hat{\xi}_j$. Using the method of moments, we obtain $\hat{\mu} = -0.793$ and $\hat{\sigma} = 0.423$. Note that the choice of a normal importance density justifies step 2; the probit transformation (or an equivalent transformation) was required to match the range of the posterior distribution to the one of the normal distribution.

Finally, as importance density we choose a normal distribution with mean $\mu = -0.793$ and standard deviation $\sigma = 0.423/1.5$. The additional division by 1.5 ensures that the importance density has thinner tails than the probit-transformed posterior distribution (for a discussion of alternative importance densities see Di Ciccio et al., 1997). We decided to divide $\hat{\sigma}$ by 1.5 for illustrative purposes only. Our importance density is displayed in Fig. 12.4 (dashed line) together with the probit-transformed posterior distribution (solid line).

A generalised harmonic mean estimate can now be obtained using either the original posterior samples $\hat{\theta}_j$ or the probit-transformed samples $\hat{\xi}_j$. Here we choose for the latter option (see also Overstall and Forster, 2010). Incorporating our specific importance density and a correction for using the probit-transformation, Eq. (12.2.22) becomes⁸

$$\hat{p}_3(d) = \left(\frac{1}{K} \sum_{j=1}^K \underbrace{\frac{\frac{1}{\hat{\sigma}} \phi\left(\frac{\hat{\xi}_j - \hat{\mu}}{\hat{\sigma}}\right)}{f(d | \Phi(\hat{\xi}_j)) \phi(\hat{\xi}_j)}}_{\text{likelihood}} \right)^{-1}, \quad \underbrace{\hat{\xi}_j = \Phi^{-1}(\hat{\theta}_j) \text{ and } \hat{\theta}_j \sim \pi(\theta | d)}_{\text{probit-transformed samples from the posterior distribution}} \quad (12.2.23)$$

Note that $\Phi(\hat{\xi}_j) = \hat{\theta}_j$, thus, to evaluate the likelihood, we can simply use the untransformed sample $\hat{\theta}_j$, but to evaluate the importance density and the prior, we have to use the probit-transformed samples $\hat{\xi}_j$ instead. For the particular samples $\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_{12}$ obtained above, we can now use the generalised harmonic

⁷Other transformation are conceivable (e.g., logit transformation).

⁸A detailed explanation is provided in the appendix. Note that using the original posterior samples $\hat{\theta}_j$ would involve transforming the importance density (e.g., the normal density on ξ) to the $(0, 1)$ interval.

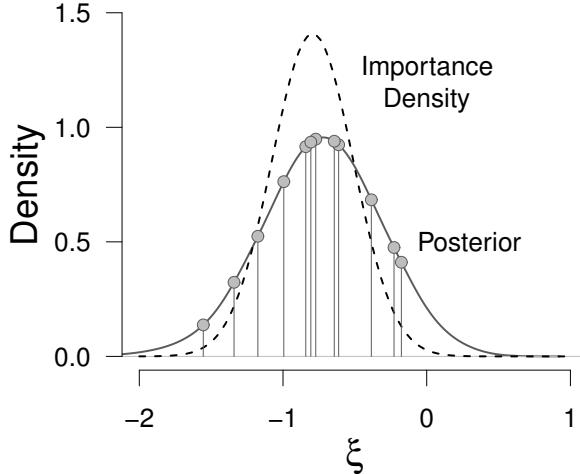


Figure 12.4: Illustration of the generalised harmonic mean estimator for the beta-binomial model. The solid line represents the probit-transformed Beta(θ ; 3, 9) posterior distribution that was obtained after having observed $y = 2$ correct responses out of $n = 10$ trials, and the dashed line represents the importance density $\mathcal{N}(\xi; \mu = -0.793, \sigma = 0.423/1.5)$. The grey dots represent the $K = 12$ probit-transformed posterior samples.

mean estimator to calculate the following estimate for $p(d)$

$$\begin{aligned}
 \hat{p}_3(d) &= \left(\frac{1}{12} \sum_{j=1}^{12} \frac{\frac{1}{0.423/1.5} \phi\left(\frac{\xi_j + 0.793}{0.423/1.5}\right)}{f(d | \Phi(\xi_j)) \phi(\xi_j)} \right)^{-1} \\
 &= \left(\frac{1}{12} \left(\frac{\frac{1}{0.423/1.5} \phi\left(\frac{-0.77 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.22^2 (1 - 0.22)^8 \phi(-0.77)} + \dots + \frac{\frac{1}{0.423/1.5} \phi\left(\frac{-1.17 + 0.793}{0.423/1.5}\right)}{\binom{10}{2} 0.12^2 (1 - 0.12)^8 \phi(-1.17)} \right) \right)^{-1} \\
 &= \left(\frac{1}{12} \frac{1}{\binom{10}{2}} (716.81 + \dots + 556.38) \right)^{-1} \\
 &= 0.092.
 \end{aligned} \tag{12.2.24}$$

For the first posterior sample $\xi_1 = -0.77$, thus, $\hat{\theta}_1 = 0.22$ in the original parameterisation, we evaluate the numerator using the R command `dnorm(x=-0.77, mean=0.793, sd=0.423/1.5)`, the prior using `dnorm(x=-0.77)`, and the likelihood using `dbinom(x=2, size=10, prob=0.22)`.

12.2.4 Method 4: The bridge sampling estimator of the marginal likelihood

As became evident in the last two sections, both the importance sampling estimator and the generalised harmonic mean estimator for $p(d)$ impose strong constraints on the tail behaviour of the importance density relative to the posterior distribution. These conditions make the choice for a suitable importance density complicated, especially when θ is high dimensional. The bridge sampler, on the other hand, alleviates such requirements (e.g., Frühwirth-Schnatter, 2004).

The bridge sampler was originally developed to directly estimate the Bayes factor, that is, the ratio of the marginal likelihoods of two models \mathcal{M}_1 and \mathcal{M}_2 (e.g., Jeffreys, 1961; Kass and Raftery, 1995). However, in this tutorial, we use a version of bridge sampling that allows us to approximate the marginal likelihood of a *single* model (for an earlier application see for example Overstall and Forster, 2010). This version is based on the following identity

$$1 = \frac{\int f(d|\theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int f(d|\theta)\pi(\theta)h(\theta)g(\theta)d\theta}, \quad (12.2.25)$$

where $g(\theta)$ is known as the proposal distribution and $h(\theta)$ the so-called bridge function. Multiplying both sides of Eq. (12.2.25) by the marginal likelihood $p(d)$ leads to

$$\begin{aligned} p(d) &= \frac{\int f(d|\theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int \frac{\int f(d|\theta)\pi(\theta)}{p(d)}h(\theta)g(\theta)d\theta} = \frac{\int f(d|\theta)\pi(\theta)h(\theta)}{h(\theta)g(\theta)} \underbrace{\frac{g(\theta)}{\pi(\theta|d)}}_{\substack{\text{proposal} \\ \text{distribution}}} d\theta \\ &= \frac{\mathbb{E}_{g(\theta)}[f(d|\theta)\pi(\theta)h(\theta)]}{\mathbb{E}_{\text{post}}[h(\theta)g(\theta)]}. \end{aligned} \quad (12.2.26)$$

Hence, the marginal likelihood is the expected value $\mathbb{E}_{g(\theta)}[f(d|\theta)\pi(\theta)h(\theta)]$ with respect to the proposal density $g(\theta)$ divided by the expected value $\mathbb{E}_{\text{post}}[h(\theta)g(\theta)]$ with respect to the posterior distribution. To estimate the population mean of the numerator, we use a sample mean by generating K_2 samples from the proposal distribution and average the integrand $f(d|\theta)\pi(\theta)h(\theta)$ at these samples. Analogously, to estimate the population mean of the denominator, we use a sample mean by generating K_1 samples from the posterior distribution and average the integrand $h(\theta)g(\theta)$ at these samples. This yields the bridge sampling estimator $\hat{p}(d)$

$$\hat{p}(d) = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i)h(\tilde{\theta}_i)}{\frac{1}{K_1} \sum_{j=1}^{K_1} h(\dot{\theta}_j)g(\dot{\theta}_j)}, \quad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\substack{\text{samples from the} \\ \text{proposal distribution}}}, \quad \underbrace{\dot{\theta}_j \sim \pi(\theta|d)}_{\substack{\text{samples from the} \\ \text{posterior distribution}}}. \quad (12.2.27)$$

Conceptually, the proposal distribution is similar to an importance density, it should resemble the posterior distribution, and should have sufficient overlap with the posterior distribution. In fact, we follow Overstall and Forster (2010) and use a normal distribution fitted to probit-transformed samples as the proposal distribution as in the case for the generalised harmonic mean estimator. In our experience, this choice for the proposal distribution works well for a wide range of scenarios. However, this proposal distribution might produce unstable estimates in case of high-dimensional posterior distributions that clearly do not follow a multivariate normal distribution. In such cases, it might be advisable to consider more sophisticated versions of bridge sampling (e.g., Frühwirth-Schnatter, 2004; Meng and Schilling, 2002; Wang and Meng, 2016).

12.2.4.1 Choosing the optimal bridge function

In this tutorial we use the bridge function from Meng and Wong (1996) defined as

$$h(\theta) = C \frac{1}{q_1 f(d | \theta) \pi(\theta) + q_2 p(d) g(\theta)}, \quad (12.2.28)$$

where $q_1 = \frac{K_1}{K_2 + K_1}$, $q_2 = \frac{K_2}{K_2 + K_1}$, and C a constant; its particular value is not required because $h(\theta)$ appears in both the numerator and the denominator of Eq. (12.2.27) and therefore cancels. This particular bridge function is referred to as the “optimal bridge function” because Meng and Wong (1996, p. 837) proved that it minimises the relative mean-squared error, Eq. (12.2.34).

Eq. (12.2.28) shows that the optimal bridge function depends on the marginal likelihood $p(d)$ which is the very entity we want to estimate. We can resolve this issue by applying an iterative scheme that updates an initial guess of the marginal likelihood until the estimate of the marginal likelihood has converged according to a predefined tolerance level. To do so, we insert the expression for the optimal bridge function, Eq. (12.2.28), into Eq. (12.2.27) as was discussed in Meng and Wong (1996). The formula to approximate the marginal likelihood on iteration $t + 1$ is then specified as

$$\hat{p}(d)^{(t+1)} = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{f(d | \tilde{\theta}_i) \pi(\tilde{\theta}_i)}{q_1 f(d | \tilde{\theta}_i) \pi(\tilde{\theta}_i) + q_2 \hat{p}(d)^{(t)} g(\tilde{\theta}_i)}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{g(\dot{\theta}_j)}{q_1 f(d | \dot{\theta}_j) \pi(\dot{\theta}_j) + q_2 \hat{p}(d)^{(t)} g(\dot{\theta}_j)}}, \quad (12.2.29)$$

$$\underbrace{\tilde{\theta}_i \sim g(\theta)}_{\text{samples from the proposal distribution}}, \quad \underbrace{\dot{\theta}_j \sim \pi(\theta | d)}_{\text{samples from the posterior distribution}}, \quad (12.2.30)$$

where $\hat{p}(d)^{(t)}$ denotes the estimate of the marginal likelihood on iteration t of the iterative scheme. Note that Eq. (12.2.29) illustrates why bridge sampling is robust to the tail behaviour of the proposal distribution relative to the posterior distribution; the difference to the importance sampling and generalised harmonic

mean estimator is that, in the case of the bridge sampling estimator, samples from the tail region cannot inflate individual summation terms and thus dominate the estimate. This is because both sums displayed in Eq. (12.2.29) involve a ratio that has a sum in the denominator. Nevertheless it should be noted that the posterior distribution and the proposal distribution need to have sufficient overlap. In the extreme scenario of no overlap the bridge sampling estimator is not defined because both sums of Eq. (12.2.29) would be zero.

To simplify matters, we multiply the numerator of the right side of Eq. (12.2.29) by $\frac{1/g(\tilde{\theta}_i)}{1/g(\tilde{\theta}_i)}$, the denominator by $\frac{1/g(\tilde{\theta}_j)}{1/g(\tilde{\theta}_j)}$, and define $l_{1,j} := \frac{f(d|\tilde{\theta}_j)\pi(\tilde{\theta}_j)}{g(\tilde{\theta}_j)}$ with samples $\tilde{\theta}_j$ from the posterior as in the generalised harmonic mean estimator, and $l_{2,i} := \frac{f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}$ with samples $\tilde{\theta}_i$ from the proposal distribution as in importance sampling. Once we calculated the values $l_{1,j}$ and $l_{2,i}$ for $j = 1, 2, \dots, K_1$ and $i = 1, 2, \dots, K_2$ respectively, we obtain the formula for the iterative scheme of the bridge sampling estimator $\hat{p}_4(d)^{(t+1)}$ at iteration $t + 1$ (Meng and Wong, 1996, p. 837), that is,

$$\begin{aligned} \hat{p}_4(d)^{(t+1)} &= \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i)}{q_1 f(d|\tilde{\theta}_i)\pi(\tilde{\theta}_i) + q_2 \hat{p}_4(d)^{(t)} g(\tilde{\theta}_i)} \frac{1/g(\tilde{\theta}_i)}{1/g(\tilde{\theta}_i)}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{g(\tilde{\theta}_j)}{q_1 f(d|\tilde{\theta}_j)\pi(\tilde{\theta}_j) + q_2 \hat{p}_4(d)^{(t)} g(\tilde{\theta}_j)} \frac{1/g(\tilde{\theta}_j)}{1/g(\tilde{\theta}_j)}} \\ &= \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{l_{2,i}}{q_1 l_{2,i} + q_2 \hat{p}_4(d)^{(t)}}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 l_{1,j} + q_2 \hat{p}_4(d)^{(t)}}}, \quad \underbrace{\tilde{\theta}_i \sim g(\theta)}_{\text{samples from the proposal distribution}}, \quad \underbrace{\tilde{\theta}_j \sim \pi(\theta|d)}_{\text{samples from the posterior distribution}}. \end{aligned} \quad (12.2.31)$$

Eq. (12.2.31) suggests that, in order to obtain a bridge sampling estimate of the marginal likelihood, a number of requirements need to be fulfilled. First, we need K_2 samples from the proposal distribution $g(\theta)$ and K_1 samples from the posterior distribution $\pi(\theta|d)$. Second, for all K_2 samples from the proposal distribution, we evaluate $l_{2,i}$. This involves obtaining the value of the unnormalised posterior (i.e., the product of the likelihood times the prior) and of the proposal distribution for all samples. Third, we evaluate $l_{1,j}$ for all K_1 samples from the posterior distribution. This is analogous to evaluating $l_{2,i}$. Fourth, we need to choose the number of samples K_1 and K_2 for the constants q_1 and q_2 . Fifth, we need an initial guess of the marginal likelihood $\hat{p}_4(d)$. Since some of these five requirements can be obtained easier than others, we will point out possible challenges.

A first challenge is that using a suitable proposal distribution may involve transforming the posterior samples. Consequently, we have to determine how the transformation affects the definition of the bridge sampling estimator for the marginal likelihood, Eq. (12.2.31).

A second challenge is how to use the K_1 samples from the posterior distribution.

One option is to use all K_1 samples for both fitting the proposal distribution and for computing the numerator of the bridge estimator. However, Overstall and Forster (2010) showed that such a procedure may result in an underestimation of the marginal likelihood. To obtain more reliable estimates they propose to divide the posterior samples into two parts; the first part is used to obtain the best-fitting proposal distribution in the numerator of $\hat{p}(d)$, and the second part is used to compute the bridge sampling estimate, thus, the denominator of $\hat{p}(d)$. Throughout this tutorial, we use two equally large parts. In the remainder we therefore state that we draw $2K_1$ samples from the posterior distribution. Out of these $2K_1$ posterior samples, we use samples with even index numbers for the first part; posterior samples with odd index numbers constitute the second part.

To summarise, the discussion of the requirements and challenges encountered in bridge sampling illustrated that the bridge sampling estimator imposes less strict requirements on the proposal distribution than the importance sampling and generalised harmonic mean estimator and allows for an almost automatic application due to the default choice of the bridge function, see also the R package `bridgesampling` by the first author.

12.2.4.2 Running example

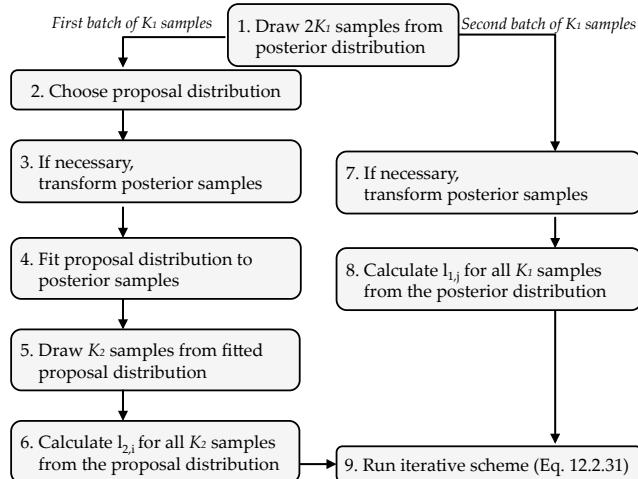


Figure 12.5: Schematic illustration of the steps involved in constructing a bridge sampling estimator for the marginal likelihood.

To obtain a bridge sampling estimate of the marginal likelihood in the beta-binomial example, we follow the eight steps illustrated in Fig. 12.5:

1. We draw $2K_1 = 24$ samples from the $\text{Beta}(\theta; 3, 9)$ posterior distribution for θ .

Using the R command `rbeta(n=24, shape1=3, shape2=9)`, we obtained the following sample of 24 values

$$\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{24}\} = \{0.22, 0.16, 0.09, 0.35, 0.06, 0.27, 0.26, 0.41, 0.20, 0.43, \\ 0.21, 0.12, 0.15, 0.21, 0.24, 0.18, 0.12, 0.22, 0.15, 0.22, \\ 0.23, 0.26, 0.29, 0.28\}.$$

Note that the first 12 samples equal the ones used in the last section.

2. *We choose a proposal distribution.*

Here we opt for an approach that can be easily generalised to models with multiple parameters and select a normal distribution as the proposal distribution $g(\theta)$.⁹

3. *We transform the first batch of K_1 posterior samples.*

Since we use a normal proposal distribution, we have to transform the posterior samples from the rate scale to the real line so that the range of the posterior distribution matches the range of the proposal distribution. This can be achieved by probit-transforming the posterior samples, that is, $\hat{\xi}_j = \Phi^{-1}(\hat{\theta}_j)$ with $j \in \{2, 4, \dots, 24\}$. Using the R command `qnorm(p=firstBatch)` we obtained

$$\{\hat{\xi}_2, \hat{\xi}_4, \dots, \hat{\xi}_{24}\} = \{-0.99, -0.39, -0.61, -0.23, -0.18, -1.17, -0.81, \\ -0.92, -0.77, -0.77, -0.64, -0.58\}.$$

4. *We fit the proposal distribution to the first batch of K_1 probit-transformed posterior samples.*

For the proposal distribution we use a normal distribution fitted with the method of moments. After probit-transforming the first batch of K_1 and applying the R commands `mean` and `sd`, we retrieve $\hat{\mu} = -0.672$ and $\hat{\sigma} = 0.298$. We therefore use $g(\xi) = \frac{1}{0.298} \phi(\frac{\xi + 0.672}{0.298})$ as the proposal density.

5. *We draw K_2 samples from the proposal distribution.*

In R we use the command `rnorm(n=12, mean=-0.672, sigma=0.298)` to sample from the fitted normal proposal and obtained

$$\{\tilde{\xi}_1, \tilde{\xi}_2, \dots, \tilde{\xi}_{12}\} = \{-0.90, -0.55, -1.16, -0.53, -0.45, -0.60, -0.63, -0.48, \\ -0.69, -1.20, -0.65, -0.79\}.$$

6. *We calculate $l_{2,i}$ for all K_2 samples from the proposal distribution.*

For this step we evaluate the likelihood and prior at the samples $\tilde{\xi}_i$ for all $i = 1, 2, \dots, K_2$. Recall that the prior was specified in the original parameterisation θ , while the samples from the proposal ξ range over the real line. The uniform prior in terms of θ implies a standard normal prior in terms of ξ due to the change-of-variables rule, see Appendix 12.C.

⁹There exist several candidates for the proposal distribution. Alternative proposal distributions are, for example, the importance density that we used for the importance sampling estimator or for the generalised harmonic mean estimator

As the likelihood function is also specified in terms of the original parameters θ , we first transform the samples ξ from the proposal distribution that range over the real line to the original parameterisation resulting in $\tilde{\theta}_i = \Phi(\tilde{\xi})$.

Thus, to calculate $l_{2,i}$ for all $i = 1, 2, \dots, K_2$ we evaluate the likelihood with the sample in the original parameterisation $\tilde{\theta}_i$, while we use the samples $\tilde{\xi}_i$ to evaluate the normal proposal density and the prior. This can be done in R with the functions `dbinom` and `dnorm` respectively.

7. *We transform the second batch of the K_1 posterior samples.*

As in step 2, we use the probit transformation and obtained

$$\{\dot{\xi}_1, \dot{\xi}_3, \dots, \dot{\xi}_{23}\} = \{-0.77, -1.34, -1.55, -0.64, -0.84, -0.81, \\ -1.04, -0.71, -1.17, -1.04, -0.74, -0.55\}.$$

8. *We calculate $l_{1,j}$ for the second batch of K_1 probit-transformed samples from the posterior distribution.*

This is analogous to step 6. We plug in the probit-transformed samples $\dot{\xi}_j$ into the prior and the proposal density in terms of ξ , while we use the untransformed samples $\dot{\theta}_j$ for the likelihood.

9. *We run the iterative scheme, Eq. (12.2.31), until our predefined tolerance criterion is reached.*

With $l_{1,j}$ and $l_{2,i}$ at hand we can now run the iterative scheme. As tolerance criterion we choose $|\hat{p}_4(d)^{(t+1)} - \hat{p}_4(d)^{(t)}| / \hat{p}_4(d)^{(t+1)} \leq 10^{-10}$. This requires an initial guess for the marginal likelihood $\hat{p}_4(d)^{(0)}$ which we set to $0.^{10}$

The simplicity of the beta-binomial model allows us to calculate a bridge sampling estimate for the $p(d)$ by hand. To determine $\hat{p}_4(d)^{(t+1)}$ according to Eq. (12.2.31), we need to calculate the constants q_1 and q_2 . Since $K_1 = K_2 = 12$, we obtain $q_1 = q_2 = 0.5$. In addition, we need to calculate $l_{2,i}$ for $i = 1, 2, \dots, 12$ for all samples from the fitted normal proposal distribution, and $l_{1,j}$ for $j = 1, 3, \dots, 23$ for the second batch of the probit-transformed samples from the posterior distribution. Here we show how to calculate $l_{2,1}$ and $l_{1,1}$ using the first sample from the proposal distribution $\tilde{\xi}_1 = -0.9$, thus, $\tilde{\theta}_1 = 0.18$, and the posterior distribution, thus, $\dot{\theta}_1 = 0.22$, thus, $\xi_1 = -0.77$, respectively

$$l_{2,1} = \frac{f(d | \tilde{\theta}_1) \phi(\tilde{\xi}_1)}{g(\tilde{\xi}_1)} = \left(\frac{\binom{10}{2} 0.18^2 (1 - 0.18)^8 \cdot 0.27}{\frac{1}{0.298} \phi\left(\frac{-0.90 + 0.672}{0.298}\right)} \right) = 0.080,$$

$$l_{1,1} = \frac{f(d | \dot{\theta}_1) \phi(\dot{\xi}_1)}{g(\dot{\xi}_1)} = \left(\frac{\binom{10}{2} 0.22^2 (1 - 0.22)^8 \cdot 0.30}{\frac{1}{0.298} \phi\left(\frac{-0.77 + 0.672}{0.298}\right)} \right) = 0.070.$$

¹⁰A better initial guess can be obtained from the generalised harmonic mean estimator explained in the previous section. In our experience, however, the exact choice of the initial value does not seem to influence the convergence of the bridge sampler much.

For $\hat{p}_4(d)^{(t+1)}$, we then get

$$\hat{p}_4(d)^{(t+1)} = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{l_{2,i}}{q_1 l_{2,i} + q_2 \hat{p}_4(d)^{(t)}}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 l_{1,2j-1} + q_2 \hat{p}_4(d)^{(t)}}}, \quad (12.2.32)$$

$$= \frac{\frac{1}{12} \left(\frac{0.080}{0.5 \cdot 0.080 + 0.5 \cdot \hat{p}_4(d)^{(t)}} + \dots + \frac{0.071}{0.5 \cdot 0.085 + 0.5 \cdot \hat{p}_4(d)^{(t)}} \right)}{\frac{1}{12} \left(\frac{1}{0.5 \cdot 0.070 + 0.5 \cdot \hat{p}_4(d)^{(t)}} + \dots + \frac{1}{0.5 \cdot 0.068 + 0.5 \cdot \hat{p}_4(d)^{(t)}} \right)}. \quad (12.2.33)$$

Using $\hat{p}_4(d)^{(0)} = 0$, we obtain as updated estimate of the marginal likelihood $\hat{p}_4(d)^{(1)} = 0.091$. This iterative procedure has to be repeated until our predefined tolerance criterion is reached. For our running example, this criterion is reached after six iterations. We now obtain $\hat{p}_4(d)^{(6)} = 0.0894$ as a bridge sampling estimate of the marginal likelihood $p(d)$.

12.2.5 Interim summary

So far we used the beta-binomial model to illustrate the computation of four different estimators of the marginal likelihood. These four estimators were discussed in order of increasing sophistication, such that the first three estimators provided the proper context for understanding the fourth—the bridge sampler. This estimator is the focus in the remainder of this tutorial. The goal of the next sections is to demonstrate that bridge sampling is particularly suitable to estimate the marginal likelihood of popular models in mathematical psychology. Importantly, bridge sampling may be used to obtain accurate estimates of the marginal likelihood of hierarchical models (for a detailed comparison of bridge sampling versus its special cases see Frühwirth-Schnatter, 2004; Sinharay and Stern, 2005).

12.2.6 Assessing the accuracy of the bridge sampling estimator

In this section we show how to quantify the accuracy of the bridge estimator. A straightforward approach would be to apply the bridge estimator multiple times and investigate the variability of the marginal likelihood estimate. In practice, however, this solution is often impractical due to the substantial computational burden of obtaining the posterior samples and evaluating the relevant quantities in the bridge sampling procedure.

Frühwirth-Schnatter (2004) proposed an alternative approach that approximates the estimator's expected relative mean-squared error

$$RE^2 = \frac{\mathbb{E}[(\hat{p}_4(d) - p(d))^2]}{p(d)^2} \quad (12.2.34)$$

The derivation of this approximate relative mean-squared error by Frühwirth-Schnatter takes into account that the samples from the proposal distribution

$g(\theta)$ are independent, whereas the MCMC samples from the posterior distribution $\pi(\theta | d)$ may be autocorrelated. The approximate relative mean-squared error is given by

$$\widehat{RE}^2 = \frac{1}{K_2} \frac{V_{g(\theta)}[r_1(\theta)]}{\mathbb{E}_{g(\theta)}^2[r_1(\theta)]} + \frac{\rho_{r_2}(0)}{K_1} \frac{V_{\text{post}}[r_2(\theta)]}{\mathbb{E}_{\text{post}}^2[r_2(\theta)]}, \quad (12.2.35)$$

where $r_1(\theta) = \frac{\pi(\theta | d)}{q_1\pi(\theta | d) + q_2g(\theta)}$, $r_2(\theta) = \frac{g(\theta)}{q_1\pi(\theta | d) + q_2g(\theta)}$, and where $V_{g(\theta)}[r_1(\theta)] = \int (r_1(\theta) - \mathbb{E}[r_1(\theta)])^2 g(\theta) d\theta$ denotes the variance of $r_1(\theta)$ with respect to the proposal distribution $g(\theta)$ (the variance $V_{\text{post}}[r_2(\theta)]$ is defined analogously), and $\rho_{r_2}(0)$ corresponds to the normalised spectral density of the autocorrelated process $r_2(\theta)$ at the frequency 0.

In practice, we approximate the unknown variances and expected values by the corresponding sample variances and means. Hence, for evaluating the variance and expected value with respect to $g(\theta)$, we use the K_2 samples for $\tilde{\theta}_i$ from the proposal distribution. To evaluate the variance and expected value with respect to the posterior distribution, we use the second batch of K_1 samples $\hat{\theta}_j$ from the posterior distribution which we also use in the iterative scheme for computing the marginal likelihood. Because the posterior samples are obtained via an MCMC procedure and are, hence, autocorrelated, the second term in Eq. (12.2.35) is adjusted by the normalised spectral density (for details see Frühwirth-Schnatter, 2004). The spectral density at frequency zero can be estimated by first fitting an autoregressive model using the `spectrum0.ar()` function as implemented in the `coda` R package (Plummer et al., 2006). To evaluate the normalised posterior density which appears in the numerator of $r_1(\theta)$ and the denominator of both $r_1(\theta)$ and $r_2(\theta)$, we use the bridge sampling estimate as normalising constant.

Note that, under the assumption that the bridge sampling procedure $\hat{p}_4(d)$ is an unbiased estimator of the marginal likelihood $p(d)$, the square root of the expected relative mean-squared error, Eq. (12.2.34), can be interpreted as the coefficient of variation (i.e., the ratio of the standard deviation and the mean; Brown, 1998). In the remainder of this chapter, we report the coefficient of variation to quantify the accuracy of the bridge sampling estimator.

12.3 Case study: Bridge sampling for reinforcement learning models

In this section, we illustrate the computation of the marginal likelihood using bridge sampling in the context of a published data set (Busemeyer and Stout, 2002) featuring the Expectancy Valence (EV) model—a popular reinforcement learning model for the Iowa gambling task (IGT; Bechara et al., 1994). We first introduce the task and the model, and then use bridge sampling to estimate the marginal likelihood of the EV model implemented in both an individual-level and a hierarchical Bayesian framework. For the individual-level framework, we compare estimates obtained from bridge sampling to importance sampling estimates published in Steingroever et al. (2016b). For the hierarchical framework, we compare our results to estimates from the Savage-Dickey density ratio test (Dickey, 1971; Dickey and Lientz, 1970; Wagenmakers et al., 2010; Wetzels et al., 2010a).

12.3.1 The Iowa gambling task

In this section we describe the IGT (see also Steingroever et al., 2013a, 2013b, 2013c, 2014, 2016a, 2016b). Originally, Bechara et al. (1994) developed the IGT to distinguish decision-making strategies of patients with lesions to the ventromedial prefrontal cortex from the ones of healthy controls (see also Bechara et al., 1998, 1999, 2000). During the last decades, the scope of application of the IGT has increased tremendously covering clinical populations with, for example, pathological gambling tendencies (Cavedini et al., 2002b), obsessive-compulsive disorder (Cavedini et al., 2002a), psychopathic tendencies (Blair et al., 2001), and schizophrenia (Bark et al., 2005; Martino et al., 2007).

The IGT is a card game that requires participants to choose, over several rounds, cards from four different decks in order to maximise their long-term net outcome Bechara et al. (1994, 1997). The four decks differ in their payoffs, and two of them result in negative long-term outcomes (i.e., the bad decks), whereas the remaining two decks result in positive long-term outcomes (i.e., the good decks). After each choice, participants receive feedback on the rewards and losses (if any) associated with that card, as well as their running tally of net outcomes over all trials so far. Unbeknownst to the participants, the task (typically) contains $N = 100$ trials.

Table 12.1: Summary of the payoff scheme of the traditional IGT as developed by Bechara et al. (1994)

	Deck A	Deck B	Deck C	Deck D
	Bad deck with frequent losses	Bad deck with infrequent losses	Good deck with frequent losses	Good deck with infrequent losses
Reward/trial	100	100	50	50
Number of losses/10 cards	5	1	5	1
Loss/10 cards	-1250	-1250	-250	-250
Net outcome/10 cards	-250	-250	250	250

A question is whether and to what extent participants eventually learn to prefer the good decks that allow them to maximise their long-term net outcome. The good decks are typically labelled as decks C and D, whereas the bad decks are labelled as decks A and B. Table 12.1 presents a summary of the traditional payoff scheme as developed by Bechara et al. (1994). This table illustrates that decks A and B yield high constant rewards, but even higher unpredictable losses, thus, a negative long-term net outcome. Decks C and D, on the other hand, yield low constant rewards, but even lower unpredictable losses: hence, the long-term net outcome is positive. In addition to the different payoff magnitudes, the decks also differ in the frequency of losses: decks A and C yield frequent losses, while decks B and D yield infrequent losses.

Table 12.2: Example data of chosen decks y_n and experienced payoffs x_n for $n = 10$ trials.

Trial n	1	2	3	4	5	6	7	8	9	10
y_n	D	C	B	A	D	C	B	A	B	B
x_n	50	50	100	100	50	50	100	100	100	-1250

12.3.2 The Expectancy Valence model

In this section, we describe the EV model (see also Steingroever et al., 2013b, 2014, 2016a, 2016b). Originally proposed by Busemeyer and Stout (2002), the EV model is arguably the most popular model for the IGT (for references see Steingroever et al., 2013b, and for alternative IGT models see Ahn et al., 2008; Dai et al., 2015; Steingroever et al., 2014; Worthy et al., 2013; Worthy and Maddox, 2014).

The model specifies how previous experienced payoffs affect the participant's next choice in the IGT through the interaction of three model parameters $\theta = (w, a, c)$ that represent distinct psychological processes. In essence, the participant's next choice at trial $n + 1$ depends on her expected utility for each deck based on the previous n trials and her willingness to exploit this knowledge. It is assumed that the deck with the highest expected utility at the trial n has the largest probability to be chosen in the next trial. The participant's expected utilities evolve over time and depend on previously choices $y^n = (y_1, y_2, \dots, y_n)$ through the subsequently observed payoffs $x^n = (x_1, x_2, \dots, x_n)$, such as the ones depicted in Table 12.2. The model assumes that the participant summarises the experienced payoffs for each deck $y \in \mathcal{Y} = \{A, B, C, D\}$ with a weighted mean of the experienced wins $W_n(y)$ and losses $L_n(y)$ to obtain the utility $v_n(y | w)$ where

$$v_n(y | w) = (1 - w)W_n(y) + wL_n(y). \quad (12.3.1)$$

Hence, at any trial n there are four utilities $v_n(y | w)$. The weight that the participant assigns to losses relative to rewards is referred to as the attention weight parameter w . A small value of w , that is, $w < 0.5$, is characteristic for decision makers who put more weight on immediate rewards and can thus be described as reward-seeking, whereas a large value of w , that is, $w > 0.5$, is characteristic for decision makers who put more weight on the immediate losses and can, thus, be described as loss averse (Ahn et al., 2008; Busemeyer and Stout, 2002).

The actually observed utility $v_n(y | w)$ corresponding to the chosen deck y at trial n after observing the payoff x_n might be higher or lower than what the decision maker expected about deck y . We write $Ev_{n-1}(y | w)$ for the expected utility of deck y extracted from information up to and including trial $n-1$. That is, before the participant has made her choice $Y_n = y$ and before she is presented with the payoff x_n at trial n . If the observed utility $v_n(y | w)$ is higher (lower) than what was expected $Ev_{n-1}(y | w)$, then the expected utility for deck y is shifted upwards (downwards) for the next trial $n + 1$. This learning process is described by the delta learning rule, also known as the Rescorla-Wagner rule (Rescorla and Wagner,

1972) and formalised as

$$Ev_n(y | w, a) = Ev_{n-1}(y | w) + a[v_n(y | w) - Ev_{n-1}(y | w)], \text{ for } y \in \mathcal{Y}, \quad (12.3.2)$$

where the parameter a quantifies the memory for rewards and losses. A value of a close to zero indicates slow forgetting and weak recency effects, whereas a value of a close to one indicates rapid forgetting and strong recency effects.

We set $Ev_0(y | w) = 0$ for every deck y to convey that the participant has no prior knowledge about the payoffs of the decks. Furthermore, we assume that at trial n only the expected utility of the chosen deck is updated and that the expected utility of the decks that are not chosen stay untouched. Consequently, Eq. (12.3.2) implies that the expected utility of each deck remains zero until the first time the deck is chosen. For instance, for the example data in Table 12.2 the expected utility of deck A is zero, until $n = 4$. The change of expected utility for deck A then plays a role relative to the expected utilities of decks B, C and D in the next trials. This value of the updated expected utility for deck A remains the same from trial $n+1 = 5$ up to trial 8, but before she is presented with the payoff at trial 8.

We assume that the probability with which the participant chooses deck y at trial $n+1$ is given by the following softmax choice rule¹¹

$$Pr(Y_{n+1} = y | x^n, \theta) = \frac{e^{u(n) \cdot Ev_n(y | w, a)}}{\sum_{y \in \mathcal{Y}} e^{u(n) \cdot Ev_n(y | w, a)}}, \text{ for } y \in \mathcal{Y}. \quad (12.3.3)$$

The function u measures how sensitive the participant is to the expected utilities collected up to trial n for the decision at trial $n+1$. Values of $u(n)$ close to zero indicate random choice behaviour (i.e., strong exploration), whereas large values of $u(n)$ indicate choice behaviour that is strongly determined by the expected utilities (i.e., strong exploitation). We parameterise the between-trial-dependent sensitivity function $u(n)$ with the following function

$$u(n) = (n/10)^c, \text{ for } n = 1, 2, \dots, N, \quad (12.3.4)$$

where $c \in [-5, 5]$. For positive c , successive choices become less random and more determined by the expected deck utilities; if c is negative, successive choices become more random and less determined by the expected deck utilities, a pattern that is clearly non-optimal. We restricted the consistency parameter c of the EV model to the range $[-2, 2]$ instead of the proposed range $[-5, 5]$ (Busemeyer and Stout, 2002). This modification improved the estimation of the EV model and prevented the choice rule from producing numbers that exceed machine precision (see also Steingroever et al., 2014).

In sum, to specify how past experience is processed for the choice in the next trial the EV model uses three parameters $\theta = (w, a, c)$: (1) the attention weight

¹¹This rule is also known as the ratio-of-strength choice rule (Luce, 1959). Furthermore, we wrote Y_{n+1} for the random choice the participant will make before seeing the payoff x_{n+1} . After we observed x_{n+1} , we write y_{n+1} for the deck that is chosen to convey that it is not random anymore.

parameter $w \in [0, 1]$ quantifies the weight of losses over rewards at each trial n , (2) the updating parameter $a \in [0, 1]$ determines how the observed utility $v_n(y | w)$ of the choosing deck y affects the expected utility for the next trial, and (3) the response consistency parameter $c \in [-2, 2]$ determines the balance between exploitation and exploration.

12.3.3 Data

We applied bridge sampling to a data set published by Busemeyer and Stout (2002). The data set consists of $S = 30$ healthy participants each contributing $N = 100$ IGT card selections (see Busemeyer and Stout for more details on the data sets).¹²

12.3.4 Application of bridge sampling to an individual-level implementation of the EV model

In this section we describe how we use bridge sampling to estimate the marginal likelihood of an individual-level implementation of the EV model. This implementation estimates model parameters for each participant separately. We therefore have as many data sets d_1, d_2, \dots, d_S as there are participants. Accordingly, we also obtain a marginal likelihood of the EV model for every participant. The likelihood of the s th participant follows from plugging in the sequence of observed choices y_s^N and payoffs x_s^N into Eq. (12.3.3) gradually resulting in

$$f(d_s | \theta_s) = Pr(Y_{s,1} = y_{s,1} | x^0, \theta_s) \times Pr(Y_{s,2} = y_{s,2} | x_s^1, \theta_s) \quad (12.3.5)$$

$$\times Pr(Y_{s,3} = y_{s,3} | x_s^2, \theta_s) \times \cdots \times Pr(Y_{s,N} = y_{s,N} | x_s^{N-1}, \theta_s), \quad (12.3.6)$$

where $x^0 = 0$ resulting in $Ev_0(y) = 0$ for every deck $y = A, B, C, D$, and $\theta_s = (w_s, a_s, c_s)$ as before. For each individual s we use the uniform priors $w_s \sim U[0, 1]$, $a_s \sim U[0, 1]$, $c_s \sim U[-2, 2]$. As a result, we have

$$p(d_s | \text{Ind}_s) = \int \int \int f(d_s | w_s, a_s, c_s) \frac{1}{4} dw_s da_s dc_s, \quad (12.3.7)$$

for every participant s , see Steingroever et al. (2016b) for more details on this prior choice and model implementations.

12.3.4.1 Schematic execution of the bridge sampler

To obtain a bridge sampling estimate of the marginal likelihood for the s th participant where $s = 1, 2, \dots, 30$ we follow the steps outlined in Fig. 12.5. We proceed as follows:

1. *For each parameter, we draw $2K_{s,1}$ samples from the posterior distribution.*
We use the posterior samples from Steingroever et al. (2016b) who fit an individual-level implementation of the EV model separately to the data of

¹²Note that we excluded three participants due to incomplete choice data.

each participant in Busemeyer and Stout (2002). For each participant we have at least 5,000 posterior samples; whenever this number of samples was insufficient to ensure convergence of the Hamiltonian Monte Carlo (HMC) chains, Steingroever et al. (2016b) repeated the fitting routine with 5,000 additional samples. Steingroever et al. (2016b) confirmed convergence of the HMC chains by reporting that all \hat{R} statistics were below 1.05. The posterior samples were split into two batches each consisting of $K_{s,1}$ number of samples.

2. *We choose a proposal distribution.*

We generalise our approach from the running example and use a multivariate normal distribution as a proposal distribution.

3. *We transform the first batch of $K_{s,1}$ posterior samples.*

Since we use a multivariate normal distribution as a proposal distribution, we transform all posterior samples to the real line using the probit function, that is, we obtain $\dot{\xi}_{s,j} = (\dot{\omega}_{s,j}, \dot{\alpha}_{s,j}, \dot{\gamma}_{s,j}) \in \mathbb{R}^3$, where $\dot{\omega}_{s,j} = \Phi^{-1}(\check{w}_{s,j})$, $\dot{\alpha}_{s,j} = \Phi^{-1}(\check{a}_{s,j})$, $\dot{\gamma}_{s,j} = \Phi^{-1}((\check{c}_{s,j} + 2)/4)$ for $j = 2, 4, \dots, 2K_1$.

4. *We fit the proposal distribution to the first batch of K_1 probit-transformed posterior samples.*

We use method-of-moment estimates and use the mean vector and the covariance matrix obtained from the first batch of $K_{s,1}$ probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. *We draw $K_{s,2}$ samples from the proposal distribution.*

We use R to randomly draw K_2 samples from the proposal distribution obtained in step 4.

6. *We calculate $l_{s,2,i}$ for all $K_{s,2}$ samples from the proposal distribution.*

For this step we evaluate the likelihood and prior at the samples $\tilde{\xi}_{s,i}$ for all $i = 1, 2, \dots, K_{s,2}$. Recall that the prior was specified in the original parameterisation, while the samples range over the real line. The uniform priors in terms of θ change into standard normal priors in terms of ξ due to the change-of-variables rule as before, see Appendix 12.C and Appendix 12.D for a more detailed explanation.

As the likelihood function is specified in terms of the parameters θ_s , we first transform the proposal samples that range over the real line to the original parameterisation resulting in $\tilde{\theta}_{s,i} = (\tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i})$, where $\tilde{w}_{s,i} = \Phi(\tilde{\omega}_{s,i})$, $\tilde{a}_{s,i} = \Phi(\tilde{\alpha}_{s,i})$ and $\tilde{c}_{s,i} = 4\Phi(\tilde{\gamma}_{s,i}) - 2$.

Thus, to calculate $l_{s,2,i}$ for all $i = 1, 2, \dots, K_{s,2}$ we evaluate the likelihood with the sample in the original parameterisation $\tilde{\theta}_{s,i}$, while we use the samples $\tilde{\xi}_{s,i}$ to evaluate the multivariate normal proposal density and the prior, that is,

$$\frac{f(d_s | \tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i}) \overbrace{\phi(\tilde{\omega}_{s,i})\phi(\tilde{\alpha}_{s,i})\phi(\tilde{\gamma}_{s,i})}^{\pi(\tilde{\xi}_{s,i})}}{g(\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})}, \quad (12.3.8)$$

where ϕ denotes the standard normal density.

7. We transform the second batch of $K_{s,1}$ posterior samples.

This is analogous to step 2.

8. We calculate $l_{s,1,j}$ for the second batch of $K_{s,1}$ probit-transformed samples from the posterior distribution.

This is analogous to step 6. We plug in the probit-transformed samples $\xi_{s,j}$ into the prior and the proposal density in terms of ξ_s , while we use the untransformed samples $\hat{\theta}_{s,j}$ for the likelihood.

9. We run the iterative scheme, Eq. (12.2.31), until our predefined tolerance criterion is reached.

With $l_{s,1,j}$ and $l_{s,2,i}$ at hand we can now run the iterative scheme. We use the same tolerance criterion and initialisation $\hat{p}_4(d)^0 = 0$ as in running example. Once convergence is reached, we obtain an estimate of the marginal likelihood for each participant, and derive the coefficient of variation for each participant using Eq. (12.2.35). The largest coefficient of variation is 1.94% suggesting that the bridge sampler has low variance.¹³

12.3.4.2 Assessing the accuracy of our implementation

To assess the accuracy of our implementation, we compared the marginal likelihood estimates obtained with our bridge sampler to the estimates obtained with importance sampling (Steingroever et al., 2016b). Fig. 12.6 shows the logarithm of the marginal likelihoods $p(d_1 | \text{Ind}_1), p(d_2 | \text{Ind}_2), \dots, p(d_S | \text{Ind}_S)$ for the $S = 30$ participants of Busemeyer and Stout (2002) obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016b; y-axis). The two sets of estimates correspond almost perfectly. These results indicate a successful implementation of the bridge sampler. Thus, this section emphasises that both the importance sampler and bridge sampler can be used to estimate the marginal likelihood for the data of individual participants. However, when we want to estimate the marginal likelihood of a Bayesian hierarchical model, it may be difficult to find a suitable importance density. The bridge sampler, on the other hand, can be applied more easily and more efficiently.

12.3.5 Application of bridge sampling to a hierarchical Implementation of the EV model

In this section we illustrate how bridge sampling is used to estimate the marginal likelihood of a hierarchical EV model. This hierarchical implementation assumes that the parameters w_s , a_s , and c_s from each participant are drawn from three separate group-level distributions. This model specification incorporates both the differences and the similarities between participants. We illustrate this application with the data from Busemeyer and Stout (2002) as we have done before, but we now also assume that these participants belong to one group.

¹³Note that this measure relates to the marginal likelihoods, not to the log marginal likelihoods.

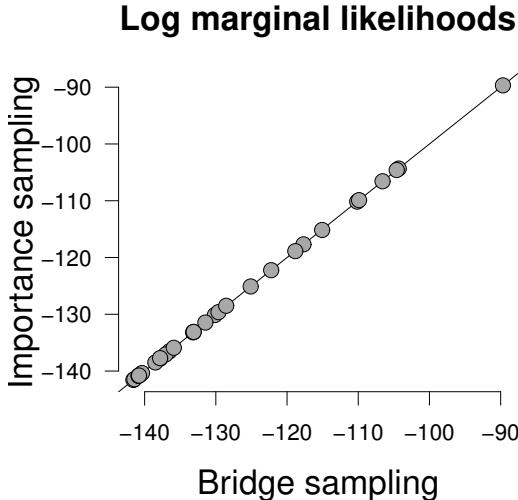


Figure 12.6: Comparison of the log marginal likelihoods obtained with bridge sampling (x-axis) and importance sampling reported by Steingroever et al. (2016b, y-axis). The main diagonal indicates perfect correspondence between the two methods.

This one group assumption is formalised using a hierarchical model, where the s th participant's parameters θ_s are drawn from a group distribution $f(\theta_s | \eta)$. Given the group-level distribution $f(\theta_s | \eta)$ with group-level parameters η , the data of the individuals are conditionally independent which implies that the likelihood is given by

$$f(d_{\text{all}} | \theta_1, \dots, \theta_S, \eta) = \prod_{s=1}^S f(d_s | \theta_s) f(\theta_s | \eta) d\theta_s, \quad (12.3.9)$$

where $f(d_s | \theta_s)$ is the likelihood of each individual as specified in Eq. (12.3.5). We focus on the group-level parameters η for inference about the group. For a posterior on η , we have to choose a prior $\pi(\eta)$, and the resulting posterior has as normalising constant

$$p(d_{\text{all}} | \text{Hier}) = \int f(d_{\text{all}} | \theta_1, \dots, \theta_S, \eta) \pi(\eta) d\eta. \quad (12.3.10)$$

We use bridge sampling to estimate $p(d_{\text{all}} | \text{Hier})$.

12.3.5.1 Schematic execution of the bridge sampler

To compute the marginal likelihood, we again follow the steps outlined in Fig. 12.5, with a few minor modifications.

1. For each parameter, that is, all individual-level and group-level parameters, we draw $2K_1 = 60,000$ samples from the posterior distribution.

To obtain the posterior samples, we fit a hierarchical Bayesian implementation of the EV model to the Busemeyer and Stout (2002) data set using the JAGS software package (Plummer, 2003).¹⁴ Each participant's parameters are assumed to be drawn from a group-level distribution. As group-level distribution we use a normal distribution characterised by the group-level means and standard deviation parameters expressed in terms of the probit-transformed parameters $\xi_s = (\omega_s, \alpha_s, \gamma_s)$.¹⁵ More specifically, we assume that the s th participant's probit-transformed parameters are drawn from the following distribution

$$f(\xi_s | \eta) = \mathcal{N}(\omega_s | \mu_\omega, \sigma_\omega^2) \mathcal{N}(\alpha_s | \mu_\alpha, \sigma_\alpha^2) \mathcal{N}(\gamma_s | \mu_\gamma, \sigma_\gamma^2), \quad (12.3.11)$$

where $\eta = (\eta_\mu, \eta_\sigma)$ with $\eta_\mu = (\mu_\omega, \mu_\alpha, \mu_\gamma)$ and $\eta_\sigma = (\sigma_\omega, \sigma_\alpha, \sigma_\gamma)$. For inference about these group-level parameters we use standard normal priors on the group means and uniform priors between 0 and 1.5 on the group-level standard deviations, that is, $\pi(\eta) = \pi(\eta_\mu)\pi(\eta_\sigma)$, where

$$\pi(\eta_\mu) = \mathcal{N}(\mu_\omega; 0, 1)\mathcal{N}(\mu_\alpha; 0, 1)\mathcal{N}(\mu_\gamma; 0, 1), \quad (12.3.12)$$

$$\pi(\eta_\sigma) = U(\sigma_\omega; 0, 1.5)U(\sigma_\alpha; 0, 1.5)U(\sigma_\gamma; 0, 1.5). \quad (12.3.13)$$

For a detailed explanation of the hierarchical implementation of the EV model, see Wetzels et al. (2010b).

To reach convergence and reduce autocorrelation, we collect two MCMC chains, each with 120,000 samples from the posterior distributions after having excluded the first 30,000 samples as burn-in. Out of these 120,000 samples per chain, we retained every 4th value yielding 30,000 samples per chain. This setting resulted in all \hat{R} statistics below 1.05 suggesting that all chains have successfully converged from their starting values to their stationary distributions. The resulting posterior samples of the six group-level parameters η together with the three individual-level parameters θ_s for every participant in the group of $S = 30$ individuals are used in the bridge sampler to estimate the marginal likelihood $p(d_{\text{all}} | \text{Hier})$.

2. We choose a proposal distribution.

We use a multivariate normal distribution as a proposal distribution.

3. We transform the first batch of K_1 posterior samples.

As before, we ensure that the range of the posterior distribution matches the range of the proposal distribution by using the probit transformation as described above. Hence, for the even samples of the posteriors $j = 2, 4, \dots, 2K_1$ and each $s = 1, \dots, 30$ we write $\dot{\xi}_{s,j} = (\dot{\omega}_{s,j}, \dot{\alpha}_{s,j}, \dot{\gamma}_{s,j})$ for the probit-transformed individual parameters, and $\dot{\zeta}_{s,j} = (\dot{\tau}_{\omega,j}, \dot{\tau}_{\alpha,j}, \dot{\tau}_{\gamma,j})$ for the the probit-transformed group-level standard deviations, where $\dot{\tau}_{\omega,j} =$

¹⁴We used a model file that is an adapted version of the model file used by Ahn et al. (2011).

¹⁵As before we define $\omega_s = \Phi^{-1}(w_s)$, $\alpha_s = \Phi^{-1}(a_s)$, $\gamma_s = \Phi^{-1}((c_s + 2)/4)$.

$\Phi^{-1}((\hat{\sigma}_{\omega,j}) / 1.5)$, $\hat{\tau}_{\alpha,j} = \Phi^{-1}((\hat{\sigma}_{\alpha,j}) / 1.5)$, $\hat{\tau}_{\gamma,j} = \Phi^{-1}((\hat{\sigma}_{\gamma,j}) / 1.5)$. The group-level mean parameters do not have to be transformed because they already range across the entire real line.

4. We fit the proposal distribution to the first batch of the K_1 probit-transformed posterior samples.

We use method-of-moment estimates for the mean vector and the covariance matrix obtained from the first batch of K_1 probit-transformed posterior samples to specify our multivariate normal proposal distribution.

5. We draw K_2 samples from the proposal distribution.

We use R to randomly draw K_2 samples from the proposal distribution fitted in step 4. For $i = 1, 2, \dots, K_2$ we obtain group-level parameters $\tilde{\eta}_i = (\tilde{\eta}_{\mu,i}, \tilde{\zeta}_{\sigma,i})$, and $S = 30$ number of individual-levels parameter each consisting of $\tilde{\xi}_{s,i} = (\tilde{\omega}_{s,i}, \tilde{a}_{s,i}, \tilde{\gamma}_{s,i})$.

6. We calculate $l_{2,i}$ for all K_2 samples from the proposal distribution.

For this step we evaluate the prior, the group-level distribution at the samples $\tilde{\eta}_{\mu,i}, \tilde{\zeta}_{\sigma,i}$ and, subsequently, all individual-level likelihood functions at the samples $\tilde{\xi}_{1,i}, \tilde{\xi}_{2,i}, \dots, \tilde{\xi}_{S,i}$ for $S = 30$. As the prior on the group-level means are already specified as normal distributions, we simply need to evaluate the normal density at the samples $\tilde{\eta}_{\mu,i}$. On the other hand, the prior on the group-level standard deviations were specified in the original parameterisation, while the samples $\tilde{\zeta}_{\sigma,i}$ range over the real line. The uniform priors in terms of the σ s change into standard normal priors in terms of ζ due to the change-of-variables rule as before, see Appendix 12.C and Appendix 12.E for a more detailed explanation.

As the likelihood $f(d_{\text{all}} | \theta_1, \dots, \theta_S, \eta)$ and the group-level distribution are in terms of the group-level σ s and the original parameterisation θ_s , we transform the proposal samples that range over the real line to the original parameterisation resulting in $\tilde{\eta}_{\sigma,i} = (\tilde{\sigma}_{\omega,i}, \tilde{\sigma}_{\alpha,i}, \tilde{\sigma}_{\gamma,i})$ for the group-level standard deviations, and $\tilde{\theta}_{s,i} = (\tilde{w}_{s,i}, \tilde{a}_{s,i}, \tilde{c}_{s,i})$, where $\tilde{w}_{s,i} = \Phi(\tilde{\omega}_{s,i})$, $\tilde{a}_{s,i} = \Phi(\tilde{\alpha}_{s,i})$, and $\tilde{c}_{s,i} = 4\Phi(\tilde{\gamma}_{s,i}) - 2$ for every individual $s = 1, 2, \dots, 30$.

Thus, to calculate $l_{2,i}$ for all $i = 1, 2, \dots, K_2$ we evaluate the individual-level likelihood functions with the sample in the original parameterisation $\tilde{\theta}_{s,i}$ and the argument of the group-level density with $\tilde{\xi}_{s,i}$ for $s = 1, 2, \dots, 30$ and group-level density parameters $\tilde{\eta}_{\mu,i}$, the original parameterisation $\tilde{\eta}_{\sigma,i}$ and the prior at $\tilde{\eta}_{\mu,i}$ and $\tilde{\zeta}_{\sigma,i}$, that is,

$$\prod_{s=1}^S f(d_s | \tilde{\theta}_{s,i}) f(\tilde{\xi}_{s,i} | \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i}) \underbrace{\phi(\tilde{\mu}_{\omega,i}) \phi(\tilde{\mu}_{\alpha,i}) \phi(\tilde{\mu}_{\gamma,i}) \phi(\tilde{\zeta}_{\omega,i}) \phi(\tilde{\zeta}_{\alpha,i}) \phi(\tilde{\zeta}_{\gamma,i})}_{g(\xi_{1,i}, \dots, \xi_{S,i}, \tilde{\eta}_{\mu,i}, \tilde{\zeta}_{\sigma,i})}, \quad (12.3.14)$$

where the function g refers to the multivariate normal distribution obtained in step 4.

7. We follow steps 7 – 9, as outlined for the bridge sampler of the individual-level implementation of the EV model.

This procedure yields a logarithm of the marginal likelihood of -3801.877 with a coefficient of variation of 10.53%.

12.3.5.2 Assessing the accuracy of our implementation

To investigate the accuracy of our implementation, we compare Bayes factors obtained with bridge sampling to Bayes factors obtained from the Savage-Dickey density ratio test (Dickey and Lientz, 1970; Dickey, 1971; for a tutorial, see Wagenaarmakers et al., 2010).

Recall that a Bayes factor is a ratio of two marginal likelihood functions. For nested model comparisons with, say, a restricted model \mathcal{M}_r within the full model \mathcal{M}_f , the Savage-Dickey density ratio¹⁶ implies that the Bayes factor can be computed as a ratio of the prior divided by the posterior of the full model at the restriction $\theta = \theta_0$, that is,

$$\text{BF}_{fr}(d) = \frac{p(d | \mathcal{M}_f)}{p(d | \mathcal{M}_r)} = \frac{\pi(\theta = \theta_0 | \mathcal{M}_f)}{\pi(\theta = \theta_0 | d, \mathcal{M}_f)}. \quad (12.3.15)$$

As the full model \mathcal{M}_f we take the EV model in which all group-level parameters are free to vary. In what follows we also consider three restricted models each with one of the three group-level mean parameters, that is, μ_ω , μ_α , and μ_γ , fixed at a predefined value. These predefined values are choosing such that the Savage-Dickey density ratio for the data at hand is one. To do so, we fit the full EV model to the Busemeyer and Stout (2002) data set (i.e., step 1 of Section 12.3.5) and then apply a nonparametric logpline density estimator (Stone et al., 1997) to the posterior samples. Fig. 12.7 shows the posterior for μ_α as the full curve, while the prior is shown as the dotted curve. Furthermore, Fig. 12.7 also shows that the prior and posterior for μ_α evaluated at $\mu_{\alpha,0} = -0.604$ have the same value, i.e., the grey dot. Hence, the Bayes factor computed using the Savage-Dickey method is one when we compare the full model in which all parameters are free to vary against the model \mathcal{M}_{r2} with μ_α fixed at $\mu_{\alpha,0}$. To compute the Bayes factor between the full model and the artificially constructed model \mathcal{M}_{r2} with $\mu_{\alpha,0} = -0.604$ using bridge sampling, we have to estimate the marginal likelihood of both \mathcal{M}_f and \mathcal{M}_{r2} . The logarithm of the marginal likelihood \mathcal{M}_f was already computed in Section 12.3.5 and presented in the top row of Table 12.3. To estimate the marginal likelihood of \mathcal{M}_{r2} we first need posterior samples of this restricted model for which we use JAGS as before. This time however, we use the likelihood Eq. (12.3.9) with μ_α fixed at $\mu_{\alpha,0} = -0.604$ in the group distribution Eq. (12.3.11). As μ_α is known and fixed within \mathcal{M}_{r2} , it is not random anymore, and therefore does not need a prior. As such, for the restricted model \mathcal{M}_{r2} we use the priors

$$\pi(\eta_\mu) = \mathcal{N}(\mu_\omega; 0, 1)\mathcal{N}(\mu_\gamma; 0, 1), \quad (12.3.16)$$

$$\pi(\eta_\sigma) = U(\sigma_\omega; 0, 1.5)U(\sigma_\alpha; 0, 1.5)U(\sigma_\gamma; 0, 1.5), \quad (12.3.17)$$

¹⁶Under certain regularity conditions that are met in our example, see Marin and Robert (2010); Verdinelli and Wasserman (1995); Wetzels et al. (2010a) for more details.

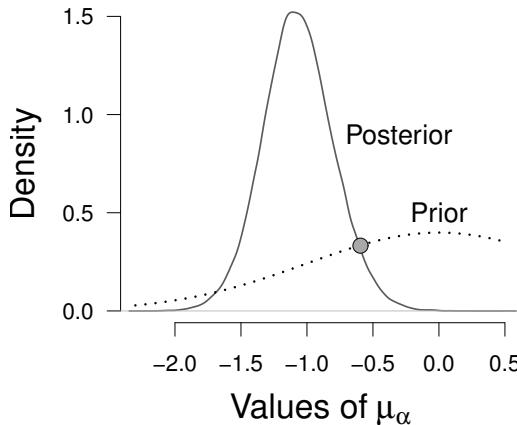


Figure 12.7: Prior and posterior distributions of the group-level mean μ_α in the Busemeyer and Stout (2002) data set. The figure shows the posterior distribution (solid line) and the prior distribution (dotted line). The grey dot indicates the intersection of the prior and the posterior distributions, for which the Savage-Dickey density ratio equals 1.

Table 12.3: Bayes factors comparing the full EV model to the restricted EV models, log marginal likelihoods, and coefficient of variation (with respect to the marginal likelihood) expressed as a percentage.

Model	$\text{BF}_{fr}(d_{\text{all}})$	log marginal likelihood	$CV[\%]$
full model	1.000	-3801.877	10.53
restricted at $\mu_\omega = -0.334$	0.729	-3801.561	14.21
restricted at $\mu_\alpha = -0.604$	0.826	-3801.686	9.99
restricted at $\mu_\gamma = 0.92$	0.710	-3801.535	13.15

instead, which is simply the prior of the full model with the prior for the fixed parameter, in this case μ_α removed. With the posterior samples for the parameters of \mathcal{M}_{r_2} at hand we proceed the estimation procedure as described in Section 12.3.5 from step 2 onwards. This lead to an estimate of the logarithm of the marginal likelihood $p(d_{\text{all}} | \mathcal{M}_{r_2})$ of -3801.686 with a coefficient of variation of 9.99%. Dividing the estimate of the marginal likelihood $p(d_{\text{all}} | \mathcal{M}_f)$ by the estimate of the marginal likelihood of $p(d_{\text{all}} | \mathcal{M}_{r_2})$ yields a Bayes factor $\text{BF}_{fr_2}(d_{\text{all}}) = 0.826$, see the third row in Table 12.3. This table also shows two other restricted models:

\mathcal{M}_{r_1} with μ_ω fixed at $\mu_{\omega,0} = -0.334$, and \mathcal{M}_{r_3} with μ_γ fixed at $\mu_{\gamma,0} = 0.92$. As before these restriction were chosen such that the Savage-Dickey density ratio equals one. The corresponding Bayes factors were derived by dividing the estimated marginal likelihood of the full model and the restricted model using the bridge sampler. It is evident that Bayes factors derived from bridge sampling closely approximate the Savage-Dickey density ratio of one. These results suggest a successful implementation of the bridge sampler. This is also reflected by the close match between the log marginal likelihoods of the four models presented in the third column of Table 12.3.

Finally, we confirm that the bridge sampler has low variance; the coefficient of variation with respect to the marginal likelihood of the full model and the three restricted models ranges between 9.99 and 14.21%.

12.4 Discussion

In this tutorial, we explained how bridge sampling can be used to estimate the marginal likelihood of popular models in mathematical psychology. As a running example, we used the beta-binomial model to illustrate step-by-step the bridge sampling estimator. To facilitate the understanding of the bridge sampler, we first discussed three of its special cases—the naive Monte Carlo estimator, the importance sampling estimator, and the generalised harmonic mean estimator. Consequently, we introduced key concepts that became gradually more complicated and sophisticated. In the second part of this tutorial, we showed how bridge sampling can be used to estimate the marginal likelihood of both an individual-level and a hierarchical implementation of the EV model (Busemeyer and Stout, 2002) for the Iowa gambling task (Bechara et al., 1994). The running example and the application of bridge sampling to the EV model demonstrated the positive aspects of the bridge sampling estimator, that is, its accuracy, reliability, practicality, and ease-of-implementation (DiCiccio et al., 1997; Frühwirth-Schnatter, 2004; Meng and Wong, 1996).

The bridge sampling estimator is superior to the naive Monte Carlo estimator, the importance sampling estimator, and the generalised harmonic mean estimator for several reasons. First, Meng and Wong (1996) showed that, among the four estimators discussed in this chapter, the bridge sampler minimises the mean-squared error because it uses the optimal bridge function. Second, in bridge sampling, choosing a suitable proposal distribution is much easier than choosing a suitable importance density for the importance sampling estimator or the generalised harmonic mean estimator because bridge sampling is more robust to the tail behaviour of the proposal distribution relative to the posterior distribution. This advantage facilitates the application of the bridge sampler to higher dimensional and complex models. This characteristic of the bridge sampler combined with the popularity of higher dimensional and complex models in mathematical psychology suggests that bridge sampling can advance model comparison exercises in many areas of mathematical psychology (e.g., reinforcement-learning models, response time models, multinomial processing tree models, etc.). Third, bridge sampling is relatively straightforward to implement. In particular, our step-by-step procedure

can be easily applied to other models with only minor changes of the code (i.e., the unnormalised posterior and potentially the proposal function have to be adapted).

Despite the numerous advantages of the bridge sampler, the take-home message of this tutorial is not that the bridge sampler should be used blindly. There exist a large variety of methods to approximate the marginal likelihood that differ in their efficiency. The most appropriate method optimises the trade-off between accuracy and implementation effort. This trade-off depends on a number of aspects such as the complexity of the model, the number of models under consideration, the statistical experience of the researcher, and the time available. This suggests that the choice of the method should be reconsidered each time a marginal likelihood needs to be obtained. Obviously, when the marginal likelihood can be determined analytically, bridge sampling is not needed at all. If the goal is to compare (at least) two nested models, the Savage-Dickey density ratio test (Dickey and Lientz, 1970; Dickey, 1971) might be a better alternative. Note however that the Savage-Dickey density ratio is not free of caveats –the full curve depicted in Fig. 12.7 is just an estimate of the posterior and also subject to estimation error.¹⁷ If only an individual-level implementation of a model is used, importance sampling may be easier to implement and may require less computational effort. If the goal is to obtain the marginal likelihood of a large number of relatively simple models, the product space or reversible jump method might be more appropriate (Carlin and Chib, 1995; Green, 1995). If a researcher with a limited programming background and/or little time resources wants to conduct a model comparison exercise, rough approximations of the Bayes factor, such as the Bayesian information criterion, might be more suitable (Schwarz, 1978). On the other hand, a researcher with an extensive background in programming and mathematical statistics might consider using path sampling—a generalisation of bridge sampling (Gelman and Meng, 1998).

To conclude, in this tutorial we showed that bridge sampling offers a reliable and easy-to-implement approach to estimate a model's marginal likelihood, see also `bridgesampling` R package of the first author. Bridge sampling can be profitably applied to a wide range of problems in mathematical psychology involving parameter estimation, model comparison, and Bayesian model averaging.

¹⁷In fact, uncertainty quantification of frequentist nonparametric methods has not yet become satisfactory.

12.A The bridge sampling estimator as a general case of methods 1 – 3

Table 12.4 shows how the naive Monte Carlo, the importance sampling, and the generalised harmonic mean estimators are special cases of the bridge sampling estimator under specific choices of the bridge function $h(\theta)$ and the proposal distribution $g(\theta)$.¹⁸

Table 12.4: Summary of the bridge sampling estimators for the marginal likelihood and its special cases: The naive Monte Carlo, importance sampling, and generalised harmonic mean estimator

Method	Estimator	Samples	Bridge function $h(\theta)$
Bridge sampling	$\frac{\frac{1}{K_2} \sum_{i=1}^{K_2} f(d \tilde{\theta}_i) \pi(\tilde{\theta}_i) h(\tilde{\theta}_i)}{\frac{1}{K_1} \sum_{j=1}^{K_1} h(\hat{\theta}_j) g(\hat{\theta}_j)}$	$\tilde{\theta}_i \sim g(\theta)$ $\hat{\theta}_j \sim \pi(\theta d)$	$h(\theta) = \frac{C}{q_1 f(d \theta) \pi(\theta) + q_2 p(d) g(\theta)}$
Naive Monte Carlo	$\frac{1}{K} \sum_{i=1}^K f(d \tilde{\theta}_i)$	$\tilde{\theta}_i \sim \pi(\theta)$	$h(\theta) = \frac{1}{g(\theta)}, g(\theta) = \pi(\theta)$
Importance sampling	$\frac{1}{K} \sum_{i=1}^K \frac{f(d \tilde{\theta}_i) \pi(\tilde{\theta}_i)}{g_{IS}(\tilde{\theta}_i)}$	$\tilde{\theta}_i \sim g_{IS}(\theta)$	$h(\theta) = \frac{1}{g_{IS}(\theta)}, g(\theta) = g_{IS}(\theta)$
Generalised harmonic mean	$\left(\frac{1}{K} \sum_{i=1}^K \frac{g_{IS}(\hat{\theta}_i)}{p(d \hat{\theta}_i) \pi(\hat{\theta}_i)} \right)^{-1}$	$\hat{\theta}_i \sim \pi(\theta d)$	$h(\theta) = \frac{1}{f(d \theta) \pi(\theta)}, g(\theta) = g_{IS}(\theta)$

In the table above $\pi(\theta)$ denotes the prior distribution, $g_{IS}(\theta)$ the importance density, $\pi(\theta | d)$ the posterior distribution, $g(\theta)$ the proposal distribution, $h(\theta)$ the bridge function, and C a constant. The last column shows the bridge function needed to obtain the special cases.

12.B Bridge sampling implementation: Avoiding numerical issues

In order to avoid numerical issues, we rewrite Eq. (12.2.31) as

¹⁸Note that bridge sampling is also a general case of the Chib and Jeliazkov (2001) method of estimating the marginal likelihood using the Metropolis-Hastings acceptance probability (Meng and Schilling, 2002; Mira and Nicholls, 2004).

$$\hat{p}_4(y)^{(t+1)} = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{l_{2,i}}{q_1 l_{2,i} + q_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 l_{1,j} + q_2 \hat{p}_4(y)^{(t)}}} \quad (12.B.1)$$

$$= \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp(\log(l_{2,i}))}{q_1 \exp(\log(l_{2,i})) + q_2 \hat{p}_4(y)^{(t)}}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 \exp(\log(l_{1,j})) + q_2 \hat{p}_4(y)^{(t)}}} \quad (12.B.2)$$

$$= \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp(\log(l_{2,i})) \exp(-l^*)}{q_1 \exp(\log(l_{2,i})) \exp(-l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{\exp(-l^*)}{q_1 \exp(\log(l_{1,j})) \exp(-l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}} \quad (12.B.3)$$

$$= \frac{1}{\exp(-l^*)} \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp(\log(l_{2,i}) - l^*)}{q_1 \exp(\log(l_{2,i}) - l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 \exp(\log(l_{1,j}) - l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}} \quad (12.B.4)$$

$$= \exp(l^*) \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp(\log(l_{2,i}) - l^*)}{q_1 \exp(\log(l_{2,i}) - l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 \exp(\log(l_{1,j}) - l^*) + q_2 \hat{p}_4(y)^{(t)} \exp(-l^*)}}. \quad (12.B.5)$$

where l^* is a constant which we can choose in a way that keeps the terms in the sums manageable. We used $l^* = \text{median}(\log(l_{1,j}))$. To further simplify matters, we defined $\hat{r}^{(t)} = \hat{p}_4(y)^{(t)} \exp(-l^*)$ and multiply the above expressions by $\exp(-l^*)$ on both sides resulting in

$$\hat{r}^{(t+1)} = \frac{\frac{1}{K_2} \sum_{i=1}^{K_2} \frac{\exp(\log(l_{2,i}) - l^*)}{q_1 \exp(\log(l_{2,i}) - l^*) + q_2 \hat{r}^{(t)}}}{\frac{1}{K_1} \sum_{j=1}^{K_1} \frac{1}{q_1 \exp(\log(l_{1,j}) - l^*) + q_2 \hat{r}^{(t)}}}. \quad (12.B.6)$$

Hence, we can run the iterative scheme with respect to \hat{r} which is more convenient because it keeps the terms in the sums manageable. We obtain an estimate of the marginal likelihood if we multiply \hat{r} by $\exp(l^*)$. Equivalently, we obtain an estimate of the logarithm of the marginal likelihood if we take the logarithm of \hat{r} and add l^* .

12.C Correcting for the probit transformation

In this section we describe how the probit transformation affects our expression of the generalised harmonic mean estimator, Eq. (12.2.22), to yield Eq. (12.2.23). Recall that we derived the generalised harmonic mean estimator using the following

equality

$$\frac{1}{p(d)} = \int \frac{g_{IS}(\theta)}{f(d|\theta)\pi(\theta)} \pi(\theta|d) d\theta. \quad (12.C.1)$$

For practical reasons, in the running example, we used a fitted normal distribution in terms of the probit-transformed parameters ξ as importance density, that is, $g_{IS}(\xi) = \frac{1}{\hat{\sigma}}\phi\left(\frac{\xi - \hat{\mu}}{\hat{\sigma}}\right)$. Note that this importance density is a function of ξ , whereas the general importance density g_{IS} in Eq. (12.C.1) is specified in terms of θ .

We now have two choices, we either (i) express the likelihood, prior and posterior in terms of ξ , or (ii) we express the importance sampler in terms of θ instead. For (i) we recall that $\theta = \Phi(\xi)$ and that the derivative of θ with respect to ξ is then $\frac{d\theta}{d\xi} = \phi(\xi)$. As such, we fill in $\theta = \Phi(\xi)$ in the likelihood, prior, posterior and $\int d\theta = \int \phi(\xi)d\xi$ resulting in

$$p(d) = \left(\mathbb{E}_{\text{post}} \begin{bmatrix} \text{importance density} \\ \frac{1}{\hat{\sigma}}\phi\left(\frac{\xi - \hat{\mu}}{\hat{\sigma}}\right) \\ \hline \underbrace{f(d|\Phi(\xi))}_{\text{likelihood}} \underbrace{\pi(\Phi(\xi))\phi(\xi)}_{\text{prior}} \end{bmatrix} \right)^{-1}, \quad (12.C.2)$$

where the expectation is with respect to the posterior in terms of ξ . Note that $\Phi(\xi) = \theta$, thus, we plugged in the untransformed θ in the likelihood, the prior and multiply it with the standard normal density ϕ to compensate for the probit-transform. In the online-provided code, we use this approach (see also Overstall and Forster, 2010). Note that in our running example with a uniform prior we have $\pi(\theta) = 1$ for every $\theta \in \Theta$ and therefore $\pi(\Phi(\xi)) = 1$ for every $\xi \in \mathbb{R}$.

Alternatively, for (ii) we recall that the integral of importance density is given by $\int g_{IS}(\xi)d\xi$ and that $\xi = \Phi^{-1}(\theta)$. The derivative of ξ with respect to θ is then $\frac{d\xi}{d\theta} = \frac{1}{\phi(\theta)}$ due to the inverse function theorem. As such, we get the equivalent expression

$$p(d) = \left(\mathbb{E}_{\text{post}} \begin{bmatrix} \text{importance density} \\ \frac{1}{\hat{\sigma}}\phi\left(\frac{\Phi^{-1}(\theta) - \hat{\mu}}{\hat{\sigma}}\right) \frac{1}{\phi(\Phi^{-1}(\theta))} \\ \hline \underbrace{f(d|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}} \end{bmatrix} \right)^{-1}, \quad (12.C.3)$$

where the expectation is with respect to the posterior in terms of θ .

These two population means can be estimated by their respective sample

means, that is,

$$\hat{p}_3(d) = \left(\frac{1}{K} \sum_{j=1}^K \frac{\frac{1}{\hat{\sigma}} \left(\frac{\hat{\xi}_j - \hat{\mu}}{\hat{\sigma}} \right)}{f(d | \Phi(\hat{\xi}_j)) \pi(\Phi(\hat{\xi}_j)) \phi(\hat{\xi}_j)} \right)^{-1}, \quad \underbrace{\hat{\xi}_j = \Phi^{-1}(\theta^*)}_{\substack{\text{probit-transformed samples} \\ \text{from the posterior distribution}}}, \quad (12.C.4)$$

$$= \left(\frac{1}{K} \sum_{j=1}^K \frac{\frac{1}{\hat{\sigma}} \phi \left(\frac{\Phi^{-1}(\hat{\theta}_j) - \hat{\mu}}{\hat{\sigma}} \right) \frac{1}{\phi(\Phi^{-1}(\hat{\theta}_j))}}{f(d | \hat{\theta}_j) \pi(\hat{\theta}_j)} \right)^{-1}, \quad \underbrace{\hat{\theta}_j \sim \pi(\theta | d)}_{\substack{\text{samples from the} \\ \text{posterior distribution}}}. \quad (12.C.5)$$

which is a result of interpreting integration as glorified summation.

12.D Details on the application of bridge sampling to the individual-level EV model

In this section, we provide more details on how we obtained the unnormalised posterior distribution for a specific participant s , for $s = 1, 2, \dots, 30$, with choices $y_s^N = (y_{s,1}, y_{s,2}, \dots, y_{s,N})$ and corresponding payoffs $x_s^N = (x_{s,1}, x_{s,2}, \dots, x_{s,N})$.

As explained in Appendix 12.B, we run the iterative scheme with respect to \hat{r} to avoid numerical issues. Consequently, we have to compute $\log(l_{1,j})$ and $\log(l_{2,i})$. We do so by transforming the priors specified with the original parameterisation to the real line using the probit transform. For the parameters $w_s \sim U[0, 1]$ and $a_s \sim U[0, 1]$ we get standard normal priors on ω_s and α_s as was elaborated on in Appendix 12.C. For the parameter c , the uniform prior $U[-2, 2]$ implies that $\int_{-2}^2 \pi(c) dc = \int_{-2}^2 0.25 dc$. To apply the change-of-variable rule we recall that $c = 4\Phi(\gamma) + 2$ and subsequently take the derivative of c with respect to γ , which results in $\frac{dc}{d\gamma} = 4\phi(\gamma)$ and, therefore, $\int dc = 4 \int \phi(\gamma) d\gamma$. Hence, the uniform prior on c in terms of γ is also just the normal density.

As such, to calculate $\log(l_{2,i})$ with $\tilde{\xi}_{s,i}$ for the i th sample from the proposal distribution we get

$$\log(l_{2,i}) = \log \left(\frac{f(d_s | \tilde{\theta}_{s,i}) \pi(\tilde{\theta}_{s,i}) \phi(\tilde{\xi}_{s,i})}{g(\tilde{\xi}_{s,i})} \right), \quad (12.D.1)$$

where $\tilde{\theta}_{s,i}$ refers to the sampled $\tilde{\xi}_{s,i}$ transformed to the original parameterisation. Taking the logarithm simplifies matters as multiplication then becomes a summation. That is,

$$\log(l_{2,i}) = \overbrace{\sum_{n=1}^N \log Pr(y_{s,n} | x^{n-1}, \tilde{\theta}_{s,i}) + \log \phi(\tilde{\omega}_{s,i}) + \log \phi(\tilde{\alpha}_{s,i}) + \log \phi(\tilde{\gamma}_{s,i})}^{\log f(d_s | \tilde{\theta}_{s,i})} - \log g(\tilde{\xi}_{s,i}), \quad (12.D.2)$$

as a result of taking independent uniform priors on the parameters and because $\log 1 = 0$.

12.E Details on the application of bridge sampling to the hierarchical EV model

Analogous to the last section, we explain here how we obtained the logarithm of the unnormalised posterior for the hierarchical implementation of the EV model. As in Appendix 12.D, we run the iterative scheme with respect to \hat{r} and, therefore, compute $\log(l_{1,j})$ and $\log(l_{2,i})$ in terms of the probit-transformed parameters. The priors on the group-level means are just standard normal, while the prior on the group-level standard deviations were given in terms of the σ s. These prior in terms of the probit-transformed τ s are also standard normal, which can be derived analogously to how we showed that the uniform prior of c on $[-2, 2]$ results in a standard normal density on γ , see Appendix 12.D. As the hierarchical model also incorporates a group-level distribution the logarithm of $l_{2,i}$ is now given by

$$\log(l_{2,i}) = \sum_{s=1}^S \left[\log f(d_s | \tilde{\theta}_{s,i}) + \log f(\tilde{\xi}_{s,i} | \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i}) \right], \quad (12.E.1)$$

$$+ \log \phi(\tilde{\mu}_{\omega,i}) + \log \phi(\tilde{\mu}_{\alpha,i}) + \log \phi(\tilde{\mu}_{\gamma,i}), \quad (12.E.2)$$

$$+ \log \phi(\tilde{\tau}_{\omega,i}) + \log \phi(\tilde{\tau}_{\alpha,i}) + \log \phi(\tilde{\tau}_{\gamma,i}) - \log g(\tilde{\xi}_{1,i}, \dots, \tilde{\xi}_{S,i}, \tilde{\eta}_i), \quad (12.E.3)$$

where $\log f(d_s | \tilde{\theta}_{s,i}) = \sum_{n=1}^N \log Pr(y_{s,n} | x^{n-1}, \tilde{\theta}_{s,i})$ and the logarithm of the group-level distribution is

$$\log f(\tilde{\xi}_{s,i} | \tilde{\eta}_{\mu,i}, \tilde{\eta}_{\sigma,i}) = \log \frac{1}{\tilde{\sigma}_{\omega,i}} \phi \left(\frac{\tilde{\omega}_{s,i} - \tilde{\mu}_i}{\tilde{\sigma}_{\omega,i}} \right) + \log \frac{1}{\tilde{\sigma}_{\alpha,i}} \phi \left(\frac{\tilde{\alpha}_{s,i} - \tilde{\mu}_i}{\tilde{\sigma}_{\alpha,i}} \right) \quad (12.E.4)$$

$$+ \log \frac{1}{\tilde{\sigma}_{\gamma,i}} \phi \left(\frac{\tilde{\gamma}_{s,i} - \tilde{\mu}_i}{\tilde{\sigma}_{\gamma,i}} \right) \quad (12.E.5)$$

Note that the hierarchical implementation implies that each draw from the proposal g , say, the i th, consists of six group-level draws $\tilde{\eta}_{\mu,i}$ and $\tilde{\eta}_{\sigma,i}$, and $S = 30$ individual-level $\tilde{\xi}_{s,i}$, each consisting of three parameters $\tilde{\xi}_{s,i} = (\tilde{\omega}_{s,i}, \tilde{\alpha}_{s,i}, \tilde{\gamma}_{s,i})$. As such, each draw of the proposal is a vector of length 96. To evaluate $\log l_{2,i}$ we transform these samples to the parameters in which the individual-level likelihood and the group-level distribution are specified.

Chapter 13

A Tutorial on Fisher Information

Abstract

In many statistical applications that concern mathematical psychologists, the concept of Fisher information plays an important role. In this tutorial we clarify the concept of Fisher information as it manifests itself across three different statistical paradigms. Firstly, in the frequentist paradigm, Fisher information is used to construct hypothesis tests and confidence intervals using maximum likelihood estimators; secondly, in the Bayesian paradigm, Fisher information is used to define a default prior; lastly, in the minimum description length paradigm, Fisher information is used to measure model complexity.

Keywords: Confidence intervals, hypothesis testing, Jeffreys's prior, minimum description length, model complexity, model selection, statistical modelling.

13.1 Introduction

Mathematical psychologists develop and apply quantitative models in order to describe human behaviour and understand latent psychological processes. Examples of such models include Stevens' law of psychophysics that describes the relation between the objective physical intensity of a stimulus and its subjectively experienced intensity (Stevens, 1957); Ratcliff's diffusion model of decision making that measures the various processes that drive behaviour in speeded response time tasks (Ratcliff, 1978); and multinomial processing tree models that decompose performance in memory tasks into the contribution of separate latent mechanisms (Batchelder and Riefer, 1980; Chechile, 1973).

This chapter is published as Ly, A., Marsman, M., Verhagen, A.J., Grasman, R.P.P.P., and Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80, 40–55. doi: <https://doi.org/10.1016/j.jmp.2017.05.006>. Also available as arXiv preprint, arXiv:1705.01064.

When applying their models to data, mathematical psychologists may operate from within different statistical paradigms and focus on different substantive questions. For instance, working within the classical or frequentist paradigm a researcher may wish to test certain hypotheses or decide upon the number of trials to be presented to participants in order to estimate their latent abilities. Working within the Bayesian paradigm a researcher may wish to know how to determine a suitable default prior on the parameters of a model. Working within the minimum description length (MDL) paradigm a researcher may wish to compare rival models and quantify their complexity. Despite the diversity of these paradigms and purposes, they are connected through the concept of Fisher information.

Fisher information plays a pivotal role throughout statistical modelling, but an accessible introduction for mathematical psychologists is lacking. The goal of this tutorial is to fill this gap and illustrate the use of Fisher information in the three statistical paradigms mentioned above: frequentist, Bayesian, and MDL. This work builds directly upon the *Journal of Mathematical Psychology* tutorial article by Myung (2003) on maximum likelihood estimation. The intended target group for this tutorial are graduate students and researchers with an affinity for cognitive modelling and mathematical statistics.

To keep this tutorial self-contained we start by describing our notation and key concepts. We then provide the definition of Fisher information and show how it can be calculated. The ensuing sections exemplify the use of Fisher information for different purposes. Section 13.2 shows how Fisher information can be used in frequentist statistics to construct confidence intervals and hypothesis tests from maximum likelihood estimators (MLEs). Section 13.3 shows how Fisher information can be used in Bayesian statistics to define a default prior on model parameters. In Section 13.4 we clarify how Fisher information can be used to measure model complexity within the MDL framework of inference.

13.1.1 Notation and key concepts

Before defining Fisher information it is necessary to discuss a series of fundamental concepts such as the nature of statistical models, probability mass functions, and statistical independence. Readers familiar with these concepts may safely skip to the next section.

A *statistical model* is typically defined through a function $f(x_i | \theta)$ that represents how a parameter θ is functionally related to potential outcomes x_i of a random variable X_i . For ease of exposition, we take θ to be one-dimensional throughout this text. The generalisation to vector-valued θ can be found in Appendix 13.A, see also Myung and Navarro (2005).

As a concrete example, θ may represent a participant's intelligence, X_i a participant's (future) performance on the i th item of an IQ test, $x_i = 1$ the potential outcome of a correct response, and $x_i = 0$ the potential outcome of an incorrect response on the i th item. Similarly, X_i is the i th trial in a coin flip experiment with two potential outcomes: heads, $x_i = 1$, or tails, $x_i = 0$. Thus, we have the binary outcome space $\mathcal{X} = \{0, 1\}$. The coin flip model is also known as the Bernoulli distribution $f(x_i | \theta)$ that relates the coin's propensity $\theta \in (0, 1)$ to land

heads to the potential outcomes as

$$f(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1-x_i}, \text{ where } x_i \in \mathcal{X} = \{0, 1\}. \quad (13.1.1)$$

Formally, if θ is known, fixing it in the functional relationship f yields a function $p_\theta(x_i) = f(x_i | \theta)$ of the potential outcomes x_i . This $p_\theta(x_i)$ is referred to as a *probability density function* (pdf) when X_i has outcomes in a continuous interval, whereas it is known as a *probability mass function* (pmf) when X_i has discrete outcomes. The pmf $p_\theta(x_i) = P(X_i = x_i | \theta)$ can be thought of as a data generative device as it specifies how θ defines the chance with which X_i takes on a potential outcome x_i . As this holds for any outcome x_i of X_i , we say that X_i is distributed according to $p_\theta(x_i)$. For brevity, we do not further distinguish the continuous from the discrete case, and refer to $p_\theta(x_i)$ simply as a pmf.

For example, when the coin's true propensity is $\theta^* = 0.3$, replacing θ by θ^* in the Bernoulli distribution yields the pmf $p_{0.3}(x_i) = 0.3^{x_i} 0.7^{1-x_i}$, a function of all possible outcomes of X_i . A subsequent replacement $x_i = 0$ in the pmf $p_{0.3}(0) = 0.7$ tells us that this coin generates the outcome 0 with 70% chance.

In general, experiments consist of n trials yielding a potential set of outcomes $x^n = (x_1, \dots, x_n)$ of the random vector $X^n = (X_1, \dots, X_n)$. These n random variables are typically assumed to be *independent and identically distributed* (iid). Identically distributed implies that each of these n random variables is governed by one and the same θ , while independence implies that the joint distribution of all these n random variables simultaneously is given by a product, that is,

$$f(x^n | \theta) = f(x_1 | \theta) \times \dots \times f(x_n | \theta) = \prod_{i=1}^n f(x_i | \theta). \quad (13.1.2)$$

As before, when θ is known, fixing it in this relationship $f(x^n | \theta)$ yields the (joint) pmf of X^n as $p_\theta(x^n) = p_\theta(x_1) \times \dots \times p_\theta(x_n) = \prod_{i=1}^n p_\theta(x_i)$.

In psychology the iid assumption is typically evoked when experimental data are analysed in which participants have been confronted with a sequence of n items of roughly equal difficulty. When the participant can be either correct or incorrect on each trial, the participant's performance X^n can then be related to an n -trial coin flip experiment governed by one single θ over all n trials. The random vector X^n has 2^n potential outcomes x^n . For instance, when $n = 10$, we have $2^n = 1,024$ possible outcomes and we write \mathcal{X}^n for the collection of all these potential outcomes. The chance of observing a potential outcome x^n is determined by the coin's propensity θ as

$$f(x^n | \theta) = f(x_1 | \theta) \times \dots \times f(x_n | \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}, \quad (13.1.3)$$

where $x^n \in \mathcal{X}^n$. When the coin's true propensity θ is $\theta^* = 0.6$, replacing θ by θ^* in Eq. (13.1.3) yields the joint pmf $p_{0.6}(x^n) = f(x^n | \theta = 0.6) = 0.6^{\sum_{i=1}^n x_i} 0.4^{n - \sum_{i=1}^n x_i}$. The pmf with a particular outcome entered, say, $x^n = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$ reveals that the coin with $\theta^* = 0.6$ generates this particular outcome with 0.18% chance.

13.1.2 Definition of Fisher information

In practice, the true value of θ is not known and has to be inferred from the observed data. The first step typically entails the creation of a data summary. For example, suppose once more that X^n refers to an n -trial coin flip experiment and suppose that we observed $x_{\text{obs}}^n = (1, 0, 0, 1, 1, 1, 0, 1, 1)$. To simplify matters, we only record the number of heads as $Y = \sum_{i=1}^n X_i$, which is a function of the data. Applying our function to the specific observations yields the realisation $y_{\text{obs}} = Y(x_{\text{obs}}^n) = 7$. Since the coin flips X^n are governed by θ , so is a function of X^n ; indeed, θ relates to the potential outcomes y of Y as follows

$$f(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \text{ where } y \in \mathcal{Y} = \{0, 1, \dots, n\}, \quad (13.1.4)$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ enumerates the possible sequences of length n that consist of y heads and $n - y$ tails. For instance, when flipping a coin $n = 10$ times, there are 120 possible sequences of zeroes and ones that contain $y = 7$ heads and $n - y = 3$ tails. The distribution $f(y | \theta)$ is known as the binomial distribution.

The summary statistic Y has $n+1$ possible outcomes, whereas X^n has 2^n . For instance, when $n = 10$ the statistic Y has only 11 possible outcomes, whereas X^n has 1,024. This reduction results from the fact that the statistic Y ignores the order with which the data are collected. Observe that the conditional probability of the raw data given $Y = y$ is equal to $P(X^n | Y = y, \theta) = 1/\binom{n}{y}$ and that it does not depend on θ . This means that after we observe $Y = y$ the conditional probability of X^n is independent of θ , even though each of the distributions of X^n and Y separately do depend on θ . We, therefore, conclude that there is no information about θ left in X^n after observing $Y = y$ (Fisher, 1920; Stigler, 1973).

More generally, we call a function of the data, say, $T = t(X^n)$ a *statistic*. A statistic is referred to as *sufficient* for the parameter θ , if the expression $P(X^n | T = t, \theta)$ does not depend on θ itself. To quantify the amount of information about the parameter θ in a sufficient statistic T and the raw data, Fisher introduced the following measure.

Definition 13.1.1 (Fisher information). The *Fisher information* $I_X(\theta)$ of a random variable X about θ is defined as¹

$$I_X(\theta) = \begin{cases} \sum_{x \in \mathcal{X}} \left(\frac{d}{d\theta} \log f(x | \theta) \right)^2 p_\theta(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathcal{X}} \left(\frac{d}{d\theta} \log f(x | \theta) \right)^2 p_\theta(x) dx & \text{if } X \text{ is continuous.} \end{cases} \quad (13.1.6)$$

The derivative $\frac{d}{d\theta} \log f(x | \theta)$ is known as the *score function*, a function of x , and describes how sensitive the model (i.e., the functional form f) is to changes in

¹Under mild regularity conditions Fisher information is equivalently defined as

$$I_X(\theta) = -E \left(\frac{d^2}{d\theta^2} \log f(X | \theta) \right) = \begin{cases} -\sum_{x \in \mathcal{X}} \left(\frac{d^2}{d\theta^2} \log f(x | \theta) \right) p_\theta(x) & \text{if } X \text{ is discrete,} \\ -\int_{\mathcal{X}} \left(\frac{d^2}{d\theta^2} \log f(x | \theta) \right) p_\theta(x) dx & \text{if } X \text{ is continuous.} \end{cases} \quad (13.1.5)$$

where $\frac{d^2}{d\theta^2} \log f(x | \theta)$ denotes the second derivative of the logarithm of f with respect to θ .

θ at a particular θ . The Fisher information measures the overall sensitivity of the functional relationship f to changes of θ by weighting the sensitivity at each potential outcome x with respect to the chance defined by $p_\theta(x) = f(x|\theta)$. The weighting with respect to $p_\theta(x)$ implies that the Fisher information about θ is an expectation.

Similarly, Fisher information $I_{X^n}(\theta)$ within the random vector X^n about θ is calculated by replacing $f(x|\theta)$ with $f(x^n|\theta)$, thus, $p_\theta(x)$ with $p_\theta(x^n)$ in the definition. Moreover, under the assumption that the random vector X^n consists of n iid trials of X it can be shown that $I_{X^n}(\theta) = nI_X(\theta)$, which is why $I_X(\theta)$ is also known as the unit Fisher information.² Intuitively, an experiment consisting of $n = 10$ trials is expected to be twice as informative about θ compared to an experiment consisting of only $n = 5$ trials. \diamond

Intuitively, we cannot expect an arbitrary summary statistic T to extract more information about θ than what is already provided by the raw data. Fisher information adheres to this rule, as it can be shown that

$$I_{X^n}(\theta) \geq I_T(\theta), \quad (13.1.7)$$

with equality if and only if T is a sufficient statistic for θ .

Example 13.1.1 (The information about θ within the raw data and a summary statistic). A direct calculation with a Bernoulli distributed random vector X^n shows that the Fisher information about θ within an n -trial coin flip experiment is given by

$$I_{X^n}(\theta) = nI_X(\theta) = n \frac{1}{\theta(1-\theta)}, \quad (13.1.8)$$

where $I_X(\theta) = \frac{1}{\theta(1-\theta)}$ is the Fisher information of θ within a single trial. As shown in Fig. 13.1, the unit Fisher information $I_X(\theta)$ depends on θ . Similarly, we can calculate the Fisher information about θ within the summary statistic Y by using the binomial model instead. This yields $I_Y(\theta) = \frac{n}{\theta(1-\theta)}$. Hence, $I_{X^n}(\theta) = I_Y(\theta)$ for any value of θ . In other words, the expected information in Y about θ is the same as the expected information about θ in X^n , regardless of the value of θ . \diamond

Observe that the information in the raw data X^n and the statistic Y are equal for every θ , and specifically also for its unknown true value θ^* . That is, there is no statistical information about θ lost when we use a sufficient statistic Y instead of the raw data X^n . This is particularly useful when the data set X^n is large and can be replaced by a single number Y .

13.2 The role of Fisher information in frequentist statistics

Recall that θ is unknown in practice and to infer its value we might: (1) provide a best guess in terms of a point estimate; (2) postulate its value and test whether this

²Note the abuse of notation – we dropped the subscript i for the i th random variable X_i and denote it simply by X instead.

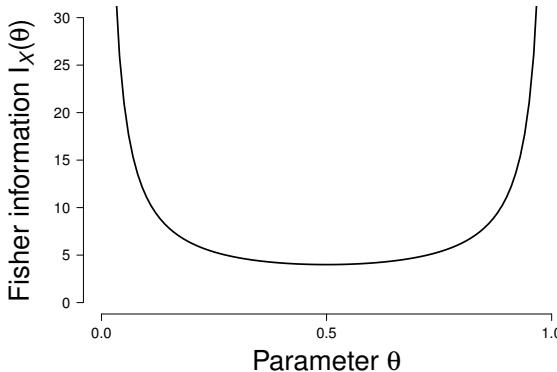


Figure 13.1: The unit Fisher information $I_X(\theta) = \frac{1}{\theta(1-\theta)}$ as a function of θ within the Bernoulli model. As θ reaches zero or one the expected information goes to infinity.

value aligns with the data, or (3) derive a confidence interval. In the frequentist framework, each of these inferential tools is related to the Fisher information and exploits the data generative interpretation of a pmf. Recall that given a model $f(x^n | \theta)$ and a known θ , we can view the resulting pmf $p_\theta(x^n)$ as a recipe that reveals how θ defines the chances with which X^n takes on the potential outcomes x^n .

This data generative view is central to Fisher's conceptualisation of the *maximum likelihood estimator* (MLE; Fisher, 1912; Fisher, 1922; Fisher, 1925; LeCam, 1990; Myung, 2003). For instance, the binomial model implies that a coin with a hypothetical propensity $\theta = 0.5$ will generate the outcome $y = 7$ heads out of $n = 10$ trials with 11.7% chance, whereas a hypothetical propensity of $\theta = 0.7$ will generate the same outcome $y = 7$ with 26.7% chance. Fisher concluded that an actual observation $y_{\text{obs}} = 7$ out of $n = 10$ is therefore more likely to be generated from a coin with a hypothetical propensity of $\theta = 0.7$ than from a coin with a hypothetical propensity of $\theta = 0.5$. Fig. 13.2 shows that for this specific observation $y_{\text{obs}} = 7$, the hypothetical value $\theta = 0.7$ is the maximum likelihood *estimate*; the number $\hat{\theta}_{\text{obs}} = 0.7$. This estimate is a realisation of the maximum likelihood *estimator* (MLE); in this case, the MLE is the function $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} Y$, i.e., the sample mean. Note that the MLE is a statistic, that is, a function of the data.

13.2.1 Using Fisher information to design an experiment

Since X^n depends on θ so will a function of X^n , in particular, the MLE $\hat{\theta}$. The distribution of the potential outcomes of the MLE $\hat{\theta}$ is known as the *sampling distribution* of the estimator and denoted as $f(\hat{\theta}_{\text{obs}} | \theta)$. As before, when θ^* is assumed to be known, fixing it in $f(\hat{\theta}_{\text{obs}} | \theta)$ yields the pmf $p_{\theta^*}(\hat{\theta}_{\text{obs}})$, a function

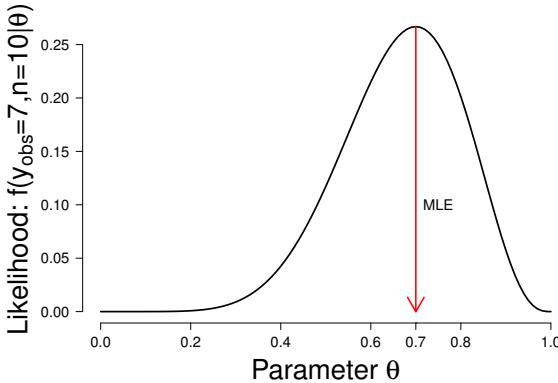


Figure 13.2: The likelihood function based on observing $y_{\text{obs}} = 7$ heads in $n = 10$ trials. For these data, the MLE is equal to $\hat{\theta}_{\text{obs}} = 0.7$, see the main text for the interpretation of this function.

of the potential outcomes of $\hat{\theta}$. This function f between the parameter θ and the potential outcomes of the MLE $\hat{\theta}$ is typically hard to describe, but for n large enough it can be characterised by the Fisher information.

For iid data and under general conditions,³ the difference between the true θ^* and the MLE converges in distribution to a normal distribution, that is,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_X^{-1}(\theta^*)), \text{ as } n \rightarrow \infty. \quad (13.2.1)$$

Hence, for large enough n , the “error” is approximately normally distributed⁴

$$(\hat{\theta} - \theta^*) \approx \mathcal{N}\left(0, 1/(nI_X(\theta^*))\right). \quad (13.2.2)$$

This means that the MLE $\hat{\theta}$ generates potential estimates $\hat{\theta}_{\text{obs}}$ around the true value θ^* with a standard error given by the inverse of the square root of the Fisher information at the true value θ^* , i.e., $1/\sqrt{nI_X(\theta^*)}$, whenever n is large enough. Note that the chances with which the estimates of $\hat{\theta}$ are generated depend on the true value θ^* and the sample size n . Observe that the standard error decreases

³Basically, when the Fisher information exists for all parameter values. For details see the advanced accounts provided by Bickel et al. (1993), Hájek (1970), Inagaki (1970), LeCam (1970) and Appendix 13.E.

⁴Note that $\hat{\theta}$ is random, while the true value θ^* is fixed. As such, the error $\hat{\theta} - \theta^*$ and the rescaled error $\sqrt{n}(\hat{\theta} - \theta^*)$ are also random. We used \xrightarrow{D} in Eq. (13.2.1) to convey that the distribution of the left-hand side goes to the distribution on the right-hand side. Similarly, \approx in Eq. (13.2.2) implies that the distribution of the left-hand side is *approximately* equal to the distribution given on the right-hand side. Hence, for finite n there will be an error due to using the normal distribution as an approximation to the true sampling distribution. This approximation error is ignored in the constructions given below, see Appendix 13.B.1 for a more thorough discussion.

when the unit information $I_X(\theta^*)$ is high or when n is large. As experimenters we do not have control over the true value θ^* , but we can affect the data generating process by choosing the number of trials n . Larger values of n increase the amount of information in X^n , heightening the chances of the MLE producing an estimate $\hat{\theta}_{\text{obs}}$ that is close to the true value θ^* . The following example shows how this can be made precise.

Example 13.2.1 (Designing a binomial experiment with the Fisher information). *Recall that the potential outcomes of a normal distribution fall within one standard error of the population mean with 68% chance. Hence, when we choose n such that $1/\sqrt{nI_X(\theta^*)} = 0.1$ we design an experiment that allows the MLE to generate estimates within 0.1 distance of the true value with 68% chance. To overcome the problem that θ^* is not known, we solve the problem for the worst case scenario. For the Bernoulli model this is given by $\theta = 1/2$, the least informative case, see Fig. 13.1. As such, we have $1/\sqrt{nI_X(\theta^*)} \leq 1/\sqrt{nI_X(1/2)} = 1/(2\sqrt{n}) = 0.1$, where the last equality is the target requirement and is solved by $n = 25$.*

This leads to the following interpretation. After simulating $k = 100$ data sets $x_{\text{obs},1}^n, \dots, x_{\text{obs},k}^n$ each with $n = 25$ trials, we can apply to each of these data sets the MLE yielding k estimates $\hat{\theta}_{\text{obs},1}, \dots, \hat{\theta}_{\text{obs},k}$. The sampling distribution implies that at least 68 of these $k = 100$ estimate are expected to be at most 0.1 distance away from the true θ^* . \diamond

13.2.2 Using Fisher information to construct a null hypothesis test

The (asymptotic) normal approximation to the sampling distribution of the MLE can also be used to construct a null hypothesis test. When we postulate that the true value equals some hypothesised value of interest, say, $\theta^* = \theta_0$, a simple plugin then allows us to construct a prediction interval based on our knowledge of the normal distribution. More precisely, the potential outcomes x^n with n large enough and generated according to $p_{\theta^*}(x^n)$ leads to potential estimates $\hat{\theta}_{\text{obs}}$ that fall within the range

$$\left(\theta^* - 1.96\sqrt{\frac{1}{n}I_X^{-1}(\theta^*)}, \theta^* + 1.96\sqrt{\frac{1}{n}I_X^{-1}(\theta^*)} \right), \quad (13.2.3)$$

with (approximately) 95% chance. This 95%-prediction interval Eq. (13.2.3) allows us to construct a point null hypothesis test based on a pre-experimental postulate $\theta^* = \theta_0$.

Example 13.2.2 (A null hypothesis test for a binomial experiment). *Under the null hypothesis $\mathcal{H}_0 : \theta^* = \theta_0 = 0.5$, we predict that an outcome of the MLE based on $n = 10$ trials will lie between $(0.19, 0.81)$ with 95% chance. This interval follows from replacing θ^* by θ_0 in the 95%-prediction interval Eq. (13.2.3). The data generative view implies that if we simulate $k = 100$ data sets each with the same $\theta^* = 0.5$ and $n = 10$, we would then have k estimates $\hat{\theta}_{\text{obs},1}, \dots, \hat{\theta}_{\text{obs},k}$ of which five are expected to be outside this 95% interval $(0.19, 0.81)$. Fisher, therefore,*

classified an outcome of the MLE that is smaller than 0.19 or larger than 0.81 as extreme under the null and would then reject the postulate $\mathcal{H}_0 : \theta_0 = 0.5$ at a significance level of .05. \diamond

The normal approximation to the sampling distribution of the MLE and the resulting null hypothesis test is particularly useful when the exact sampling distribution of the MLE is unavailable or hard to compute.

Example 13.2.3 (An MLE null hypothesis test for the Laplace model). *Suppose that we have n iid samples from the Laplace distribution*

$$f(x_i | \theta) = \frac{1}{2b} \exp\left(-\frac{|x_i - \theta|}{b}\right), \quad (13.2.4)$$

where θ denotes the population mean and the population variance is given by $2b^2$. It can be shown that the MLE for this model is the sample median, $\hat{\theta} = \hat{M}$, and the unit Fisher information is $I_X(\theta) = b^{-2}$. The exact sampling distribution of the MLE is unwieldy (Kotz et al., 2001) and not presented here. Asymptotic normality of the MLE is practical, as it allows us to discard the unwieldy exact sampling distribution and, instead, base our inference on a more tractable (approximate) normal distribution with a mean equal to the true value θ^* and a variance equal to b^2/n . For $n = 100$, $b = 1$ and repeated sampling under the hypothesis $\mathcal{H}_0 : \theta^* = \theta_0$, approximately 95% of the estimates (the observed sample medians) are expected to fall in the range $(\theta_0 - 0.196, \theta_0 + 0.196)$. \diamond

13.2.3 Using Fisher information to compute confidence intervals

An alternative to both point estimation and null hypothesis testing is interval estimation. In particular, a 95%-confidence interval can be obtained by replacing in the prediction interval Eq. (13.2.3) the unknown true value θ^* by an estimate $\hat{\theta}_{\text{obs}}$. Recall that a simulation with $k = 100$ data sets each with n trials leads to $\hat{\theta}_{\text{obs},1}, \dots, \hat{\theta}_{\text{obs},k}$ estimates, and each estimate leads to a different 95%-confidence interval. It is then expected that 95 of these $k = 100$ intervals encapsulate the true value θ^* .⁵ Note that these intervals are centred around different points whenever the estimates differ and that their lengths differ, as the Fisher information depends on θ .

Example 13.2.4 (An MLE confidence interval for the Bernoulli model). *When we observe $y_{\text{obs},1} = 7$ heads in $n = 10$ trials, the MLE then produces the estimate $\hat{\theta}_{\text{obs},1} = 0.7$. Replacing θ^* in the prediction interval Eq. (13.2.3) with $\theta^* = \hat{\theta}_{\text{obs},1}$ yields an approximate 95%-confidence interval $(0.42, 0.98)$ of length 0.57. On the other hand, had we instead observed $y_{\text{obs},2} = 6$ heads, the MLE would then yield $\hat{\theta}_{\text{obs},2} = 0.6$ resulting in the interval $(0.29, 0.90)$ of length 0.61.* \diamond

In sum, Fisher information can be used to approximate the sampling distribution of the MLE when n is large enough. Knowledge of the Fisher information

⁵But see Brown et al. (2001).

can be used to choose n such that the MLE produces an estimate close to the true value, construct a null hypothesis test, and compute confidence intervals.

13.3 The role of Fisher information in Bayesian statistics

This section outlines how Fisher information can be used to define the Jeffreys's prior, a default prior commonly used for estimation problems and for nuisance parameters in a Bayesian hypothesis test (e.g., Bayarri et al., 2012; Dawid, 2011; Gronau et al., 2017a; Jeffreys, 1961; Liang et al., 2008; Li and Clyde, 2015; Ly et al., 2016a, 2016b; Ly et al., 2017d; Ly et al., 2017e; Robert, 2016). To illustrate the desirability of the Jeffreys's prior we first show how the naive use of a uniform prior may have undesirable consequences, as the uniform prior depends on the representation of the inference problem, that is, on how the model is parameterised. This dependence is commonly referred to as lack of invariance: different parameterisations of the same model result in different posteriors and, hence, different conclusions. We visualise the representation problem using simple geometry and show how the geometrical interpretation of Fisher information leads to the Jeffreys's prior that is parameterisation-invariant.

13.3.1 Bayesian updating

Bayesian analysis centres on the observations x_{obs}^n for which a generative model f is proposed that functionally relates the observed data to an unobserved parameter θ . Given the observations x_{obs}^n , the functional relationship f is inverted using Bayes' rule to infer the relative plausibility of the values of θ . This is done by replacing the potential outcome part x^n in f by the actual observations yielding a *likelihood function* $f(x_{\text{obs}}^n | \theta)$, which is a function of θ . In other words, x_{obs}^n is known, thus, fixed, and the true θ is unknown, therefore, free to vary. The candidate set of possible values for the true θ is denoted by Θ and referred to as the parameter space. Our knowledge about θ is formalised by a distribution $g(\theta)$ over the parameter space Θ . This distribution is known as the prior on θ , as it is set before any datum is observed. We can use Bayes' theorem to calculate the posterior distribution over the parameter space Θ given the data that were actually observed as follows

$$g(\theta | X^n = x_{\text{obs}}^n) = \frac{f(x_{\text{obs}}^n | \theta)g(\theta)}{\int_{\Theta} f(x_{\text{obs}}^n | \theta)g(\theta) d\theta}. \quad (13.3.1)$$

This expression is often verbalised as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \quad (13.3.2)$$

The posterior distribution is a combination of what we knew before we saw the data (i.e., the information in the prior), and what we have learned from the observations in terms of the likelihood (e.g., Lee and Wagenmakers, 2013). Note that the integral is now over θ and not over the potential outcomes.

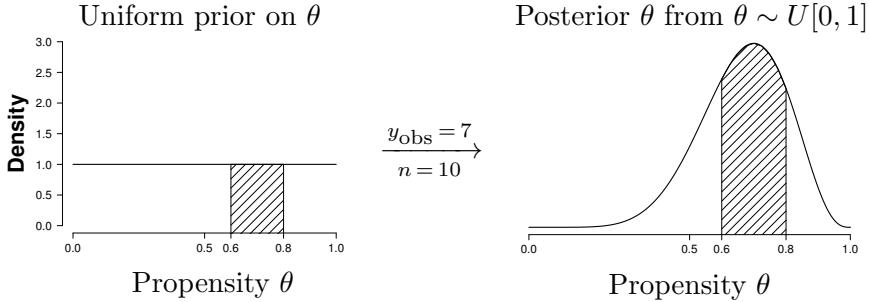


Figure 13.3: Bayesian updating based on observations x_{obs}^n with $y_{\text{obs}} = 7$ heads out of $n = 10$ tosses. In the left panel, the uniform prior distribution assigns equal probability to every possible value of the coin’s propensity θ . In the right panel, the posterior distribution is a compromise between the prior and the observed data.

13.3.2 Failure of the uniform distribution on the parameter as a non-informative prior

When little is known about the parameter θ that governs the outcomes of X^n , it may seem reasonable to express this ignorance with a uniform prior distribution $g(\theta)$, as no parameter value of θ is then favoured over another. This leads to the following type of inference:

Example 13.3.1 (Uniform prior on θ). *Before data collection, θ is assigned a uniform prior, that is, $g(\theta) = 1/V_\Theta$ with a normalisation constant of $V_\Theta = 1$ as shown in the left panel of Fig. 13.3. Suppose that we observe coin flip data x_{obs}^n with $y_{\text{obs}} = 7$ heads out of $n = 10$ trials. To relate these observations to the coin’s propensity θ we use the Bernoulli distribution as our $f(x^n | \theta)$. A replacement of x^n by the data actually observed yields the likelihood function $f(x_{\text{obs}}^n | \theta) = \theta^7(1-\theta)^3$, which is a function of θ . Bayes’ theorem now allows us to update our prior to the posterior that is plotted in the right panel of Fig. 13.3.* ◇

Note that a uniform prior on θ has the length, more generally, volume, of the parameter space as the normalisation constant; in this case, $V_\Theta = 1$, which equals the length of the interval $\Theta = (0, 1)$. Furthermore, a uniform prior can be characterised as the prior that gives equal probability to all sub-intervals of equal length. Thus, the probability of finding the true value θ^* within a sub-interval $J_\theta = (\theta_a, \theta_b) \subset \Theta = (0, 1)$ is given by the relative length of J_θ with respect to the length of the parameter space, that is,

$$P(\theta^* \in J_\theta) = \int_{J_\theta} g(\theta) d\theta = \frac{1}{V_\Theta} \int_{\theta_a}^{\theta_b} 1 d\theta = \frac{\theta_b - \theta_a}{V_\Theta}. \quad (13.3.3)$$

Hence, before any datum is observed, the uniform prior expresses the belief $P(\theta^* \in J_\theta) = 0.20$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x_{obs}^n with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $P(\theta^* \in J_\theta | x_{\text{obs}}^n) = 0.54$, see the shaded areas in Fig. 13.3.

Although intuitively appealing, it can be unwise to choose the uniform distribution by default, as the results are highly dependent on how the model is parameterised. In what follows, we show how a different parameterisation leads to different posteriors and, consequently, different conclusions.

Example 13.3.2 (Different representations, different conclusions). *The propensity of a coin landing heads up is related to the angle ϕ with which that coin is bent. Suppose that the relation between the angle ϕ and the propensity θ is given by the function $\theta = h(\phi) = \frac{1}{2} + \frac{1}{2}(\frac{\phi}{\pi})^3$, chosen here for mathematical convenience.⁶ When ϕ is positive the tail side of the coin is bent inwards, which increases the coin's chances to land heads. As the function $\theta = h(\phi)$ also admits an inverse function $h^{-1}(\theta) = \phi$, we have an equivalent formulation of the problem in Example 13.3.1, but now described in terms of the angle ϕ instead of the propensity θ .*

As before, in order to obtain a posterior distribution, Bayes' theorem requires that we specify a prior distribution. As the problem is formulated in terms of ϕ , one may believe that a non-informative choice is to assign a uniform prior $\tilde{g}(\phi)$ on ϕ , as this means that no value of ϕ is favoured over another. A uniform prior on ϕ is in this case given by $\tilde{g}(\phi) = 1/V_\Phi$ with a normalisation constant $V_\Phi = 2\pi$, because the parameter ϕ takes on values in the interval $\Phi = (-\pi, \pi)$. This uniform distribution expresses the belief that the true ϕ^* can be found in any of the intervals $(-1.0\pi, -0.8\pi), (-0.8\pi, -0.6\pi), \dots, (0.8\pi, 1.0\pi)$ with 10% probability, because each of these intervals is 10% of the total length, see the top-left panel of Fig. 13.4. For the same data as before, the posterior calculated from Bayes' theorem is given in top-right panel of Fig. 13.4. As the problem in terms of the angle ϕ is equivalent to that of $\theta = h(\phi)$ we can use the function h to translate the posterior in terms of ϕ to a posterior on θ , see the bottom-right panel of Fig. 13.4. This posterior on θ is noticeably different from the posterior on θ shown in Figure 13.3.

Specifically, the uniform prior on ϕ corresponds to the prior belief $\tilde{P}(\theta^* \in J_\theta) = 0.13$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x_{obs}^n with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $\tilde{P}(\theta^* \in J_\theta | x_{\text{obs}}^n) = 0.29$,⁷ see the shaded areas in Fig. 13.4. Crucially, the earlier analysis that assigned a uniform prior to the propensity θ yielded a posterior probability $P(\theta^* \in J_\theta | x_{\text{obs}}^n) = 0.54$, which is markedly different from the current analysis that assigns a uniform prior to the angle ϕ .

The same posterior on θ is obtained when the prior on ϕ is first translated into a prior on θ (bottom-left panel) and then updated to a posterior with Bayes' theorem. Regardless of the stage at which the transformation is applied, the resulting posterior on θ differs substantially from the result plotted in the right panel of Fig. 13.3. ◇

Thus, the uniform prior distribution is not a panacea for the quantification of prior ignorance, as the conclusions depend on how the problem is parameterised.

⁶Another example involves the logit formulation of the Bernoulli model, that is, in terms of $\phi = \log(\frac{\theta}{1-\theta})$, where $\Phi = \mathbb{R}$. This logit formulation is the basic building block in item response theory. We did not discuss this example as the uniform prior on the logit cannot be normalised and, therefore, not easily represented in the plots.

⁷The tilde makes explicit that the prior and posterior are derived from the uniform prior $\tilde{g}(\phi)$ on ϕ .

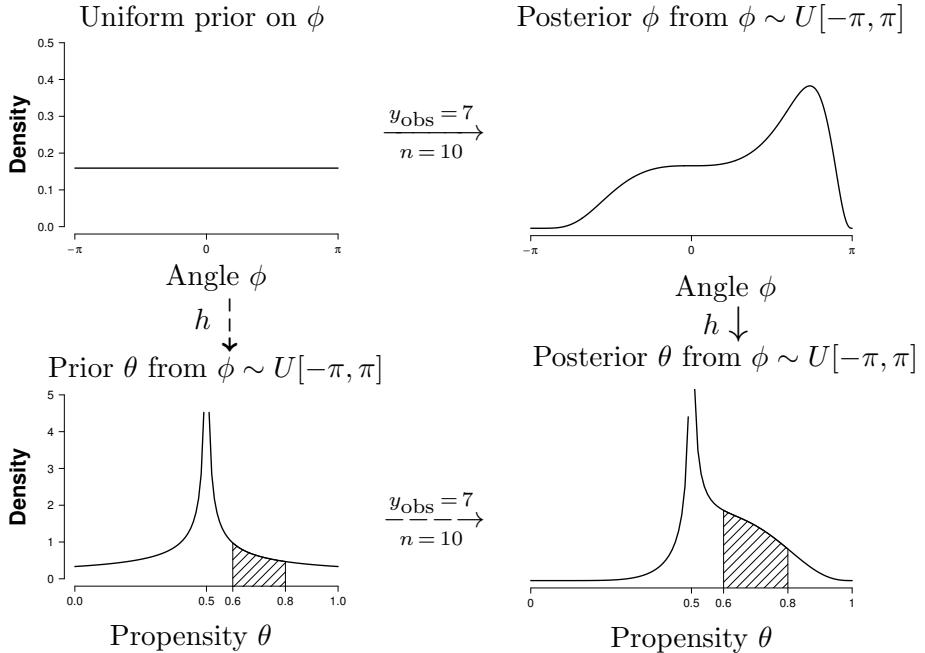


Figure 13.4: Bayesian updating based on observations x_{obs}^n with $y_{\text{obs}} = 7$ heads out of $n = 10$ tosses when a uniform prior distribution is assigned to the coin's angle ϕ . The uniform distribution is shown in the top-left panel. Bayes' theorem results in a posterior distribution for ϕ that is shown in the top-right panel. This posterior $\tilde{g}(\phi | x_{\text{obs}}^n)$ is transformed into a posterior on θ (bottom-right panel) using $\theta = h(\phi)$. The same posterior on θ is obtained if we proceed via an alternative route in which we first transform the uniform prior on ϕ to the corresponding prior on θ and then apply Bayes' theorem with the induced prior on θ . A comparison to the results from Fig. 13.3 reveals that posterior inference differs notably depending on whether a uniform distribution is assigned to the angle ϕ or to the propensity θ .

In particular, a uniform prior on the coin's angle $\tilde{g}(\phi) = 1/V_\Phi$ yields a highly informative prior in terms of the coin's propensity θ . This lack of invariance caused Karl Pearson, Ronald Fisher and Jerzy Neyman to reject 19th century Bayesian statistics that was based on the uniform prior championed by Pierre-Simon Laplace. This rejection resulted in, what is now known as, frequentist statistics, see also Hald (2008), Lehmann (2011), and Stigler (1986).

13.3.3 A default prior by Jeffreys's rule

Unlike the other fathers of modern statistical thoughts, Harold Jeffreys continued to study Bayesian statistics based on formal logic and his philosophical convictions of scientific inference (see, e.g., Aldrich, 2005; Etz and Wagenmakers, 2017; Jeffreys, 1961; Ly et al., 2016a, 2016b; Robert et al., 2009; Wrinch and Jeffreys,

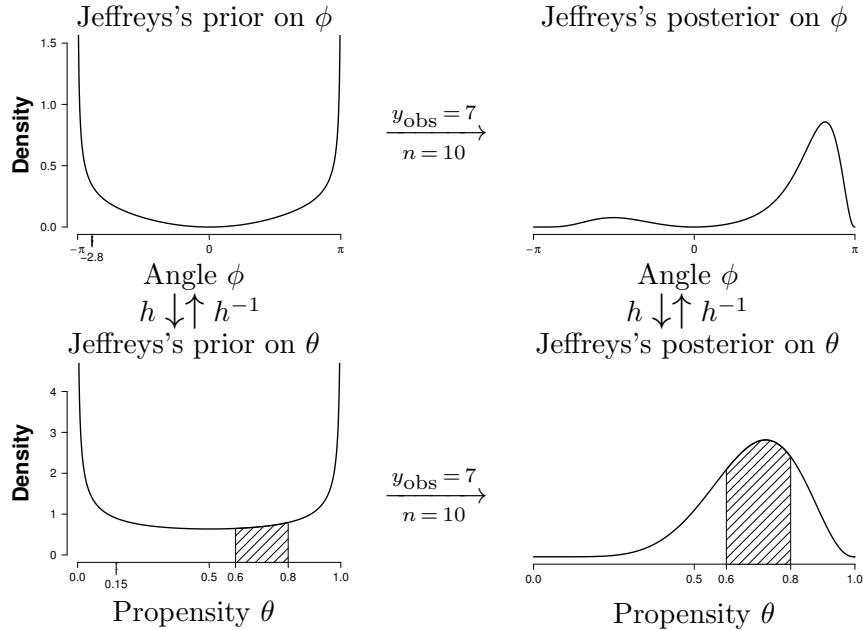


Figure 13.5: For priors constructed through Jeffreys's rule it does not matter whether the problem is represented in terms of the angles ϕ or its propensity θ . Thus, not only is the problem equivalent due to the transformations $\theta = h(\phi)$ and its backwards transformation $\phi = h^{-1}(\theta)$, the prior information is the same in both representations. This also holds for the posteriors.

1919, 1921, 1923). Jeffreys concluded that the uniform prior is unsuitable as a default prior due to its dependence on the parameterisation. As an alternative, Jeffreys (1946) proposed the following prior based on Fisher information

$$g_J(\theta) = \frac{1}{V} \sqrt{I_X(\theta)}, \text{ where } V = \int_{\Theta} \sqrt{I_X(\theta)} d\theta, \quad (13.3.4)$$

which is known as the prior derived from Jeffreys's rule or the *Jeffreys's prior* in short. The Jeffreys's prior is parameterisation-invariant, which implies that it leads to the same posteriors regardless of how the model is represented.

Example 13.3.3 (Jeffreys's prior). *The Jeffreys's prior of the Bernoulli model in terms of ϕ is*

$$g_J(\phi) = \frac{3\phi^2}{V \sqrt{\pi^6 - \phi^6}}, \text{ where } V = \pi, \quad (13.3.5)$$

which is plotted in the top-left panel of Fig. 13.5. The corresponding posterior is plotted in the top-right panel, which we transformed into a posterior in terms of θ using the function $\theta = h(\phi)$ shown in the bottom-right panel.⁸

⁸The subscript J makes explicit that the prior and posterior are based on the prior derived from Jeffreys's rule, i.e., $g_J(\theta)$ on θ , or equivalently, $g_J(\phi)$ on ϕ .

Similarly, we could have started with the Jeffreys's prior in terms of θ instead, that is,

$$g_J(\theta) = \frac{1}{V\sqrt{\theta(1-\theta)}}, \text{ where } V = \pi. \quad (13.3.6)$$

The Jeffreys's prior and posterior on θ are plotted in the bottom-left and the bottom-right panel of Fig. 13.5, respectively. The Jeffreys's prior on θ corresponds to the prior belief $P_J(\theta^* \in J_\theta) = 0.14$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x_{obs}^n with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $P_J(\theta^* \in J_\theta | x_{\text{obs}}^n) = 0.53$, see the shaded areas in Fig. 13.5. The posterior is identical to the one obtained from the previously described updating procedure that starts with the Jeffreys's prior on ϕ instead of on θ . \diamond

This example shows that the Jeffreys's prior leads to the same posterior knowledge regardless of how we as researcher represent the problem. Hence, the same conclusions about θ are drawn regardless of whether we (1) use Jeffreys's rule to construct a prior on θ and update with the observed data, or (2) use Jeffreys's rule to construct a prior on ϕ , update to a posterior distribution on ϕ , which is then transformed to a posterior on θ .

13.3.4 Geometrical properties of Fisher information

In the remainder of this section we make intuitive that the Jeffreys's prior is in fact uniform in the model space. We elaborate on what is meant by model space and how this can be viewed geometrically. This geometric approach illustrates (1) the role of Fisher information in the definition of the Jeffreys's prior, (2) the interpretation of the shaded area, and (3) why the normalisation constant is $V = \pi$, regardless of the chosen parameterisation.

13.3.4.1 The model space \mathcal{M}

Before we describe the geometry of statistical models, recall that a pmf can be thought of as a data generating device of X , as the pmf specifies the chances with which X takes on the potential outcomes 0 and 1. Each such pmf has to fulfil two conditions: (i) the chances have to be non-negative, that is, $0 \leq p(x) = P(X = x)$ for every possible outcome x of X , and (ii) to explicitly convey that there are $w = 2$ outcomes, and none more, the chances have to sum to one, that is, $p(0) + p(1) = 1$. We call the largest set of functions that adhere to conditions (i) and (ii) the complete set of pmfs \mathcal{P} .

As any pmf from \mathcal{P} defines $w = 2$ chances, we can represent such a pmf as a vector in w dimensions. To simplify notation, we write $p(X)$ for all w chances simultaneously, hence, $p(X)$ is the vector $p(X) = [p(0), p(1)]$ when $w = 2$. The two chances with which a pmf $p(X)$ generates outcomes of X can be simultaneously represented in the plane with $p(0) = P(X = 0)$ on the horizontal axis and $p(1) = P(X = 1)$ on the vertical axis. In the most extreme case, we have the pmf $p(X) = [1, 0]$ or $p(X) = [0, 1]$. These two extremes are linked by a straight line in

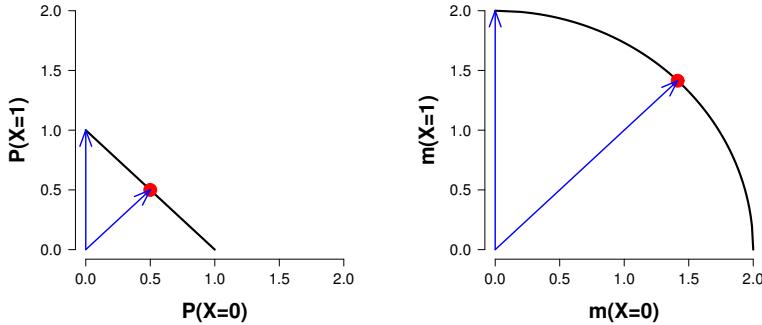


Figure 13.6: The true pmf of X with the two outcomes $\{0, 1\}$ has to lie on the line (left panel) or more naturally on the positive part of the circle (right panel). The dot represents the pmf $p_e(X)$.

the left panel of Fig. 13.6. *Any* pmf –and the true pmf $p^*(X)$ of X in particular– can be uniquely identified with a vector on the line and vice versa. For instance, the pmf $p_e(X) = [1/2, 1/2]$ (i.e., the two outcomes are generated with the same chance) is depicted as the dot on the line.

This vector representation allows us to associate to each pmf of X a norm, that is, a length. Our intuitive notion of length is based on the *Euclidean norm* and entails taking the root of the sums of squares. For instance, we can associate to the pmf $p_e(X)$ the length $\|p_e(X)\|_2 = \sqrt{(1/2)^2 + (1/2)^2} = 1/\sqrt{2} \approx 0.71$. On the other hand, the length of the pmf that states that $X = 1$ is generated with 100% chance has length one. Note that by eye, we conclude that $p_e(X)$, the arrow pointing to the dot in the left panel in Fig. 13.6 is indeed much shorter than the arrow pointing to extreme pmf $p(X) = [0, 1]$.

This mismatch in lengths can be avoided when we represent each pmf $p(X)$ by two times its square root instead (Kass, 1989), that is, by $m(X) = 2\sqrt{p(X)} = [2\sqrt{p(0)}, 2\sqrt{p(1)}]$.⁹ A pmf that is identified as the vector $m(X)$ is now two units away from the origin, that is, $\|m(X)\|_2 = \sqrt{m(0)^2 + m(1)^2} = \sqrt{4(p(0) + p(1))} = 2$. For instance, the pmf $p_e(X)$ is now represented as $m_e(X) \approx [1.41, 1.41]$. The model space \mathcal{M} is the collection of all transformed pmfs and represented as the surface of (the positive part of) a circle, see the right panel of Fig. 13.6.¹⁰ By representing the set of all possible pmfs of X as vectors $m(X) = 2\sqrt{p(X)}$ that reside on the sphere \mathcal{M} , we adopted our intuitive notion of distance. As a result, we can now, by simply looking at the figures, clarify that a uniform prior on the parameter space may lead to a very informative prior in the model space \mathcal{M} .

⁹The factor two is used to avoid a scaling of a quarter, though, its precise value is not essential for the ideas conveyed here. To simplify matters, we also call $m(X)$ a pmf.

¹⁰Hence, the model space \mathcal{M} is the collection of all functions on \mathcal{X} such that (i) $m(x) \geq 0$ for every outcome x of X , and (ii) $\sqrt{m(0)^2 + m(1)^2} = 2$. This vector representation of all the pmfs on X has the advantage that it also induces an inner product, which allows one to project one vector onto another, see Rudin (1991, p. 4), van der Vaart (1998, p. 94) and Appendix 13.E.

13.3.4.2 Uniform on the parameter space versus uniform on the model space

As \mathcal{M} represents the largest set of pmfs, any model defines a subset of \mathcal{M} . Recall that the function $f(x|\theta)$ represents how we believe a parameter θ is functionally related to an outcome x of X . For each θ this parameterisation yields a pmf $p_\theta(X)$ and, thus, also the vector $m_\theta(X) = 2\sqrt{p_\theta(X)}$. We denote the resulting set of vectors $m_\theta(X)$ so created by \mathcal{M}_Θ . For instance, the Bernoulli model $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ consists of pmfs given by $p_\theta(X) = [f(0|\theta), f(1|\theta)] = [1-\theta, \theta]$, which we represent as the vectors $m_\theta(X) = [2\sqrt{1-\theta}, 2\sqrt{\theta}]$. Doing this for every θ in the parameter space Θ yields the candidate set of pmfs \mathcal{M}_Θ . In this case, we obtain a saturated model, since $\mathcal{M}_\Theta = \mathcal{M}$, see the left panel in Fig. 13.7, where the right most square on the curve corresponds to $m_0(X) = [2, 0]$. By following the curve in an anti-clockwise manner we encounter squares that represent the pmfs $m_\theta(X)$ corresponding to $\theta = 0.1, 0.2, \dots, 1.0$ respectively. In the right panel

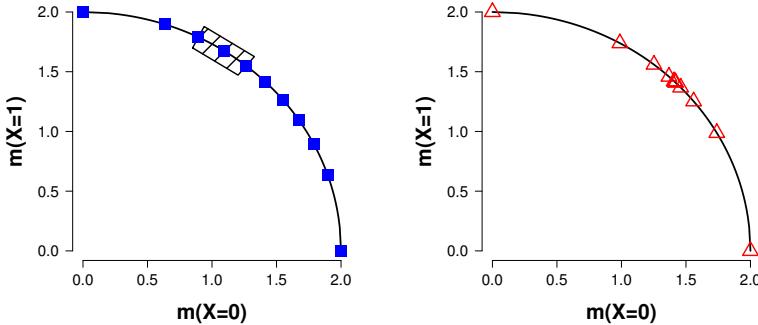


Figure 13.7: The parameterisation in terms of propensity θ (left panel) and angle ϕ (right panel) differ from each other substantially, and from a uniform prior in the model space. Left panel: The eleven squares (starting from the right bottom going anti-clockwise) represent pmfs that correspond to $\theta = 0.0, 0.1, 0.2, \dots, 0.9, 1.0$. The shaded area corresponds to the shaded area in the bottom-left panel of Fig. 13.5 and accounts for 14% of the model's length. Right panel: Similarly, the eleven triangles (starting from the right bottom going anti-clockwise) represent pmfs that correspond to $\phi = -1.0\pi, -0.8\pi, \dots, 0.8\pi, 1.0\pi$.

of Fig. 13.7 the same procedure is repeated, but this time in terms of ϕ at $\phi = -1.0\pi, -0.8\pi, \dots, 1.0\pi$. Indeed, filling in the gaps shows that the Bernoulli model in terms of θ and ϕ fully overlap with the largest set of possible pmfs, thus, $\mathcal{M}_\Theta = \mathcal{M} = \mathcal{M}_\Phi$. Fig. 13.7 makes precise what is meant when we say that the models \mathcal{M}_Θ and \mathcal{M}_Φ are equivalent; the two models define the same candidate set of pmfs that we believe to be viable data generating devices for X .

However, θ and ϕ represent \mathcal{M} in a substantially different manner. As the representation $m(X) = 2\sqrt{p(X)}$ respects our natural notion of distance, we conclude, by eye, that a uniform division of θ s with distance, say, $d\theta = 0.1$ does not lead to a uniform partition of the model. More extremely, a uniform division of ϕ with distance $d\phi = 0.2\pi$ (10% of the length of the parameter space) also does not

lead to a uniform partition of the model. In particular, even though the intervals $(-\pi, -0.8\pi)$ and $(-0.2\pi, 0)$ are of equal length in the parameter space Φ , they do not have an equal displacement in the model \mathcal{M}_Φ . In effect, the right panel of Fig. 13.7 shows that the 10% probability that the uniform prior on ϕ assigns to $\phi^* \in (-\pi, -0.8\pi)$ in parameter space is redistributed over a larger arc length of the model \mathcal{M}_Φ compared to the 10% assigned to $\phi^* \in (-0.2\pi, 0)$. Thus, a uniform distribution on ϕ favours the pmfs $m_\phi(X)$ with ϕ close to zero. Note that this effect is cancelled by the Jeffreys's prior, as it puts more mass near the end points compared to $\phi = 0$, see the top-left panel of Fig. 13.5. Similarly, the left panel of Fig. 13.7 shows that the uniform prior $g(\theta)$ also fails to yield an equiprobable assessment of the pmfs in model space. Again, the Jeffreys's prior in terms of θ compensates for the fact that the interval $(0, 0.1)$ as compared to $(0.5, 0.6)$ in Θ is more spread out in model space. However, it does so less severely compared to the Jeffreys's prior on ϕ . To illustrate, we added additional tick marks on the horizontal axis of the priors in the left panels of Fig. 13.5. The tick mark at $\phi = -2.8$ and $\theta = 0.15$ both indicate the 25% quantiles of their respective Jeffreys's priors. Hence, the Jeffreys's prior allocates more mass to the boundaries of ϕ than to the boundaries of θ to compensate for the difference in geometry, see Fig. 13.7. More generally, the Jeffreys's prior uses Fisher information to convert the geometry of the model to the parameter space.

Note that because the Jeffreys's prior is specified using the Fisher information, it takes the functional relationship $f(x|\theta)$ into account. The functional relationship makes precise how the parameter is linked to the data and, thus, gives meaning and context to the parameter. On the other hand, a prior on ϕ specified without taking the functional relationship $f(x|\phi)$ into account is a prior that neglects the context of the problem. For instance, the right panel of Fig. 13.7 shows that this neglect with a uniform prior on ϕ results in having the geometry of $\Phi = (-\pi, \pi)$ forced onto the model \mathcal{M}_Φ .

13.3.5 Uniform prior on the model

Fig. 13.7 shows that neither a uniform prior on θ , nor a uniform prior on ϕ yields a uniform prior on the model. Alternatively, we can begin with a uniform prior on the model \mathcal{M} and convert this into priors on the parameter spaces Θ and Φ . This uniform prior on the model translated to the parameters is exactly the Jeffreys's prior.

Recall that a prior on a space S is uniform, if it has the following two defining features: (i) the prior is proportional to one, and (ii) a normalisation constant given by $V_S = \int_S 1 ds$ that equals the length, more generally, volume of S . For instance, a replacement of s by ϕ and S by $\Phi = (-\pi, \pi)$ yields the uniform prior on the angles with the normalisation constant $V_\Phi = \int_\Phi 1 d\phi = 2\pi$. Similarly, a replacement of s by the pmf $m_\theta(X)$ and S by the function space \mathcal{M}_Θ yields a uniform prior on the model \mathcal{M}_Θ . The normalisation constant then becomes a daunting looking integral in terms of displacements $dm_\theta(X)$ between vectors in model space \mathcal{M}_Θ . Fortunately, it can be shown, see Appendix 13.C, that V

simplifies to

$$V = \int_{\mathcal{M}_\Theta} 1 dm_\theta(X) = \int_{\Theta} \sqrt{I_X(\theta)} d\theta. \quad (13.3.7)$$

Thus, V can be computed in terms of θ by multiplying the distances $d\theta$ in Θ by the root of the Fisher information. Heuristically, this means that the root of the Fisher information translates displacements $dm_\theta(X)$ in the model \mathcal{M}_Θ to distances $\sqrt{I_X(\theta)}d\theta$ in the parameter space Θ .

Recall from Example 13.3.3 that regardless of the parameterisation, the normalisation constant of the Jeffreys's prior was π . To verify that this is indeed the length of the model, we use the fact that the circumference of a quarter circle with radius $r = 2$ can also be calculated as $V = (2\pi r)/4 = \pi$.

Given that the Jeffreys's prior corresponds to a uniform prior on the model, we deduce that the shaded area in the bottom-left panel of Fig. 13.5 with $P_J(\theta^* \in J_\theta) = 0.14$, implies that the model interval $J_m = (m_{0.6}(X), m_{0.8}(X))$, the shaded area in the left panel of Fig. 13.7, accounts for 14% of the model's length. After updating the Jeffreys's prior with the observations x_{obs}^n consisting of $y_{\text{obs}} = 7$ out of $n = 10$ the probability of finding the true data generating pmf $m^*(X)$ in this interval of pmfs J_m is increased to 53%.

In conclusion, we verified that the Jeffreys's prior is a prior that leads to the same conclusion regardless of how we parameterise the problem. This property is a direct result of shifting our focus from finding the true parameter value within the parameter space to the proper formulation of the estimation problem –as discovering the true data generating pmf $m_{\theta^*}(X) = 2\sqrt{p_{\theta^*}(X)}$ in \mathcal{M}_Θ and by expressing our prior ignorance as a uniform prior on the model \mathcal{M}_Θ .

13.4 The role of Fisher information in minimum description length

In this section we graphically show how Fisher information is used as a measure of model complexity and its role in model selection within the minimum description length framework (MDL; de Rooij and Grünwald, 2011; Grünwald et al., 2005; Grünwald, 2007; Myung et al., 2000c; Myung et al., 2006; Pitt et al., 2002).

The primary aim of a model selection procedure is to select a single model from a set of competing models, say, models \mathcal{M}_1 and \mathcal{M}_2 , that best suits the observed data x_{obs}^n . Many model selection procedures have been proposed in the literature, but the most popular methods are those based on penalised maximum likelihood criteria, such as the Akaike information criterion (AIC; Akaike, 1974; Burnham and Anderson, 2002), the Bayesian information criterion (BIC; Raftery, 1995; Schwarz, 1978), and the Fisher information approximation (FIA; Grünwald,

2007; Rissanen, 1996). These criteria are defined as follows

$$\text{AIC} = -2 \log f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n)) + 2d_j, \quad (13.4.1)$$

$$\text{BIC} = -2 \log f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n)) + d_j \log(n), \quad (13.4.2)$$

$$\text{FIA} = \underbrace{-\log f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n))}_{\text{Goodness-of-fit}} + \underbrace{\frac{d_j}{2} \log \frac{n}{2\pi}}_{\text{Dimensionality}} + \underbrace{\log \left(\int_{\Theta} \sqrt{\det I_{\mathcal{M}_j}(\theta_j)} d\theta_j \right)}_{\text{Geometric complexity}},$$

where n denotes the sample size, d_j the number of free parameters, $\hat{\theta}_j$ the MLE, $I_{\mathcal{M}_j}(\theta_j)$ the unit Fisher information, and f_j the functional relationship between the potential outcome x^n and the parameters θ_j within model \mathcal{M}_j .¹¹ Hence, except for the observations x_{obs}^n , all quantities in the formulas depend on the model \mathcal{M}_j . We made this explicit using a subscript j to indicate that the quantity, say, θ_j belongs to model \mathcal{M}_j .¹² For all three criteria, the model yielding the lowest criterion value is perceived as the model that generalises best (Myung and Pitt, 2016).

Each of the three model selection criteria tries to strike a balance between model fit and model complexity. Model fit is expressed by the goodness-of-fit terms, which involves replacing the potential outcomes x^n and the unknown parameter θ_j of the functional relationships f_j by the actually observed data x_{obs}^n , as in the Bayesian setting, and the maximum likelihood estimate $\hat{\theta}_j(x_{\text{obs}}^n)$, as in the frequentist setting.

The positive terms in the criteria account for model complexity. A penalisation of model complexity is necessary, because the support in the data cannot be assessed by solely considering goodness-of-fit, as the ability to fit observations increases with model complexity (e.g., Roberts and Pashler, 2000). As a result, the more complex model necessarily leads to better fits but may in fact overfit the data. The overly complex model then captures idiosyncratic noise rather than general structure, resulting in poor model generalisability (Myung et al., 2000c; Wagenmakers and Waldorp, 2006b).

The focus in this section is to make intuitive how FIA acknowledges the trade-off between goodness-of-fit and model complexity in a principled manner by graphically illustrating this model selection procedure, see also Balasubramanian (1996), Kass (1989), Klaassen and Lenstra (2003), Myung et al. (2000a), and Rissanen (1996). We exemplify the concepts with simple multinomial processing tree (MPT) models (e.g., Batchelder and Riefer, 1999; Klauer and Kellen, 2011; Wu et al., 2010). For a more detailed treatment of the subject we refer to Appendix 13.D, de Rooij and Grünwald (2011), Grünwald (2007), Myung et al. (2006), and the references therein.

¹¹For vector-valued parameters θ_j , we have a Fisher information matrix and $\det I_{\mathcal{M}_j}(\theta_j)$ refers to the determinant of this matrix. This determinant is always non-negative, because the Fisher information matrix is always a positive semidefinite symmetric matrix. Intuitively, volumes and areas cannot be negative (Appendix 13.C.3.3).

¹²For the sake of clarity, we will use different notations for the parameters within the different models. We introduce two models in this section: the model \mathcal{M}_1 with parameter $\theta_1 = \vartheta$ which we pit against the model \mathcal{M}_2 with parameter $\theta_2 = \alpha$.

13.4.0.1 The description length of a model

Recall that each model specifies a functional relationship f_j between the potential outcomes of X and the parameters θ_j . This f_j is used to define a so-called *normalised maximum likelihood* (NML) code. For the j th model its NML code is defined as

$$p_{\text{NML}}(x_{\text{obs}}^n | \mathcal{M}_j) = \frac{f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n))}{\sum_{x^n \in \mathcal{X}^n} f_j(x^n | \hat{\theta}_j(x^n))}, \quad (13.4.3)$$

where the sum in the denominator is over all possible outcomes x^n in \mathcal{X}^n , and where $\hat{\theta}_j$ refers to the MLE within model \mathcal{M}_j . The NML code is a relative goodness-of-fit measure, as it compares the observed goodness-of-fit term against the sum of all possible goodness-of-fit terms. Note that the actual observations x_{obs}^n only affect the numerator, by a plugin of x_{obs}^n and its associated maximum likelihood estimate $\hat{\theta}(x_{\text{obs}}^n)$ into the functional relationship f_j belonging to model \mathcal{M}_j . The sum in the denominator consists of the same plugins, but for every possible realisation of X^n .¹³ Hence, the denominator can be interpreted as a measure of the model's collective goodness-of-fit or the model's fit capacity. Consequently, for every set of observations x_{obs}^n , the NML code outputs a number between zero and one that can be transformed into a non-negative number by taking the negative logarithm as¹⁴

$$-\log p_{\text{NML}}(x_{\text{obs}}^n | \mathcal{M}_j) = -\log f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n)) + \underbrace{\log \sum_{x^n \in \mathcal{X}^n} f_j(x^n | \hat{\theta}_j(x^n))}_{\text{Model complexity}}, \quad (13.4.4)$$

which is called the description length of model \mathcal{M}_j . Within the MDL framework, the model with the shortest description length is the model that best describes the observed data x_{obs}^n .

The model complexity term is typically hard to compute, but Rissanen (1996) showed that it can be well-approximated by the dimensionality and the geometrical complexity terms. That is,

$$\text{FIA} = -\log f_j(x_{\text{obs}}^n | \hat{\theta}_j(x_{\text{obs}}^n)) + \frac{d_j}{2} \log \frac{n}{2\pi} + \log \left(\int_{\Theta} \sqrt{\det I_{\mathcal{M}_j}(\theta_j)} d\theta_j \right),$$

is an approximation of the description length of model \mathcal{M}_j . The determinant is simply the absolute value when the number of free parameters d_j is equal to one. Furthermore, the integral in the geometrical complexity term coincides with the normalisation constant of the Jeffreys's prior, which represented the volume of the model. In other words, a model's fit capacity is proportional to its volume in model space as one would expect.

In sum, within the MDL philosophy, a model is selected if it yields the shortest description length, as this model uses the functional relationship f_j that best

¹³As before, for continuous data, the sum is replaced by an integral.

¹⁴Quite deceptively the minus sign actually makes this measure positive, as $-\log(y) = \log(1/y) \geq 0$ if $0 \leq y \leq 1$.

extracts the regularities from x_{obs}^n . As the description length is often hard to compute, we approximate it with FIA instead (Heck et al., 2014). To do so, we have to characterise (1) all possible outcomes of X , (2) propose at least two models which will be pitted against each other, and (3) identify the model characteristics: the MLE $\hat{\theta}_j$ corresponding to \mathcal{M}_j , and its volume $V_{\mathcal{M}_j}$. In the remainder of this section we show that FIA selects the model that is closest to the data with an additional penalty for model complexity.

13.4.1 A new running example and the geometry of a random variable with $w = 3$ outcomes

To graphically illustrate the model selection procedure underlying MDL we introduce a random variable X that has $w = 3$ number of potential outcomes.

Example 13.4.1 (A psychological task with three outcomes). *In the training phase of a source-memory task, the participant is presented with two lists of words on a computer screen. List \mathcal{L} is projected on the left-hand side and list \mathcal{R} is projected on the right-hand side. In the test phase, the participant is presented with two words, side by side, that can stem from either list, thus, ll, lr, rl, rr. At each trial, the participant is asked to categorise these pairs as either:*

- L meaning both words come from the left list, i.e., “ll”,
- M meaning the words are mixed, i.e., “lr” or “rl”,
- R meaning both words come from the right list, i.e., “rr”.

For simplicity we assume that the participant will be presented with n test pairs X^n of equal difficulty. \diamond

For the graphical illustration of this new running example, we generalise the ideas presented in Section 13.3.4.1 from $w = 2$ to $w = 3$ dimensions. Recall that a pmf of X with w number of outcomes can be written as a w -dimensional vector. For the task described above we know that a data generating pmf defines the three chances $p(X) = [p(L), p(M), p(R)]$ with which X generates the outcomes $[L, M, R]$ respectively.¹⁵ As chances cannot be negative, (i) we require that $0 \leq p(x) = P(X = x)$ for every outcome x in \mathcal{X} , and (ii) to explicitly convey that there are $w = 3$ outcomes, and none more, these $w = 3$ chances have to sum to one, that is, $\sum_{x \in \mathcal{X}} p(x) = 1$. We call the largest set of functions that adhere to conditions (i) and (ii) the complete set of pmfs \mathcal{P} . The three chances with which a pmf $p(X)$ generates outcomes of X can be simultaneously represented in three-dimensional space with $p(L) = P(X = L)$ on the left most axis, $p(M) = P(X = M)$ on the right most axis and $p(R) = P(X = R)$ on the vertical axis as shown in the left panel of Fig. 13.8.¹⁶ In the most extreme case, we have the pmf $p(X) = [1, 0, 0]$, $p(X) = [0, 1, 0]$ or $p(X) = [0, 0, 1]$, which correspond to the corners of the triangle

¹⁵As before we write $p(X) = [p(L), p(M), p(R)]$ with a capital X to denote all the w number of chances simultaneously and we used the shorthand notation $p(L) = p(X = L)$, $p(M) = p(X = M)$ and $p(R) = p(X = R)$.

¹⁶This is the three-dimensional generalisation of Fig. 13.6.

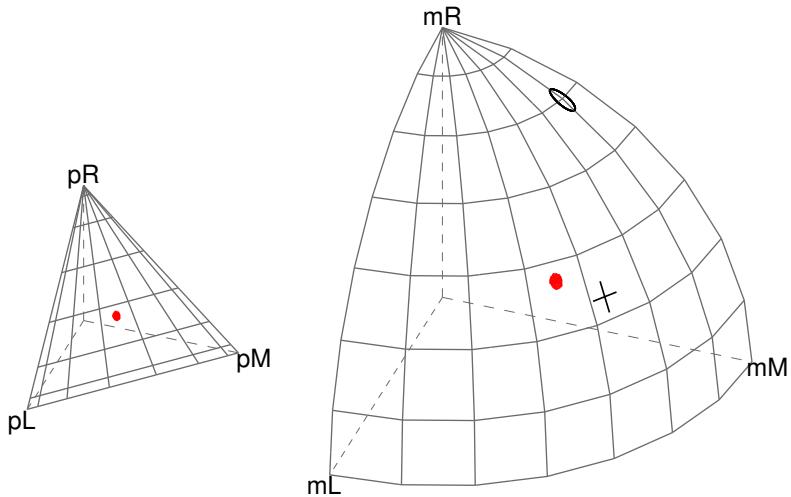


Figure 13.8: Every point on the sphere corresponds to a pmf of a categorical distribution with $w = 3$ categories. In particular, the (red) dot refers to the pmf $p_e(x) = [1/3, 1/3, 1/3]$, the circle represents the pmf given by $p(X) = [0.01, 0.18, 0.81]$, while the cross represents the pmf $p(X) = [0.25, 0.5, 0.25]$.

indicated by pL , pM and pR , respectively. These three extremes are linked by a triangular plane in the left panel of Fig. 13.8. Any pmf –and the true pmf $p^*(X)$ in particular– can be uniquely identified with a vector on the triangular plane and vice versa. For instance, a possible true pmf of X is $p_e(X) = [1/3, 1/3, 1/3]$ (i.e., the outcomes L , M and R are generated with the same chance) and depicted as a (red) dot on the simplex.

This vector representation allows us to associate to each pmf of X the Euclidean norm. For instance, the representation in the left panel of Fig. 13.8 leads to an extreme pmf $p(X) = [1, 0, 0]$ that is one unit long, while $p_e(X) = [1/3, 1/3, 1/3]$ is only $\sqrt{(1/3)^2 + (1/3)^2 + (1/3)^2} \approx 0.58$ units away from the origin. As before, we can avoid this mismatch in lengths by considering the vectors $m(X) = 2\sqrt{p(X)}$, instead. Any pmf that is identified as $m(X)$ is now two units away from the origin. The model space \mathcal{M} is the collection of all transformed pmfs and represented as the surface of (the positive part of) the sphere in the right panel of Fig. 13.8. By representing the set of all possible pmfs of X as $m(X) = 2\sqrt{p(X)}$, we adopted our intuitive notion of distance. As a result, the selection mechanism underlying MDL can be made intuitive by simply looking at the forthcoming plots.

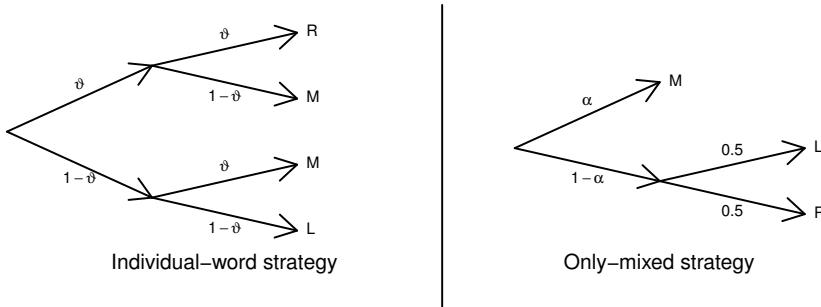


Figure 13.9: Two MPT models that theorise how a participant chooses the outcomes L , M , or R in the source-memory task described in the main text. The left panel schematically describes the individual-word strategy, while the right model schematically describes the only-mixed strategy.

13.4.2 The individual-word and the only-mixed strategy

To ease the exposition, we assume that both words presented to the participant come from the right list \mathcal{R} , thus, “*rr*” for the two models introduced below. As model \mathcal{M}_1 we take the so-called individual-word strategy. Within this model \mathcal{M}_1 , the parameter is $\theta_1 = \vartheta$, which we interpret as the participant’s “right-list recognition ability”. With chance ϑ the participant then correctly recognises that the first word originates from the right list and repeats this procedure for the second word, after which the participant categorises the word pair as L , M , or R , see the left panel of Fig. 13.9 for a schematic description of this strategy as a processing tree. Fixing the participant’s “right-list recognition ability” ϑ yields the following pmf

$$f_1(X | \vartheta) = [(1 - \vartheta)^2, 2\vartheta(1 - \vartheta), \vartheta^2]. \quad (13.4.5)$$

For instance, when the participant’s true ability is $\vartheta^* = 0.9$, the three outcomes $[L, M, R]$ are then generated with the following three chances $f_1(X | 0.9) = [0.01, 0.18, 0.81]$, which is plotted as a circle in Fig. 13.8. On the other hand, when $\vartheta^* = 0.5$ the participant’s generating pmf is then $f_1(X | \vartheta = 0.5) = [0.25, 0.5, 0.25]$, which is depicted as the cross in model space \mathcal{M} . The set of pmfs so defined forms a curve that goes through both the cross and the circle, see the left panel of Fig. 13.10.

As a competing model \mathcal{M}_2 , we take the so-called only-mixed strategy. For the task described in Example 13.4.1, we might pose that participants from a certain clinical group are only capable of recognising mixed word pairs and that they are unable to distinguish the pairs “*rr*” from “*ll*” resulting in a random guess between the responses L and R , see the right panel of Fig. 13.9 for the processing tree. Within this model \mathcal{M}_2 the parameter is $\theta_2 = \alpha$, which is interpreted as the participant’s “mixed-list differentiability skill” and fixing it yields the following pmf

$$f_2(X | \alpha) = [(1 - \alpha)/2, \alpha, (1 - \alpha)/2]. \quad (13.4.6)$$

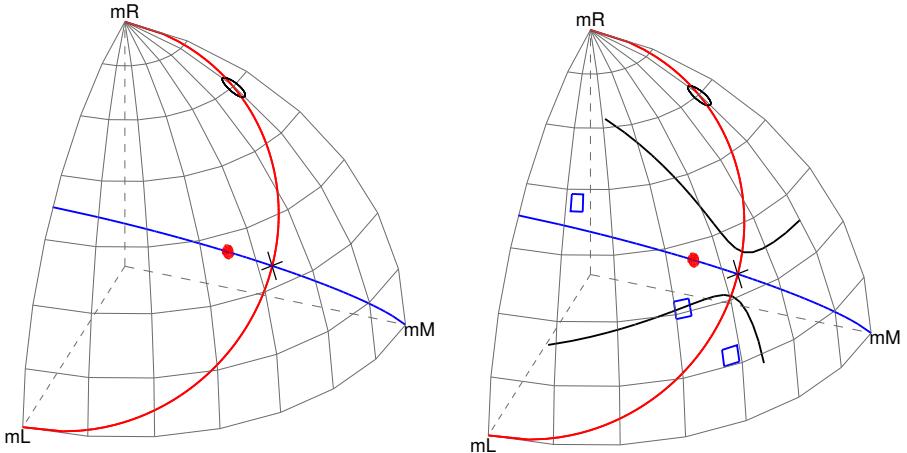


Figure 13.10: Left panel: The set of pmfs that are defined by the individual-list strategy \mathcal{M}_1 forms a curve that goes through both the cross and the circle, while the pmfs of the only-mixed strategy \mathcal{M}_2 correspond to the curve that goes through both the cross and the dot. Right panel: The model selected by FIA can be thought of as the model closest to the empirical pmf with an additional penalty for model complexity. The selection between the individual-list and the only-mixed strategy by FIA based on $n = 30$ trials is formalised by the additional curves – the only-mixed strategy is preferred over the individual-list strategy, when the observations yield an empirical pmf that lies between the two non-decision curves. The top, middle and bottom squares corresponding to the data sets $x_{\text{obs},1}^n, x_{\text{obs},2}^n$ and $x_{\text{obs},3}^n$ in Table 13.1, which are best suited to \mathcal{M}_2 , either, and \mathcal{M}_1 , respectively. The additional penalty is most noticeable at the cross, where the two models share a pmf. Observations with $n = 30$ yielding an empirical pmf in this area are automatically assigned to the simpler model, i.e., the only-mixed strategy \mathcal{M}_2 .

For instance, when the participant's true differentiability is $\alpha^* = 1/3$, we have $f_2(X | 1/3) = [1/3, 1/3, 1/3]$, which, as before, is plotted as the dot in Fig. 13.10. On the other hand, when $\alpha^* = 0.5$ the participant's generating pmf is then given by $f_2(X | \alpha = 0.5) = [0.25, 0.5, 0.25]$, i.e., the cross. The set of pmfs so defined forms a curve that goes through both the dot and the cross, see the left panel of Fig. 13.10.

The plots show that the models \mathcal{M}_1 and \mathcal{M}_2 are neither saturated nor nested, as the two models define proper subsets of \mathcal{M} and only overlap at the cross. Furthermore, the plots also show that \mathcal{M}_1 and \mathcal{M}_2 are both one-dimensional, as each model is represented as a line in model space. Hence, the dimensionality terms in all three information criteria are the same. Consequently, AIC and BIC will only discriminate these two models based on goodness-of-fit alone. This particular model comparison, thus, allows us to highlight the role Fisher information plays

in the MDL model selection philosophy.

13.4.3 Model characteristics

13.4.3.1 The maximum likelihood estimators

For FIA we need to compute the goodness-of-fit terms, thus, we need to identify the MLEs for the parameters within each model. For the models at hand, the MLEs are

$$\hat{\theta}_1 = \hat{\vartheta} = (Y_M + 2Y_R)/(2n) \text{ for } \mathcal{M}_1, \text{ and } \hat{\theta}_2 = \hat{\alpha} = Y_M/n \text{ for } \mathcal{M}_2, \quad (13.4.7)$$

where Y_L , Y_M and $Y_R = n - Y_L - Y_M$ are the number of L , M and R responses in the data consisting of n trials.

Estimation is a within model operation and it can be viewed as projecting the so-called *empirical* (*i.e.*, *observed*) pmf corresponding to the data onto the model. For iid data with $w = 3$ outcomes the empirical pmf corresponding to x_{obs}^n is defined as $\hat{p}_{\text{obs}}(X) = [y_L/n, y_M/n, y_R/n]$. Hence, the empirical pmf gives the relative occurrence of each outcome in the sample. For instance, the observations x_{obs}^n consisting of $[y_L = 3, y_M = 3, y_R = 3]$ responses correspond to the observed pmf $\hat{p}_{\text{obs}}(X) = [1/3, 1/3, 1/3]$, *i.e.*, the dot in Fig. 13.10. Note that this observed pmf $\hat{p}_{\text{obs}}(X)$ does not reside on the curve of \mathcal{M}_1 .

Nonetheless, when we use the MLE $\hat{\vartheta}$ of \mathcal{M}_1 , we as researchers bestow the participant with a “right-list recognition ability” ϑ and implicitly assume that she used the individual-word strategy to generate the observations. In other words, we only consider the pmfs on the curve of \mathcal{M}_1 as viable explanations of how the participant generated her responses. For the data at hand, we have the estimate $\hat{\vartheta}_{\text{obs}} = 0.5$. If we were to generalise the observations x_{obs}^n under \mathcal{M}_1 , we would then plug this estimate into the functional relationship f_1 resulting in the predictive pmf $f_1(X | \hat{\vartheta}_{\text{obs}}) = [0.25, 0.5, 0.25]$. Hence, even though the number of L , M and R responses were equal in the observations x_{obs}^n , under \mathcal{M}_1 we expect that this participant will answer with twice as many M responses compared to the L and R responses in a next set of test items. Thus, for predictions, part of the data is ignored and considered as noise.

Geometrically, the generalisation $f_1(X | \hat{\vartheta}_{\text{obs}})$ is a result of projecting the observed pmf $\hat{p}_{\text{obs}}(X)$, *i.e.*, the dot, onto the cross that does reside on the curve of \mathcal{M}_1 .¹⁷ Observe that amongst all pmfs on \mathcal{M}_1 , the projected pmf is closest to the empirical pmf $\hat{p}_{\text{obs}}(X)$. Under \mathcal{M}_1 the projected pmf $f_1(X | \hat{\vartheta}_{\text{obs}})$, *i.e.*, the cross, is perceived as structural, while any deviations from the curve of \mathcal{M}_1 is labelled as noise. When generalising the observations, we ignore noise. Hence, by estimating the parameter ϑ , we implicitly restrict our predictions to only those pmfs that are defined by \mathcal{M}_1 . Moreover, evaluating the prediction at x_{obs}^n and, subsequently, taking the negative logarithm yields the goodness-of-fit term; in this case, $-\log f_1(x_{\text{obs}}^n | \hat{\vartheta}_{\text{obs}} = 0.5) = 10.4$.

¹⁷This resulting pmf $f_1(X | \hat{\vartheta}_{\text{obs}})$ is also known as the Kullback-Leibler projection of the empirical pmf $\hat{p}_{\text{obs}}(X)$ onto the model \mathcal{M}_1 . White (1982) used this projection to study the behaviour of the MLE under model misspecification.

Which part of the data is perceived as structural or as noise depends on the model. For instance, when we use the MLE $\hat{\alpha}$, we restrict our predictions to the pmfs of \mathcal{M}_2 . For the data at hand, we get $\hat{\alpha}_{\text{obs}} = 1/3$ and the plugin yields $f_2(X | \hat{\alpha}_{\text{obs}}) = [1/3, 1/3, 1/3]$. Again, amongst all pmfs on \mathcal{M}_2 , the projected pmf is closest to the empirical pmf $\hat{p}_{\text{obs}}(X)$. In this case, the generalisation under \mathcal{M}_2 coincides with the observed pmf $\hat{p}_{\text{obs}}(X)$. Hence, under \mathcal{M}_2 there is no noise, as the empirical pmf $\hat{p}_{\text{obs}}(X)$ was already on the model. Geometrically, this means that \mathcal{M}_2 is closer to the empirical pmf than \mathcal{M}_1 , which results in a lower goodness-of-fit term $-\log f_2(x_{\text{obs}}^n | \hat{\alpha}_{\text{obs}} = 1/3) = 9.9$.

This geometric interpretation allows us to make intuitive that data sets with the same goodness-of-fit terms will be as far from \mathcal{M}_1 as from \mathcal{M}_2 . Equivalently, \mathcal{M}_1 and \mathcal{M}_2 identify the same amount of noise within x_{obs}^n , when the two models fit the observations equally well. For instance, Fig. 13.10 shows that observations x_{obs}^n with an empirical pmf $\hat{p}_{\text{obs}}(X) = [0.25, 0.5, 0.25]$ are equally far from \mathcal{M}_1 as from \mathcal{M}_2 . Note that the closest pmf on \mathcal{M}_1 and \mathcal{M}_2 are both equal to the empirical pmf, as $f_1(X | \hat{\vartheta}_{\text{obs}} = 0.5) = \hat{p}_{\text{obs}}(X) = f_2(X | \hat{\alpha}_{\text{obs}} = 1/2)$. As a result, the two goodness-of-fit terms will be equal to each other.

In sum, goodness-of-fit measures a model's proximity to the observed data. Consequently, models that take up more volume in model space will be able to be closer to a larger number of data sets. In particular, when, say, \mathcal{M}_3 is nested within \mathcal{M}_4 , this means that the distance between $\hat{p}_{\text{obs}}(X)$ and \mathcal{M}_3 (noise) is at least the distance between $\hat{p}_{\text{obs}}(X)$ and \mathcal{M}_4 . Equivalently, for any data set, \mathcal{M}_4 will automatically label more of the observations as structural. Models that excessively identify parts of the observations as structural are known to overfit the data. Overfitting has an adverse effect on generalisability, especially when n is small, as $\hat{p}_{\text{obs}}(X)$ is then dominated by sampling error. In effect, the more voluminous model will then use this sampling error, rather than the structure, for its predictions. To guard ourselves from overfitting, thus, bad generalisability, the information criteria AIC, BIC and FIA all penalise for model complexity. AIC and BIC only do this via the dimensionality terms, while FIA also take the models' volumes into account.

13.4.3.2 Geometrical complexity

For both models the dimensionality term is given by $\frac{1}{2} \log(\frac{n}{2\pi})$. Recall that the geometrical complexity term is the logarithm of the model's volume, which for the individual-word and the only-mixed strategy are given by

$$V_{\mathcal{M}_1} = \int_0^1 \sqrt{I_{\mathcal{M}_1}(\theta)} d\theta = \sqrt{2}\pi \quad \text{and} \quad V_{\mathcal{M}_2} = \int_0^1 \sqrt{I_{\mathcal{M}_2}(\alpha)} d\alpha = \pi, \quad (13.4.8)$$

respectively. Hence, the individual-word strategy is a more complex model, because it has a larger volume, thus, capacity to fit data compared to the only-mixed strategy. After taking logs, we see that the individual-word strategy incurs an additional penalty of $1/2 \log(2)$ compared to the only-mixed strategy.

13.4.4 Model selection based on the minimum description length principle

With all model characteristics at hand, we only need observations to illustrate that MDL model selection boils down to selecting the model that is closest to the observations with an additional penalty for model complexity. Table 13.1 shows

Table 13.1: The description lengths for three observations $x_{\text{obs}}^n = [y_L, y_M, y_R]$, where y_L, y_M, y_R are the number of observed responses L, M and R respectively.

$x_{\text{obs}}^n = [y_L, y_M, y_R]$	$\text{FIA}_{\mathcal{M}_1}(x_{\text{obs}}^n)$	$\text{FIA}_{\mathcal{M}_2}(x_{\text{obs}}^n)$	Preferred model
$x_{\text{obs},1}^n = [12, 1, 17]$	42	26	\mathcal{M}_2
$x_{\text{obs},2}^n = [14, 10, 6]$	34	34	tie
$x_{\text{obs},3}^n = [12, 16, 2]$	29	32	\mathcal{M}_1

three data sets $x_{\text{obs},1}^n, x_{\text{obs},2}^n, x_{\text{obs},3}^n$ with $n = 30$ observations. The three associated empirical pmfs are plotted as the top, middle and lower rectangles in the right panel of Fig. 13.10, respectively. Table 13.1 also shows the approximation of each model's description length using FIA. Note that the first observed pmf, the top rectangle in Fig. 13.10, is closer to \mathcal{M}_2 than to \mathcal{M}_1 , while the third empirical pmf, the lower rectangle, is closer to \mathcal{M}_1 . Of particular interest is the middle rectangle, which lies on an additional black curve that we refer to as a non-decision curve; observations that correspond to an empirical pmf that lies on this curve are described equally well by \mathcal{M}_1 and \mathcal{M}_2 . For this specific comparison, we have the following decision rule: FIA selects \mathcal{M}_2 as the preferred model whenever the observations correspond to an empirical pmf between the two non-decision curves, otherwise, FIA selects \mathcal{M}_1 . Fig. 13.10 shows that FIA, indeed, selects the model that is closest to the data except in the area where the two models overlap – observations consisting of $n = 30$ trials with an empirical pmf near the cross are considered better described by the simpler model \mathcal{M}_2 . Hence, this yields an incorrect decision even when the empirical pmf is exactly equal to the true data generating pmf that is given by, say, $f_1(X | \vartheta = 0.51)$. This automatic preference for the simpler model, however, decreases as n increases. The left and right panel of Fig. 13.11 show the non-decision curves when $n = 120$ and n (extremely) large, respectively. As a result of moving non-decision bounds, the data set $x_{\text{obs},4}^n = [56, 40, 24]$ that has the same observed pmf as $x_{\text{obs},2}^n$, i.e., the middle rectangle, will now be better described by model \mathcal{M}_1 .

For (extremely) large n , the additional penalty due to \mathcal{M}_1 being more voluptuous than \mathcal{M}_2 becomes irrelevant and the sphere is then separated into quadrants: observations corresponding to an empirical pmf in the top-left or bottom-right quadrant are better suited to the only-mixed strategy, while the top-right and bottom-left quadrants indicate a preference for the individual-word strategy \mathcal{M}_1 . Note that pmfs on the non-decision curves in the right panel of Fig. 13.11 are as far apart from \mathcal{M}_1 as from \mathcal{M}_2 , which agrees with our geometric interpretation of goodness-of-fit as a measure of the model's proximity to the data. This quadrant division is only based on the two models' goodness-of-fit terms and yields the same selection as one would get from BIC (e.g., Rissanen, 1996). For large n , FIA, thus,

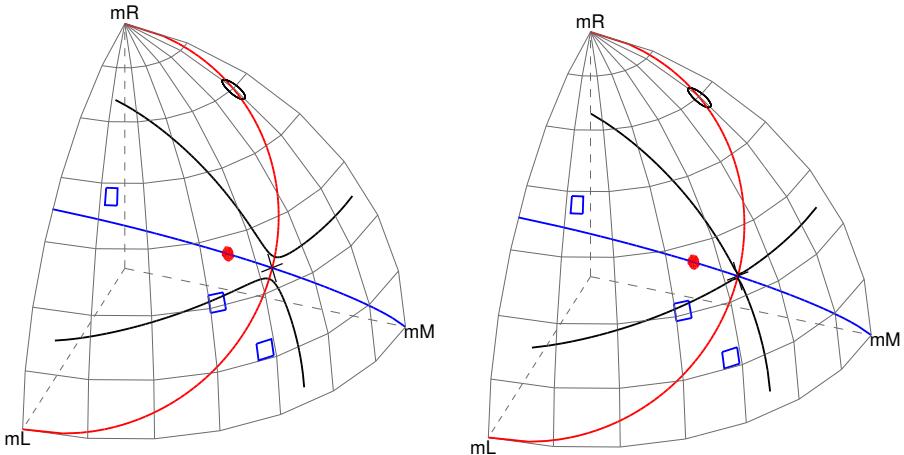


Figure 13.11: For n large the additional penalty for model complexity becomes irrelevant. The plotted non-decision curves are based on $n = 120$ and $n = 10,000$ trials in the left and right panel respectively. In the right panel only the goodness-of-fit matters in the model comparison. The model selected is then the model that is closest to the observations.

selects the model that is closest to the empirical pmf. This behaviour is desirable, because asymptotically the empirical pmf is not distinguishable from the true data generating pmf. As such, the model that is closest to the empirical pmf will then also be closest to the true pmf. Hence, FIA asymptotically selects the model that is closest to the true pmf. As a result, the projected pmf within the closest model is then expected to yield the best predictions amongst the competing models.

13.4.5 Fisher information and generalisability

Model selection by MDL is sometimes perceived as a formalisation of Occam's razor (e.g., Balasubramanian, 1996; Grünwald, 1998), a principle that states that the most parsimonious model should be chosen when the models under consideration fit the observed data equally well. This preference for the parsimonious model is based on the belief that the simpler model is better at predicting new (as yet unseen) data coming from the same source, as was shown by Pitt et al. (2002) with simulated data.

To make intuitive why the more parsimonious model, on average, leads to better predictions, we assume, for simplicity, that the true data generating pmf is given by $f(X | \theta^*)$, thus, the existence of a true parameter value θ^* . As the observations are expected to be contaminated with sampling error, we also expect an estimation error, i.e., a distance $d\theta$ between the maximum likelihood estimate $\hat{\theta}_{\text{obs}}$ and the true θ^* . Recall that in the construction of Jeffreys's prior Fisher information was used to convert displacement in model space to distances on

parameter space. Conversely, Fisher information transforms the estimation error in parameter space to a generalisation error in model space. Moreover, the larger the Fisher information at θ^* is, the more it will expand the estimation error into a displacement between the prediction $f(X | \hat{\theta}_{\text{obs}})$ and the true pmf $f(X | \theta^*)$. Thus, a larger Fisher information at θ^* will push the prediction further from the true pmf resulting in a bad generalisation. Smaller models have, on average, a smaller Fisher information at θ^* and will therefore lead to more stable predictions that are closer to the true data generating pmf. Note that the generalisation scheme based on the MLE plugin $f(X | \hat{\theta}_{\text{obs}})$ ignores the error at each generalisation step. The Bayesian counterpart, on the other hand, does take these errors into account, see Dawid (2011), Ly et al. (2017b), Marsman et al. (2016a) and see van Erven et al. (2012), Grünwald and Mehta (2016), van der Pas and Grünwald (2014), Wagenmakers et al. (2006) for a prequential view of generalisability.

13.5 Concluding comments

Fisher information is a central statistical concept that is of considerable relevance for mathematical psychologists. We illustrated the use of Fisher information in three different statistical paradigms: in the frequentist paradigm, Fisher information was used to construct hypothesis tests and confidence intervals; in the Bayesian paradigm, Fisher information was used to specify a default prior that does not depend on how the model is parameterised; lastly, in the paradigm of information theory, data compression, and minimum description length, Fisher information was used to measure model complexity. Note that these three paradigms highlight three uses of the functional relationship f between potential observations x^n and the parameters θ . Firstly, in the frequentist setting, the second argument was fixed at a supposedly known parameter value θ_0 or $\hat{\theta}_{\text{obs}}$ resulting in a probability mass function, a function of the potential outcomes $f(\cdot | \theta_0)$. Secondly, in the Bayesian setting, the first argument was fixed at the observed data resulting in a likelihood function, a function of the parameters $f(x_{\text{obs}} | \cdot)$. Lastly, in the information geometric setting both arguments were free to vary, i.e., $f(\cdot | \cdot)$ and plugged in by the observed data and the maximum likelihood estimate.

To ease the exposition we only considered Fisher information of one-dimensional parameters. The generalisation of the concepts introduced here to vector valued θ can be found in the appendix. A complete treatment of all the uses of Fisher information throughout statistics would require a book (e.g., Frieden, 2004) rather than a tutorial. Due to the vastness of the subject, the present account is by no means comprehensive. Our goal was to use concrete examples to provide more insight about Fisher information, something that may benefit psychologists who propose, develop, and compare mathematical models for psychological processes. Other uses of Fisher information are in the detection of model misspecification (Golden, 1995; Golden, 2000; Waldorp et al., 2005; Waldorp, 2009; Waldorp et al., 2011; White, 1982), in the reconciliation of frequentist and Bayesian estimation methods through the Bernstein-von Mises theorem (Bickel and Kleijn, 2012; Rivoirard and Rousseau, 2012; van der Vaart, 1998; Yang and Le Cam, 2000), in statistical decision theory (e.g., Berger, 1985; Hájek, 1972; Korostelev and Korosteleva,

2011; Ray and Schmidt-Hieber, 2016; Wald, 1949), in the specification of objective priors for more complex models (e.g., Ghosal et al., 1997; Grazian and Robert, 2015; Kleijn and Zhao, 2017), and computational statistics and generalised MCMC sampling in particular (e.g., Banterle et al., 2015; Girolami and Calderhead, 2011; Grazian and Liseo, 2014; Gronau et al., 2017b).

In sum, Fisher information is a key concept in statistical modelling. We hope to have provided an accessible and concrete tutorial that explains the concept and some of its uses for applications that are of particular interest to mathematical psychologists.

13.A Generalisation to vector-valued parameters: The Fisher information matrix

Let X be a random variable, $\vec{\theta} = (\theta_1, \dots, \theta_d)$ a vector of parameters, and f a functional relationship that relates $\vec{\theta}$ to the potential outcomes x of X . As before, it is assumed that by fixing $\vec{\theta}$ in f we get the pmf $p_{\vec{\theta}}(x) = f(x | \vec{\theta})$, which is a function of x . The pmf $p_{\vec{\theta}}(x)$ fully determines the chances with which X takes on the events in the outcome space \mathcal{X} . The Fisher information of the vector $\vec{\theta} \in \mathbb{R}^d$ is a positive semidefinite symmetric matrix of dimension $d \times d$ with the entry at the i th row and j th column given by

$$I_X(\vec{\theta})_{i,j} = \text{Cov}\left(l(X | \vec{\theta}), l^T(X | \vec{\theta})\right)_{i,j}, \quad (13.A.1)$$

$$= \begin{cases} \sum_{x \in \mathcal{X}} \left(\frac{\partial}{\partial \theta_i} l(x | \vec{\theta}), \frac{\partial}{\partial \theta_j} l(x | \vec{\theta}) \right) p_{\vec{\theta}}(x) & \text{if } X \text{ is discrete,} \\ \int_{x \in \mathcal{X}} \left(\frac{\partial}{\partial \theta_i} l(x | \vec{\theta}), \frac{\partial}{\partial \theta_j} l(x | \vec{\theta}) \right) p_{\vec{\theta}}(x) dx & \text{if } X \text{ is continuous.} \end{cases} \quad (13.A.2)$$

where $l(x | \vec{\theta}) = \log f(x | \vec{\theta})$ is the log-likelihood function, $\frac{\partial}{\partial \theta_i} l(x | \vec{\theta})$ is the score function, that is, the partial derivative with respect to the i th component of the vector $\vec{\theta}$ and the dot is short-hand notation for the vector of the partial derivatives with respect to $\theta = (\theta_1, \dots, \theta_d)$. Thus, $l(x | \vec{\theta})$ is a $d \times 1$ column vector of score functions, while $l^T(x | \vec{\theta})$ is a $1 \times d$ row vector of score functions at the outcome x . The partial derivative is evaluated at $\vec{\theta}$, the same $\vec{\theta}$ that is used in the pmf $p_{\vec{\theta}}(x)$ for the weighting. In Appendix 13.E it is shown that the score functions are expected to be zero, which explains why $I_X(\vec{\theta})$ is a covariance matrix.

Under mild regularity conditions the i, j th entry of the Fisher information matrix can be equivalently calculated via the negative expectation of the second order partial derivates, that is,

$$I_X(\vec{\theta})_{i,j} = -E\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(X | \vec{\theta})\right), \quad (13.A.3)$$

$$= \begin{cases} -\sum_{x \in \mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x | \vec{\theta}) p_{\vec{\theta}}(x) & \text{if } X \text{ is discrete,} \\ -\int_{x \in \mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x | \vec{\theta}) p_{\vec{\theta}}(x) dx & \text{if } X \text{ is continuous.} \end{cases} \quad (13.A.4)$$

Note that the sum (thus, integral in the continuous case) is with respect to the outcomes x of X .

Example 13.A.1 (Fisher information for normally distributed random variables). *When X is normally distributed, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, it has the following probability density function (pdf)*

$$f(x | \vec{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \quad (13.A.5)$$

where the parameters are collected into the vector $\vec{\theta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$, with $\mu \in \mathbb{R}$ and $\sigma > 0$.

The score vector at a specific $\theta = (\begin{smallmatrix} \mu \\ \sigma \end{smallmatrix})$ is the following vector of functions of x

$$l(x | \vec{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \mu} l(x | \vec{\theta}) \\ \frac{\partial}{\partial \sigma} l(x | \vec{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{x-\mu}{\sigma^2} \\ \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma} \end{pmatrix}. \quad (13.A.6)$$

The unit Fisher information matrix $I_X(\vec{\theta})$ is a 2×2 symmetric positive semidefinite matrix, consisting of expectations of partial derivatives. Equivalently, $I_X(\vec{\theta})$ can be calculated using the second order partials derivatives

$$I_X(\vec{\theta}) = -E \begin{pmatrix} \frac{\partial^2}{\partial \mu \partial \mu} \log f(x | \mu, \sigma^2) & \frac{\partial^2}{\partial \mu \partial \sigma} \log f(x | \mu, \sigma) \\ \frac{\partial^2}{\partial \sigma \partial \mu} \log f(x | \mu, \sigma) & \frac{\partial^2}{\partial \sigma \partial \sigma} \log f(x | \mu, \sigma) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \quad (13.A.7)$$

The off-diagonal elements are in general not zero. If the i, j -th entry is zero we say that θ_i and θ_j are orthogonal to each other, see Appendix 13.C.3.3 below. \diamond

For iid trials $X^n = (X_1, \dots, X_n)$ with $X \sim p_\theta(x)$, the Fisher information matrix for X^n is given by $I_{X^n}(\vec{\theta}) = n I_X(\vec{\theta})$. Thus, for vector-valued parameters $\vec{\theta}$ the Fisher information matrix remains additive.

In the remainder of the text, we simply use θ for both one-dimensional and vector-valued parameters. Similarly, depending on the context it should be clear whether $I_X(\theta)$ is a number or a matrix.

13.B Frequentist statistics based on asymptotic normality

The construction of the hypothesis tests and confidence intervals in the frequentist section were all based on the MLE being asymptotically normal.

13.B.1 Asymptotic normality of the MLE for vector-valued parameters

For so-called regular parametric models, see Appendix 13.E, the MLE for vector-valued parameters θ converges in distribution to a multivariate normal distribution, that is,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}_d(0, I_X^{-1}(\theta^*)), \text{ as } n \rightarrow \infty, \quad (13.B.1)$$

where \mathcal{N}_d is a d -dimensional multivariate normal distribution, and $I_X^{-1}(\theta^*)$ the inverse Fisher information matrix at the true value θ^* . For n large enough, we can, thus, approximate the sampling distribution of the “error” of the MLE by a normal distribution, thus,

$$(\hat{\theta} - \theta^*) \approx \mathcal{N}_d\left(0, \frac{1}{n} I_X^{-1}(\theta^*)\right), \text{ we repeat, approximately.} \quad (13.B.2)$$

In practice, we fix n and replace the true sampling distribution by this normal distribution. Hence, we incur an approximation error that is only negligible whenever n is large enough. What constitutes n large enough depends on the true data

generating pmf $p^*(x)$ that is unknown in practice. In other words, the hypothesis tests and confidence intervals given in the main text based on the replacement of the true sampling distribution by this normal distribution might not be appropriate. In particular, this means that a hypothesis test at a significance level of 5% based on the asymptotic normal distribution, instead of the true sampling distribution, might actually yield a type 1 error rate of, say, 42%. Similarly, as a result of the approximation error, a 95%-confidence interval might only encapsulate the true parameter in, say, 20% of the time that we repeat the experiment.

13.B.2 Asymptotic normality of the MLE and the central limit theorem

Asymptotic normality of the MLE can be thought of as a refinement of the central limit theorem. The (Lindeberg-Lévy) CLT is a general statement about the sampling distribution of the sample mean estimator $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ based on iid trials of X with common population mean $\theta = E(X)$ and variance $\text{Var}(X) < \infty$. More specifically, the CLT states that, with a proper scaling, the sample mean \bar{X} centred around the true θ^* will converge in distribution to a normal distribution, that is, $\sqrt{n}(\bar{X} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \text{Var}(X))$. In practice, we replace the true sampling distribution by this normal distribution at fixed n and hope that n is large enough. Hence, for fixed n we then suppose that the “error” is distributed as $(\bar{X} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \frac{1}{n}\text{Var}(X))$ and we ignore the approximation error. In particular, when we know that the population variance is $\text{Var}(X) = 1$, we then know that we require an experiment with $n = 100$ samples for \bar{X} to generate estimates within 0.196 distance from θ^* with approximately 95% chance, that is, $P(|\bar{X} - \theta^*| \leq 0.196) \approx 0.95$.¹⁸ This calculation was based on our knowledge of the normal distribution $\mathcal{N}(0, 0.01)$, which has its 97.5% quantile at 0.196. In the examples below we re-use this calculation by matching the asymptotic variances to 0.01.¹⁹ The 95% statement only holds approximately, because we do not know whether $n = 100$ is large enough for the CLT to hold, i.e., this probability could be well below 23%. Note that the CLT holds under very general conditions; the population mean and variance both need to exist, i.e., be finite. The distributional form of X is irrelevant for the statement of the CLT.

On the other hand, to even compute the MLE we not only require that the population quantities to exists and be finite, but we also need to know the functional relationship f that relates these parameters to the outcomes of X . When we assume more (and nature adheres to these additional conditions), we know more, and are then able to give stronger statements. We give three examples.

Example 13.B.1 (Asymptotic normality of the MLE vs the CLT: The Gaussian distribution). *If X has a Gaussian (normal) distribution, i.e., $X \sim \mathcal{N}(\theta, \sigma^2)$, with*

¹⁸As before, chance refers to the relative frequency, that is, when we repeat the experiment $k = 200$ times, each with $n = 100$, we get k number of estimates and approximately 95% of these k number of estimates are then expected to be within 0.196 distance away from the true population mean θ^* .

¹⁹Technically, an asymptotic variance is free of n , but we mean the approximate variance at finite n . For the CLT this means $\frac{1}{n}\sigma^2$.

σ^2 known, then the MLE is the sample mean and the unit Fisher information is $I_X(\theta) = 1/\sigma^2$. Asymptotic normality of the MLE leads to the same statement as the CLT, that is, $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$. Hence, asymptotically we do not gain anything by going from the CLT to asymptotic normality of the MLE. The additional knowledge of $f(x|\theta)$ being normal does, however, allow us to come to the rare conclusion that the normal approximation holds exactly for every finite n , thus, $(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \frac{1}{n}\sigma^2)$. In all other cases, whenever $X \not\sim \mathcal{N}(\theta, \sigma^2)$, we always have an approximation.²⁰ Thus, whenever $\sigma^2 = 1$ and $n = 100$ we know that $P(|\hat{\theta} - \theta^*| \leq 0.196) = 0.95$ holds exactly. \diamond

Example 13.B.2 (Asymptotic normality of the MLE vs the CLT: The Laplace distribution). If X has a Laplace distribution with scale b , i.e., $X \sim \text{Laplace}(\theta, b)$, then its population mean and variance are $\theta = E(X)$ and $2b^2 = \text{Var}(X)$, respectively.

In this case, the MLE is the sample median \hat{M} and the unit Fisher information is $I_X(\theta) = 1/b^2$. Asymptotic normality of the MLE implies that we can approximate the sampling distribution by the normal distribution, that is, $(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \frac{1}{n}b^2)$, when n is large enough. Given that the population variance is $\text{Var}(X) = 1$, we know that $b = 1/\sqrt{2}$, yielding a variance of $\frac{1}{2n}$ in our normal approximation to the sampling distribution. Matching this variance to 0.01 shows that we now require only $n = 50$ samples for the estimator to generate estimates within 0.196 distance away from the true value θ^* with 95% chance. As before, the validity of this statement only holds approximately, i.e., whenever the normal approximation to the sampling distribution of the MLE at $n = 50$ is not too bad.

Hence, the additional knowledge of $f(x|\theta)$ being Laplace allows us to use an estimator, i.e., the MLE, that has a lower asymptotic variance. Exploiting this knowledge allowed us to design an experiment with twice as few participants. \diamond

Example 13.B.3 (Asymptotic normality of the MLE vs the CLT: The Cauchy distribution). If X has a Cauchy distribution centred around θ with scale 1, i.e., $X \sim \text{Cauchy}(\theta, 1)$, then X does not have a finite population variance, nor a finite population mean. As such, the CLT cannot be used. Even worse, Fisher (1922) showed that the sample mean as an estimator for θ is in this case useless, as the sampling distribution of the sample mean is a Cauchy distribution that does not depend on n , namely, $\bar{X} \sim \text{Cauchy}(\theta, 1)$. As such, using the first observation alone to estimate θ is as good as combining the information of $n = 100$ samples in the sample mean estimator. Hence, after seeing the first observation no additional information about θ is gained using the sample mean \bar{X} , not even if we increase n .

The sample median estimator \hat{M} performs better. Again, Fisher (1922) already knew that for n large enough that $(\hat{M} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \frac{1}{n}\frac{\pi^2}{2})$. The MLE is even better, but unfortunately, in this case, it cannot be given as an explicit function of the

²⁰This is a direct result of Cramér's theorem that states that whenever X is independent of Y and $Z = X + Y$ with Z a normal distribution, then X and Y themselves are necessarily normally distributed.

*data.*²¹ The Fisher information can be given explicitly, namely, $I_X(\theta) = 1/2$. Asymptotic normality of the MLE implies that $(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, \frac{1}{n}2)$, when n is large enough. Matching the variances in the approximation based on the normal distribution to 0.01 shows that we require $n = 25\pi^2 \approx 247$ for the sample median and $n = 200$ samples for the MLE to generate estimates within 0.196 distance away from the true value of value θ^* with approximate 95% chance. ◇

13.B.3 Efficiency of the MLE: The Hájek-LeCam convolution theorem and the Cramér-Fréchet-Rao information lower bound

The previous examples showed that the MLE is an estimator that leads to a smaller sample size requirement, because it is the estimator with the lower asymptotic variance. This lower asymptotic variance is a result of the MLE making explicit use of the functional relationship between the samples x_{obs}^n and the target θ in the population. Given any such f , one might wonder whether the MLE is the estimator with the *lowest possible* asymptotic variance. The answer is affirmative, whenever we restrict ourselves to the broad class of so-called regular estimators.

A *regular estimator* $T_n = t_n(X_n)$ is a function of the data that has a limiting distribution that does not change too much, whenever we change the parameters in the neighbourhood of the true value θ^* , see van der Vaart (1998, p. 115) for a precise definition. The Hájek-LeCam convolution theorem characterises the aforementioned limiting distribution as a convolution, i.e., a sum of the independent statistics Δ_{θ^*} and Z_{θ^*} . That is, for any regular estimator T_n and every possible true value θ^* we have

$$\sqrt{n}(T_n - \theta^*) \xrightarrow{D} \Delta_{\theta^*} + Z_{\theta^*}, \text{ as } n \rightarrow \infty, \quad (13.B.3)$$

where $Z_{\theta^*} \sim \mathcal{N}(0, I_X^{-1}(\theta^*))$ and where Δ_{θ^*} has an arbitrary distribution. By independence, the variance of the asymptotic distribution is simply the sum of the variances. As the variance of Δ_{θ^*} cannot be negative, we know that the asymptotic variance of any regular estimator T_n is bounded from below, that is, $\text{Var}(\Delta_{\theta^*}) + I_X^{-1}(\theta^*) \geq I_X^{-1}(\theta^*)$.

The MLE is a regular estimator with Δ_{θ^*} equal to the fixed true value θ^* , thus, $\text{Var}(\Delta_{\theta^*}) = 0$. As such, the MLE has an asymptotic variance $I_X^{-1}(\theta^*)$ that is equal to the lower bound given above. Hence, amongst the broad class of regular estimators, the MLE performs best. This result was already foreshadowed by Fisher (1922), though it took another 50 years before this statement was made mathematically rigorous (Hájek, 1970; Inagaki, 1970; LeCam, 1970; van der Vaart, 2002; Yang, 1999), see also Ghosh (1985) for a beautiful review.

We stress that the normal approximation to the true sampling distribution only holds when n is large enough. In practice, n is relatively small and the replacement of the true sampling distribution by the normal approximation can,

²¹Given observations x_{obs}^n the maximum likelihood estimate $\hat{\theta}_{\text{obs}}$ is the number for which the score function $\dot{l}(x_{\text{obs}}^n | \theta) = \sum_{i=1}^n \frac{2(x_{\text{obs},i} - \theta)}{1 + (x_{\text{obs},i} - \theta)^2}$ is zero. This optimisation cannot be solved analytically and there are $2n$ solutions to this equation.

thus, lead to confidence intervals and hypothesis tests that perform poorly (Brown et al., 2001). This can be very detrimental, especially, when we are dealing with hard decisions such as the rejection or non-rejection of a hypothesis.

A simpler version of the Hájek-LeCam convolution theorem is known as the Cramér-Fréchet-Rao information lower bound (Cramér, 1946; Fréchet, 1943; Rao, 1945), which also holds for finite n . This theorem states that the variance of an unbiased estimator T_n cannot be lower than the inverse Fisher information, that is, $n\text{Var}(T_n) \geq I_X^{-1}(\theta^*)$. We call an estimator $T_n = t(X^n)$ *unbiased* if for every possible true value θ^* and at each fixed n , its expectation is equal to the true value, that is, $E(T_n) = \theta^*$. Hence, this lower bound shows that Fisher information is not only a concept that is useful for large samples.

Unfortunately, the class of unbiased estimators is rather restrictive (in general, it does not include the MLE) and the lower bound cannot be attained whenever the parameter is of more than one dimensions (Wijsman, 1973). Consequently, for vector-valued parameters θ , this information lower bound does not inform us, whether we should stop our search for a better estimator.

The Hájek-LeCam convolution theorem implies that for n large enough the MLE $\hat{\theta}$ is the best performing statistic. For the MLE to be superior, however, the data do need to be generated as specified by the functional relationship f . In reality, we do not know whether the data are indeed generated as specified by f , which is why we should also try to empirically test such an assumption. For instance, we might believe that the data are normally distributed, while in fact they were generated according to a Cauchy distribution. This incorrect assumption implies that we should use the sample mean, but Example 13.B.3 showed the futility of such estimator. Model misspecification, in addition to hard decisions based on the normal approximation, might be the main culprit of the crisis of replicability. Hence, more research on the detection of model misspecification is desirable and expected (e.g., Grünwald, 2016; Grünwald and van Ommen, 2014; van Ommen et al., 2016).

13.C Bayesian use of the Fisher-Rao metric: The Jeffreys's prior

We make intuitive that the Jeffreys's prior is a uniform prior on the model \mathcal{M}_Θ , i.e.,

$$P(m^* \in J_m) = \frac{1}{V} \int_{J_m} 1 dm_\theta(X) = \int_{\theta_a}^{\theta_b} \sqrt{I_X(\theta)} d\theta, \quad (13.C.1)$$

where $J_m = (m_{\theta_a}(X), m_{\theta_b}(X))$ is an interval of pmfs in model space. To do so, we explain why the differential $dm_\theta(X)$, a displacement in model space, is converted into $\sqrt{I_X(\theta)} d\theta$ in parameter space. The elaboration below boils down to an explanation of arc length computations using integration by substitution.

13.C.1 Tangent vectors

First note that we swapped the area of integration by substituting the interval $J_m = (m_{\theta_a}(X), m_{\theta_b}(X))$ consisting of pmfs in function space \mathcal{M}_Θ by the interval (θ_a, θ_b) in parameter space. This is made possible by the parameter functional ν with domain \mathcal{M}_Θ and range Θ that uniquely assigns to any (transformed) pmf $m_a(X) \in \mathcal{M}_\Theta$ a parameter value $\theta_a \in \Theta$. In this case, we have $\theta_a = \nu(m_a(X)) = (\frac{1}{2}m_a(1))^2$. Uniqueness of the assignment implies that the resulting parameter values θ_a and θ_b in Θ differ from each other whenever $m_a(X)$ and $m_b(X)$ in \mathcal{M}_Θ differ from each other. For example, the map $\nu : \mathcal{M}_\Theta \rightarrow \Theta$ implies that

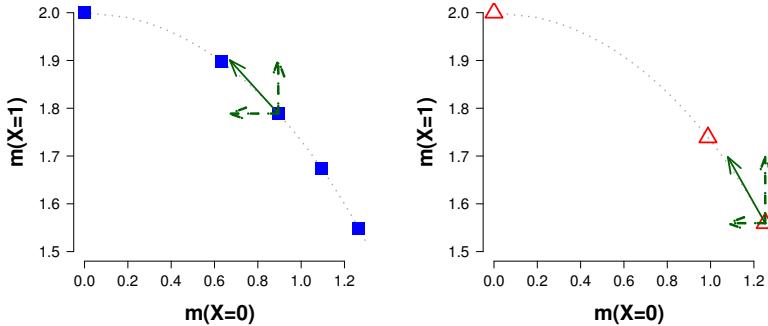


Figure 13.12: The full arrow represents the simultaneous displacement in model space based on the Taylor approximation Eq. (13.C.3) in terms of θ at $m_{\theta_a}(X)$, where $\theta_a = 0.8$ (left panel) and in terms of ϕ at $m_{\phi_a}(X)$ where $\phi_a = 0.6\pi$ (right panel). The dotted line represents a part of the Bernoulli model and note that the full arrow is tangent to the model.

in the left panel of Fig. 13.12 the third square from the left with coordinates $m_a(X) = [0.89, 1.79]$ can be labelled by $\theta_a = 0.8 \approx (\frac{1}{2}(1.79))^2$, while the second square from the left with coordinates $m_b(X) = [0.63, 1.90]$ can be labelled by $\theta_b = 0.9 \approx (\frac{1}{2}(1.90))^2$.

To calculate the arc length of the curve J_m consisting of functions in \mathcal{M}_Θ , we first approximate J_m by a finite sum of tangent vectors, i.e., straight lines. The approximation of the arc length is the sum of the length of these straight lines. The associated approximation error goes to zero, when we increase the number of tangent vectors and change the sum into an integral sign, as in the usual definition of an integral. First we discuss tangent vectors.

In the left panel in Fig. 13.12, we depicted the tangent vector at $m_{\theta_a}(X)$ as the full arrow. This full arrow is constructed from its components: one broken arrow that is parallel to the horizontal axis associated with the outcome $x = 0$, and one broken arrow that is parallel to the vertical axis associated with the outcome $x = 1$. The arrows parallel to the axes are derived by first fixing $X = x$ followed by a Taylor expansion of the parameterisation $\theta \mapsto m_\theta(x)$ at θ_a . The Taylor expansion is derived by differentiating with respect to θ at θ_a yielding the

following “linear” function of the distance $d\theta = |\theta_b - \theta_a|$ in parameter space,

$$dm_{\theta_a}(x) = m_{\theta_b}(x) - m_{\theta_a}(x) = \underbrace{\frac{dm_{\theta_a}(x)}{d\theta}}_{A_{\theta_a}(x)} d\theta + \underbrace{o(d\theta)}_{B_{\theta_a}(x)}, \quad (13.C.2)$$

where the slope, a function of x , $A_{\theta_a}(x)$ at $m_{\theta_a}(x)$ in the direction of x is given by

$$A_{\theta_a}(x) = \frac{dm_{\theta_a}(x)}{d\theta} = \frac{1}{2} \underbrace{\left\{ \frac{d}{d\theta} \log f(x | \theta_a) \right\}}_{\text{score function}} m_{\theta_a}(x), \quad (13.C.3)$$

and with an “intercept” $B_{\theta_a}(x) = o(d\theta)$ that goes fast to zero whenever $d\theta \rightarrow 0$. Thus, for $d\theta$ small, the intercept $B_{\theta_a}(x)$ is practically zero. Hence, we approximate the displacement between $m_{\theta_a}(x)$ and $m_{\theta_b}(x)$ by a straight line.

Example 13.C.1 (Tangent vectors). *In the right panel of Fig. 13.12 the right most triangle is given by $m_{\phi_a}(X) = [1.25, 1.56]$, while the triangle in the middle refers to $m_{\phi_b}(X) = [0.99, 1.74]$. Using the functional $\tilde{\nu}$, i.e., the inverse of the parameterisation, $\phi \mapsto 2\sqrt{f(x|\phi)}$, where $f(x|\phi) = (\frac{1}{2} + \frac{1}{2}(\frac{\phi}{\pi})^3)^x (\frac{1}{2} - \frac{1}{2}(\frac{\phi}{\pi})^3)^{1-x}$, we find that these two pmfs correspond to $\phi_a = 0.6\pi$ and $\phi_b = 0.8\pi$.*

The tangent vector at $m_{\phi_a}(X)$ is constructed from its components. For the horizontal displacement, we fill in $x = 0$ in $\log f(x|\phi)$ followed by the derivation with respect to ϕ at ϕ_a and a multiplication by $m_{\phi_a}(x)$ resulting in

$$\frac{dm_{\phi_a}(0)}{d\phi} d\phi = \frac{1}{2} \left\{ \frac{d}{d\phi} \log f(0 | \phi_a) \right\} m_{\phi_a}(0) d\phi, \quad (13.C.4)$$

$$= - \frac{3\phi_a^2}{\sqrt{2\pi^3(\pi^3 + \phi_a^3)}} d\phi, \quad (13.C.5)$$

where $d\phi = |\phi_b - \phi_a|$ is the distance in parameter space Φ . The minus sign indicates that the displacement along the horizontal axis is from right to left. Filling in $d\phi = |\phi_b - \phi_a| = 0.2\pi$ and $\phi_a = 0.6\pi$ yields a horizontal displacement of 0.17 at $m_{\phi_a}(0)$ from right to left in model space. Similarly, the vertical displacement in terms of ϕ is calculated by first filling in $x = 1$ and leads to

$$\frac{dm_{\phi_a}(1)}{d\phi} d\phi = \frac{1}{2} \left\{ \frac{d}{d\phi} \log f(1 | \phi_a) \right\} m_{\phi_a}(1) d\phi, \quad (13.C.6)$$

$$= - \frac{3\phi_a^2}{\sqrt{2\pi^3(\pi^3 - \phi_a^3)}} d\phi. \quad (13.C.7)$$

By filling in $d\phi = 0.2$ and $\phi_a = 0.6\pi$, we see that a change of $d\phi = 0.2\pi$ at $\phi_a = 0.6\pi$ in the parameter space corresponds to a vertical displacement of 0.14 at $m_{\phi_a}(1)$ from bottom to top in model space. Note that the axes in Fig. 13.12 are scaled differently.

The combined displacement $\frac{dm_{\phi_a}(X)}{d\phi} d\phi$ at $m_{\phi_a}(X)$ is the sum of the two broken arrows and plotted as a full arrow in the right panel of Fig. 13.12. ◇

The length of the tangent vector $\frac{dm_{\theta_a}(X)}{d\theta}$ at the vector $m_{\theta_a}(X)$ is calculated by taking the root of the sum of its squared component, the natural measure of distance we adopted above and this yields

$$\left\| \frac{dm_{\theta_a}(X)}{d\theta} d\theta \right\|_2 = \sqrt{\sum_{x \in \mathcal{X}} \left(\frac{dm_{\theta_a}(x)}{d\theta} \right)^2 (d\theta)^2}, \quad (13.C.8)$$

$$= \sqrt{\sum_{x \in \mathcal{X}} \left(\frac{d}{d\theta} \log f(x | \theta_a) \right)^2 p_{\theta_a}(x) d\theta} = \sqrt{I_X(\theta_a)} d\theta. \quad (13.C.9)$$

The second equality follows from the definition of $\frac{dm_{\theta_a}(X)}{d\theta}$, i.e., Eq. (13.C.3), and the last equality is due to the definition of Fisher information.

Example 13.C.2 (Length of the tangent vectors). *The length of the tangent vector in the right panel of Fig. 13.12 can be calculated as the root of the sums of squares of its components, that is, $\left\| \frac{dm_{\phi_a}(X)}{d\phi} d\phi \right\|_2 = \sqrt{(-0.14)^2 + 0.17^2} = 0.22$. Alternatively, we can first calculate the square root of the Fisher information at $\phi_a = 0.6\pi$, i.e.,*

$$\sqrt{I(\phi_a)} = \frac{3\phi_a^2}{\sqrt{\pi^6 - \phi_a^6}} = 0.35, \quad (13.C.10)$$

and a multiplication by $d\phi = 0.2\pi$ results in $\left\| \frac{dm_{\phi_a}(X)}{d\phi} \right\|_2 d\phi = 0.22$. \diamond

More generally, to approximate the length between pmfs $m_{\theta_a}(X)$ and $m_{\theta_b}(X)$, we first identify $\nu(m_{\theta_a}(X)) = \theta_a$ and multiply this with the distance $d\theta = |\theta_a - \nu(m_{\theta_b}(X))|$ in parameter space, i.e.,

$$dm_\theta(X) = \left\| \frac{dm_\theta(X)}{d\theta} \right\|_2 d\theta = \sqrt{I_X(\theta)} d\theta. \quad (13.C.11)$$

In other words, the root of the Fisher information converts a small distance $d\theta$ at θ_a to a displacement in model space at $m_{\theta_a}(X)$.

13.C.2 The Fisher-Rao metric

By virtue of the parameter functional ν , we send an interval of pmfs $J_m = (m_{\theta_a}(X), m_{\theta_b}(X))$ in the function space \mathcal{M}_Θ to the interval (θ_a, θ_b) in the parameter space Θ . In addition, with the conversion of $dm_\theta(X) = \sqrt{I_X(\theta)} d\theta$ we integrate by substitution, that is,

$$P(m^*(X) \in J_m) = \frac{1}{V} \int_{m_{\theta_a}(X)}^{m_{\theta_b}(X)} 1 dm_\theta(X) = \frac{1}{V} \int_{\theta_a}^{\theta_b} \sqrt{I_X(\theta)} d\theta. \quad (13.C.12)$$

In particular, choosing $J_\theta = \mathcal{M}_\Theta$ yields the normalisation constant $V = \int_0^1 \sqrt{I_X(\theta)} d\theta$. The interpretation of V as being the total length of \mathcal{M}_Θ is due to the use of $dm_\theta(X)$ as the metric, a measure of distance, in model space. To

honour Calyampudi Radhakrishna Rao's (1945) contribution to the theory, this metric is also known as the Fisher-Rao metric (e.g., Amari et al., 1987; Atkinson and Mitchell, 1981; Burbea, 1984; Burbea and Rao, 1982, 1984; Dawid, 1977; Efron, 1975; Kass and Vos, 2011).

13.C.3 Fisher-Rao metric for vector-valued parameters

13.C.3.1 The parameter functional $\nu : \mathcal{P} \rightarrow B$ and the categorical distribution

For random variables with w number of outcomes, the largest set of pmfs \mathcal{P} is the collection of functions p on \mathcal{X} such that (i) $0 \leq p(x) = P(X = x)$ for every outcome x in \mathcal{X} , and (ii) to explicitly convey that there are w outcomes, and none more, these w chances have to sum to one, that is, $\sum_{x \in \mathcal{X}} p(x) = 1$. The complete set of pmfs \mathcal{P} can be parameterised using the functional ν that assigns to each w -dimensional pmf $p(X)$ a parameter $\beta \in \mathbb{R}^{w-1}$.

For instance, given a pmf $p(X) = [p(L), p(M), p(R)]$ we typically use the functional $\nu : \mathcal{P} \rightarrow \mathbb{R}^2$ that takes the first two coordinates, that is, $\nu(p(X)) = \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$, where $\beta_1 = p(L)$ and $\beta_2 = p(M)$. The range of this functional ν is the parameter space $B = [0, 1] \times [0, \beta_1]$. Conversely, the inverse of the functional ν is the parameterisation $\beta \mapsto p_\beta(X) = [\beta_1, \beta_2, 1 - \beta_1 - \beta_2]$, where (i') $0 \leq \beta_1, \beta_2$ and (ii') $\beta_1 + \beta_2 \leq 1$. The restrictions (i') and (ii') imply that the parameterisation has domain B and the largest set of pmfs \mathcal{P} as its range. By virtue of the functional ν and its inverse, that is, the parameterisation $\beta \mapsto p_\beta(X)$, we conclude that the parameter space B and the complete set of pmfs \mathcal{P} are isomorphic. This means that each pmf $p(X) \in \mathcal{P}$ can be uniquely identified with a parameter $\beta \in B$ and vice versa. The inverse of ν implies that the parameters $\beta \in B$ are functionally related to the potential outcomes x of X as

$$f(x | \beta) = \beta_1^{x_L} \beta_2^{x_M} (1 - \beta_1 - \beta_2)^{x_R}, \quad (13.C.13)$$

where x_L, x_M and x_R are the number of L, M and R responses in one trial – we either have $x = [x_L, x_M, x_R] = [1, 0, 0]$, $x = [0, 1, 0]$, or $x = [0, 0, 1]$. The model $f(x | \beta)$ can be regarded as the generalisation of the Bernoulli model to $w = 3$ categories. In effect, the parameters β_1 and β_2 can be interpreted as a participant's propensity of choosing L and M , respectively. If X^n consists of n iid categorical random variables with the outcomes $[L, M, R]$, the joint pmf of X^n is then

$$f(x^n | \beta) = \beta_1^{y_L} \beta_2^{y_M} (1 - \beta_1 - \beta_2)^{y_R}, \quad (13.C.14)$$

where y_L, y_M and $y_R = n - y_L - y_M$ are the number of L, M and R responses in n trials. As before, the representation of the pmfs as the vectors $m_\beta(X) = [2\sqrt{\beta_1}, 2\sqrt{\beta_2}, 2\sqrt{1 - \beta_1 - \beta_2}]$ form the surface of (the positive part of) the sphere of radius two, thus, $\mathcal{M} = \mathcal{M}_B$, see Fig. 13.13. The extreme pmfs indicated by mL, mM and mR in the figure are indexed by the parameter values $\beta = (1, 0)$, $\beta = (0, 1)$ and $\beta = (0, 0)$, respectively.

13.C.3.2 The stick-breaking parameterisation of the categorical distribution

Alternatively, we could also have used a “stick-breaking” parameter functional $\tilde{\nu}$ that sends each pmf in \mathcal{P} to the vector of parameters $\tilde{\nu}(p(X)) = \binom{\gamma_1}{\gamma_2}$, where $\gamma_1 = p_L$ and $\gamma_2 = p_M/(1-p_L)$.²² Again the parameter $\gamma = \binom{\gamma_1}{\gamma_2}$ is only a label, but this time the range of $\tilde{\nu}$ is the parameter space $\Gamma = [0, 1] \times [0, 1]$. The functional relationship f associated to γ is given by

$$f(x | \gamma) = \gamma_1^{x_L} ((1 - \gamma_1)\gamma_2)^{x_M} ((1 - \gamma_1)(1 - \gamma_2))^{x_R}. \quad (13.C.15)$$

For each γ we can transform the pmf into the vector

$$m_\gamma(X) = [2\sqrt{\gamma_1}, 2\sqrt{(1 - \gamma_1)\gamma_2}, 2\sqrt{(1 - \gamma_1)(1 - \gamma_2)}], \quad (13.C.16)$$

and write \mathcal{M}_Γ for the collection of vectors so defined. As before, this collection coincides with the full model, i.e., $\mathcal{M}_\Gamma = \mathcal{M}$. In other words, by virtue of the functional $\tilde{\nu}$ and its inverse $\gamma \mapsto p_\gamma(x) = f(x | \gamma)$ we conclude that the parameter space Γ and the complete set of pmfs \mathcal{M} are isomorphic. Because $\mathcal{M} = \mathcal{M}_B$ this means that we also have an isomorphism between the parameter space B and Γ via \mathcal{M} , even though B is a strict subset of Γ . Note that this equivalence goes via parameterisation $\beta \mapsto m_\beta(X)$ and the functional $\tilde{\nu}$.

13.C.3.3 Multidimensional Jeffreys's prior via the Fisher information matrix and orthogonal parameters

The multidimensional Jeffreys's prior is parameterisation-invariant and has as normalisation constant $V = \int \sqrt{\det I_X(\theta)} d\theta$, where $\det I_X(\theta)$ is the determinant of the Fisher information matrix.

In the previous subsection we argued that the categorical distribution in terms of β or parameterised with γ are equivalent to each other, that is, $\mathcal{M}_B = \mathcal{M} = \mathcal{M}_\Gamma$. However, these two parameterisations describe the model space \mathcal{M} quite differently. In this subsection we use the Fisher information to show that the parameterisation in terms of γ is sometimes preferred over β .

The complete model \mathcal{M} is easier described by γ , because the parameters are orthogonal. We say that two parameters are *orthogonal to each other* whenever the corresponding off-diagonal entries in the Fisher information matrix are zero. The Fisher information matrices in terms of β and γ are

$$I_X(\beta) = \frac{1}{1 - \beta_1 - \beta_2} \begin{pmatrix} 1 - \beta_2 & 1 \\ 1 & 1 - \beta_1 \end{pmatrix} \text{ and } I_X(\gamma) = \begin{pmatrix} \frac{1}{\gamma_1(1-\gamma_1)} & 0 \\ 0 & \frac{1-\gamma_1}{\gamma_2(1-\gamma_2)} \end{pmatrix},$$

respectively. The left panel of Fig. 13.13 shows the tangent vectors at $p_{\beta^*}(X) = [1/3, 1/3, 1/3]$ in model space, where $\beta^* = (1/3, 1/3)$. The green tangent vector corresponds to $\frac{\partial m_{\beta^*}(X)}{\partial \beta_1}$, thus, with $\beta_2 = 1/3$ fixed and β_1 free to vary, while the red tangent vector corresponds to $\frac{\partial m_{\beta^*}(X)}{\partial \beta_2}$, thus, with $\beta_1 = 1/3$ and β_2

²²This only works if $p_L < 1$. When $p(x_1) = 1$, we simply set $\gamma_2 = 0$, thus, $\gamma = (1, 0)$.

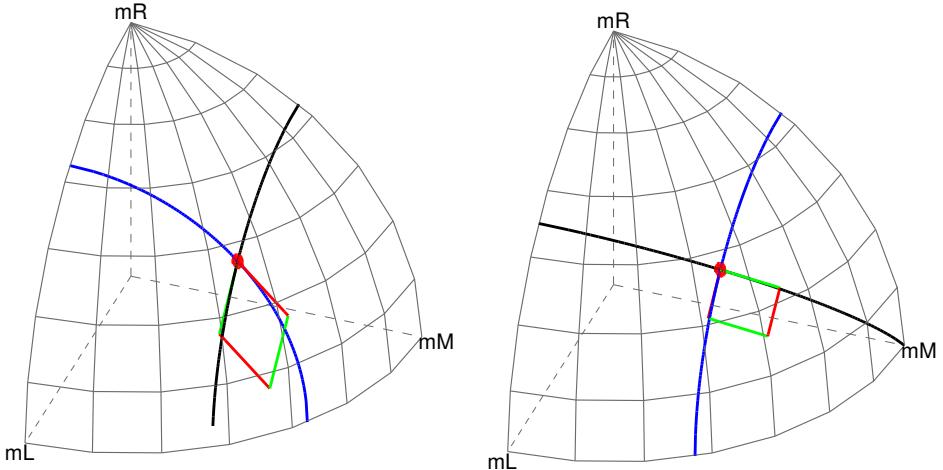


Figure 13.13: When the off-diagonal entries are zero, the tangent vectors are orthogonal. Left panel: The tangent vectors at $p_{\beta^*}(X) = [1/3, 1/3, 1/3]$ span a diamond with an area given by $\sqrt{\det I(\beta^*)}d\beta$. The black curve is the submodel with $\beta_2 = 1/3$ fixed and β_1 free to vary and yields a green tangent vector. The blue curve is the submodel with $\beta_1 = 1/3$ fixed and β_2 free to vary. Right panel: The tangent vectors at the same pmf in terms of γ , thus, $p_{\gamma^*}(X)$, span a rectangle with an area given by $\sqrt{\det I(\gamma^*)}d\gamma$. The black curve is the submodel with $\gamma_2 = 1/2$ fixed and γ_1 free to vary and yields a green tangent vector. The blue curve is the submodel with $\gamma_1 = 1/3$ fixed and γ_2 free to vary.

free to vary. The area of the diamond spanned by these two tangent vectors is $\sqrt{\det I(\beta^*)}d\beta_1 d\beta_2$, where we have taken $d\beta_1 = 0.1$ and $d\beta_2 = 0.1$.

The right panel of Fig. 13.13 shows the tangent vectors at the same point $p_{\gamma^*}(X) = [1/3, 1/3, 1/3]$, where $\gamma^* = (1/3, 1/2)$. The green tangent vector corresponds to $\frac{\partial m_{\gamma^*}(X)}{\partial \gamma_1}$, thus, with $\gamma_2 = 1/2$ fixed and γ_1 free to vary, while the red tangent vector corresponds to $\frac{\partial m_{\gamma^*}(X)}{\partial \gamma_2}$, thus, with $\gamma_1 = 1/3$ and γ_2 free to vary. By glancing over the plots, we see that the two tangent vectors are indeed orthogonal. The area of the rectangle spanned by these these two tangent vectors is $\sqrt{\det I(\gamma^*)}d\gamma_1 d\gamma_2$, where we have taken $d\gamma_1 = d\gamma_2 = 0.1$.

There are now two ways to calculate the normalisation constant of the Jeffreys's prior, the area, more generally volume, of the model \mathcal{M} . In terms of β this leads to

$$V = \int_0^1 \left(\int_0^{\beta_1} \frac{1}{1 - \beta_1 - \beta_2} \sqrt{\beta_1 \beta_2 - \beta_1 - \beta_2} d\beta_2 \right) d\beta_1. \quad (13.C.17)$$

Observe that the inner integral depends on the value of β_1 from the outer integral. This coupling is reflected by the non-zero off-diagonal term of the Fisher informa-

tion matrix $I_X(\beta)$ corresponding to β_1 and β_2 . On the other hand, orthogonality implies that the two parameters can be treated independently of each other. That is, knowing and fixing γ_1 and changing γ_2 will not affect $m_\gamma(X)$ via γ_1 . This means that the double integral decouples

$$V = \int_0^1 \left(\int_0^1 \frac{1}{\sqrt{\gamma_1 \gamma_2 (1 - \gamma_2)}} d\gamma_1 \right) d\gamma_2 = \int_0^1 \frac{1}{\sqrt{\gamma_1}} d\gamma_1 \int_0^1 \frac{1}{\sqrt{\gamma_2 (1 - \gamma_2)}} d\gamma_2 = 2\pi. \quad (13.C.18)$$

Using standard geometry we verify that this is indeed the area of \mathcal{M} , as an eighth of the surface area of a sphere of radius two is given by $\frac{1}{8}4\pi 2^2 = 2\pi$.

Orthogonality is relevant in Bayesian analysis, as it provides an argument to choose a prior on a vector-valued parameter that factorises (e.g., Berger et al., 1998; Huzurbazar, 1950, 1956; Jeffreys, 1961; Kass and Vaidyanathan, 1992; Ly et al., 2016a, 2016b), see also Cox and Reid (1987); Mitchell (1962).

By taking a random variable X with $w = 3$ outcomes, we were able to visualise the geometry of model space. For more general X these plots get more complicated and perhaps even impossible to draw. Nonetheless, the ideas conveyed here extend, even to continuous X , whenever the model adheres to the regularity conditions given in Appendix 13.E.

13.D MDL: Coding theoretical background

13.D.1 Coding theory, code length and log-loss

A coding system translates words, i.e., outcomes of a random variable X , into code words with code lengths that behave like a pmf. Code lengths can be measured with a logarithm, which motivates the adoption of log-loss, defined below, as the decision criterion within the MDL paradigm. The coding theoretical terminologies introduced here are illustrated using the random variable X with $w = 3$ potential outcomes.

13.D.1.1 Kraft-McMillan inequality: From code lengths of a specific coding system to a pmf

For the source-memory task we encoded the outcomes as L , M and R , but when we communicate a participant's responses x_{obs}^n to a collaborator over the internet, we have to encode the observations x_{obs}^n as zeroes and ones. For instance, we might use a coding system \tilde{C} with code words $\tilde{C}(X = L) = 00$, $\tilde{C}(X = M) = 01$ and $\tilde{C}(X = R) = 10$. This coding system \tilde{C} will transform any set of responses x_{obs}^n into a code string $\tilde{C}(x_{\text{obs}}^n)$ consisting of $2n$ bits. Alternatively, we can use a coding system C with code words $C(X = L) = 10$, $C(X = M) = 0$ and $C(X = R) = 11$, instead. Depending on the actual observations x_{obs}^n , this coding system outputs code strings $C(x_{\text{obs}}^n)$ with varying code lengths that range from n to $2n$ bits. For example, if a participant responded with $x_{\text{obs}}^n = (M, R, M, L, L, M, M, M)$ in $n = 8$ trials, the coding system C would then output the 11-bit long code string $C(x_{\text{obs}}^n) = 01101010000$. In contrast, the first coding system \tilde{C} will always output

a 16-bit long code string when $n = 8$. Shorter code strings are desirable as they will lead to a smaller load on the communication network and they are less likely to be intercepted by “competing” researchers.

Note that the shorter code length $C(x_{\text{obs}}^n) = 01101010000$ of 11-bits is a result of having code words of unequal lengths. The fact that one of the code word is shorter does not interfere with the decoding, since no code word is a prefix of another code word. As such, we refer to C as a prefix (free) coding system. This implies that the 11-bit long code string $C(x_{\text{obs}}^n)$ is self-punctuated and that it can be uniquely deciphered by simply reading the code string from left to right resulting in the retrieval of x_{obs}^n . Note that the code lengths of C inherit the randomness of the data. In particular, the coding system C produces a shorter code string with high chance, if the participant generates the outcome M with high chance. In the extreme case, the coding system C produces the 8-bits long code string $C(x^n) = 00000000$ with 100% (respectively, 0%) chance, if the participant generates the outcome M with 100% (respectively, 0%) chance. More generally, Kraft and McMillan (Kraft, 1949; McMillan, 1956) showed that *any* uniquely decipherable (prefix) coding system from the outcome space \mathcal{X} with w outcomes to an alphabet with D elements must satisfy the inequality

$$\sum_{i=1}^w D^{-l_i} \leq 1, \quad (13.D.1)$$

where l_i is the code length of the outcome w . In our example, we have taken $D = 2$ and code length of 2, 1 and 2 bits for the response L, M and R respectively. Indeed, $2^{-2} + 2^{-1} + 2^{-2} = 1$. Hence, code lengths behave like the logarithm (with base D) of a pmf.

13.D.1.2 Shannon-Fano algorithm: From a pmf to a coding system with specific code lengths

Given a data generating pmf $p^*(X)$, we can use the so-called *Shannon-Fano algorithm* (e.g., Cover and Thomas, 2006, Ch. 5) to construct a prefix coding system C^* . The idea behind this algorithm is to give the outcome x that is generated with the highest chance the shortest code length. To do so, we encode the outcome x as a code word $C^*(x)$ that consists of $-\log_2 p^*(x)$ bits.²³

For instance, when a participant generates the outcomes $[L, M, R]$ according to the chances $p^*(X) = [0.25, 0.5, 0.25]$ the Shannon-Fano algorithm prescribes that we should encode the outcome L with $-\log_2(0.25) = 2$, M with $-\log_2(0.5) = 1$ and R with 2 bits; the coding system C given above.²⁴ The Shannon-Fano algorithm works similarly for any other given pmf $p_\beta(X)$. Hence, the Kraft-McMillan inequality and its inverse, i.e., the Shannon-Fano algorithm imply that pmfs and

²³When we use the logarithm with base two, $\log_2(y)$, we get the code length in bits, while the natural logarithm, $\log(y)$, yields the code length in nats. Any result in terms of the natural logarithm can be equivalently described in terms of the logarithm with base two, as $\log(y) = \log(2) \log_2(y)$.

²⁴Due to rounding, the Shannon-Fano algorithm actually produces code words $C(x)$ that are at most one bit larger than the ideal code length $-\log_2 p^*(x)$. We avoid further discussions on rounding. Moreover, in the following we consider the natural logarithm instead.

uniquely decipherable coding systems are equivalent to each other. As such we have an additional interpretation of a pmf. To distinguish the different uses, we write $f(X | \beta)$ when we view the pmf as a coding system, while we retain the notation $p_\beta(X)$ when we view the pmf as a data generating device. In the remainder of this section we will not explicitly construct any other coding system, as the coding system itself is irrelevant for the discussion at hand –only the code lengths matter.

13.D.1.3 Entropy, cross entropy, log-loss

With the true data generating pmf $p^*(X)$ at hand, thus, also the true coding system $f(X | \beta^*)$, we can calculate the (population) average code length per trial

$$H(p^*(X)) = H\left(p^*(X) \| f(X | \beta^*)\right) = \sum_{x \in \mathcal{X}} -\log f(x | \beta^*) p^*(x). \quad (13.D.2)$$

Whenever we use the logarithm with base 2, we refer to this quantity $H(p^*(X))$ as the Shannon entropy.²⁵ If the true pmf is $p^*(X) = [0.25, 0.5, 0.25]$ we have an average code length of 1.5 bits per trial whenever we use the true coding system $f(X | \beta^*)$. Thus, we expect to use 12 bits to encode observations consisting of $n = 8$ trials.

As coding theorists, we have no control over the true data generating pmf $p^*(X)$, but we can choose the coding system $f(X | \beta)$ to encode the observations. The (population) average code length per trial is given by

$$H(p^*(X) \| \beta) = H\left(p^*(X) \| f(X | \beta)\right) = \sum_{x \in \mathcal{X}} -\log f(x | \beta) p^*(x). \quad (13.D.3)$$

The quantity $H(p^*(X) \| \beta)$ is also known as the cross entropy from the true pmf $p^*(X)$ to the postulated $f(X | \beta)$.²⁶ For instance, if the pmf $f(X | \beta) = [0.01, 0.18, 0.81]$ is used to encode data that are generated according to $p^*(X) = [0.25, 0.5, 0.25]$, we will use 2.97 bits on average per trial. Clearly, this is much more than the 1.5 bits per trial that we get from using the true coding system $f(X | \beta^*)$.

More generally, Shannon (1948) showed that the cross entropy can never be smaller than the entropy, i.e., $H(p^*(X)) \leq H(p^*(X) \| \beta)$. In other words, we always get a larger average code length, whenever we use the wrong coding system $f(X | \beta)$. To see why this holds, we decompose the cross entropy as a sum of the entropy and the Kullback-Leibler divergence,²⁷ and show that the latter cannot be negative. This decomposition follows from the definition of cross entropy and

²⁵Shannon denoted this quantity with an H to refer to the capital Greek letter for eta. It seems that John von Neumann convinced Claude Shannon to call this quantity entropy rather than information (Tribus and McIrvine, 1971).

²⁶Observe that the entropy $H(p^*(X))$ is the just the cross entropy from the true $p^*(X)$ to the true coding system $f(X | \beta^*)$.

²⁷The KL-divergence is also known as the relative entropy.

a subsequent addition and subtraction of the entropy resulting in

$$H(p^*(X) \parallel \beta) = H(p^*(X)) + \underbrace{\sum_{x \in \mathcal{X}} (\log \frac{p^*(x)}{f(x \mid \beta^*)}) p^*(x)}_{D(p^*(X) \parallel \beta)}, \quad (13.D.4)$$

where $D(p^*(X) \parallel \beta)$ defines the *Kullback-Leibler divergence* from the true pmf $p^*(X)$ to the postulated coding system $f(X \mid \beta)$. Using the so-called *Jensen's inequality* it can be shown that the KL-divergence is non-negative and that it is only zero whenever $f(X \mid \beta) = p^*(X)$. Thus, the cross entropy can never be smaller than the entropy. Consequently, to minimise the load on the communication network, we have to minimise the cross entropy with respect to the parameter β . Unfortunately, however, we cannot do this in practice, because the cross entropy is a population quantity based on the unknown true pmf $p^*(X)$. Instead, we do the next best thing by replacing the true $p^*(X)$ in Eq. (13.D.3) by the empirical pmf that gives the relative occurrences of the outcomes in the sample rather than in the population. Hence, for any postulated $f(X \mid \beta)$, with β fixed, we approximate the population average defined in Eq. (13.D.3) by the sample average

$$H(x_{\text{obs}}^n \parallel \beta) = H(\hat{p}_{\text{obs}}(X) \parallel f(X \mid \beta)) = \sum_{i=1}^n -\log f(x_{\text{obs},i} \mid \beta) = -\log f(x_{\text{obs}}^n \mid \beta). \quad (13.D.5)$$

We call the quantity $H(x_{\text{obs}}^n \parallel \beta)$ the log-loss from the observed data x_{obs}^n , i.e., the empirical pmf $\hat{p}_{\text{obs}}(X)$, to the coding system $f(X \mid \beta)$.

13.D.2 Data compression and statistical inference

The entropy inequality $H(p^*(X)) \leq H(p^*(X) \parallel \beta)$ implies that the coding theorist's goal of finding the coding system $f(X \mid \beta)$ with the shortest average code length is in fact equivalent to the statistical goal of finding the true data generating process $p^*(X)$. The coding theorist's best guess is the coding system $f(X \mid \beta)$ that minimises the log-loss from x_{obs}^n to the model \mathcal{M}_B . Note that minimising the negative log-likelihood is the same as maximising the likelihood. Hence, the log-loss is minimised by the coding system associated with the MLE, thus, the predictive pmf $f(X \mid \hat{\beta}_{\text{obs}})$. Furthermore, the cross entropy decomposition shows that minimisation of the log-loss is equivalent to minimisation of the KL-divergence from the observations x_{obs}^n to the model \mathcal{M}_B . The advantage of having the optimisation problem formulated in terms of KL-divergence is that it has a known lower bound, namely, zero. Moreover, whenever the KL-divergence from x_{obs}^n to the code $f(X \mid \hat{\beta}_{\text{obs}})$ is larger than zero, we then know that the empirical pmf associated to the observations does not reside on the model. In particular, Section 13.4.3.1 showed that the MLE plugin, $f(X \mid \hat{\beta}_{\text{obs}})$ is the pmf on the model that is closest to the data. This geometric interpretation is due to the fact that we retrieve the Fisher-Rao metric, when we take the second derivative of the KL-divergence with respect to β (Kullback and Leibler, 1951). This connection between the KL-divergence and Fisher information is exploited in Ghosal et al.

(1997) to generalise the Jeffreys's prior to nonparametric models, see also van Erven and Harremos (2014) for the relationship between KL-divergence and the broader class of divergence measures developed by Rényi (1961), see also Campbell (1965).

13.E Regularity conditions

A more mathematically rigorous exposition of the subject would have had this section as the starting point, rather than the last section of the appendix. The regularity conditions given below can be seen as a summary, and guidelines for model builders. If we as scientists construct models such that these conditions are met, we can then use the results presented in the main text. We first give a more general notion of statistical models, then state the regularity conditions followed by a brief discussion on these conditions.

The goal of statistical inference is to find the true probability measure P^* that governs the chances with which X takes on its events. A model \mathcal{P}_Θ defines a subset of \mathcal{P} , the largest collection of all possible probability measures. We as model builders choose \mathcal{P}_Θ and perceive each probability measure P within \mathcal{P}_Θ as a possible explanation of how the events of X were or will be generated. When $P^* \in \mathcal{P}_\Theta$ we have a well-specified model and when $P^* \notin \mathcal{P}_\Theta$, we say that the model is misspecified.

By taking \mathcal{P}_Θ to be equal to the largest possible collection \mathcal{P} , we will not be misspecified. Unfortunately, this choice is not helpful as the complete set is hard to track and leads to uninterpretable inferences. Instead, we typically construct the candidate set \mathcal{P}_Θ using a parameterisation that sends a label $\theta \in \Theta$ to a probability measure P_θ . For instance, we might take the label $\theta = (\begin{smallmatrix} \mu \\ \sigma^2 \end{smallmatrix})$ from the parameter space $\Theta = \mathbb{R} \times (0, \infty)$ and interpret these two numbers as the population mean and variance of a normal probability P_θ . This distributional choice is typical in psychology, because it allows for very tractable inference with parameters that are generally over-interpreted. Unfortunately, the normal distribution comes with rather stringent assumptions resulting in a high risk of misspecification. More specifically, the normal distribution is far too ideal, as it supposes that the population is nicely symmetrically centred at its population mean and outliers are practically not expected due to its tail behaviour.

Statistical modelling is concerned with the intelligent construction of the candidate set \mathcal{P}_Θ such that it encapsulates the true probability measure P^* . In other words, the restriction of \mathcal{P} to \mathcal{P}_Θ in a meaningful manner. Consequently, the goal of statistical inference is to give an informed guess \tilde{P} within \mathcal{P}_Θ for P^* based on the data. This guess should give us insights to how the data *were* generated and how yet unseen data *will be generated*. Hence, the goal is not to find the parameters as they are mere labels. Of course parameters can be helpful, but they should not be the goal of inference.

Note that our general description of a model as a candidate set \mathcal{P}_Θ does not involve any structure –thus, the members of \mathcal{P}_Θ do not need to be related to each other in any sense. We use the parameterisation to transfer the structure of our labels Θ to a structure on \mathcal{P}_Θ . To do so, we require that Θ is a nice

open subset of \mathbb{R}^d . Furthermore, we require that each label defines a member P_θ of \mathcal{P}_Θ unambiguously. This means that if θ^* and θ differ from each other that the resulting pair of probability measure P_{θ^*} and P_θ also differ from each other. Equivalently, we call a parameterisation identifiable whenever $\theta^* = \theta$ leads to $P_{\theta^*} = P_\theta$. Conversely, identifiability implies that when we know everything about P_θ , we can then also use the inverse of the parameterisation to pinpoint the unique θ that corresponds to P_θ . We write $\nu : \mathcal{P}_\Theta \rightarrow \Theta$ for the functional that attaches to each probability measure P a label θ . For instance, ν could be defined on the family of normal distribution such that $P \mapsto \nu(P) = \begin{pmatrix} E_P(X) \\ \text{Var}_P(X) \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. In this case we have $\nu(\mathcal{P}_\Theta) = \Theta$ and, therefore, a one-to-one correspondence between the probability measures $P_\theta \in \mathcal{P}_\Theta$ and the parameters $\theta \in \Theta$.

By virtue of the parameterisation and its inverse ν , we can now transfer additional structure from Θ to \mathcal{P}_Θ . We assume that each probability measure P_θ that is defined on the events of X can be identified with a probability density function (pdf) $p_\theta(x)$ that is defined on the outcomes of X . For this assumption, we require that the set \mathcal{P}_Θ is dominated by a so-called countably additive measure λ . When X is continuous, we usually take for λ the Lebesgue measure that assigns to each interval of the form (a, b) a length of $b - a$. Domination allows us to express the probability of X falling in the range (a, b) under P_θ by the ‘‘area under the curve of $p_\theta(x)$ ’’, that is, $P_\theta(X \in (a, b)) = \int_a^b p_\theta(x)dx$. For discrete variables X taking values in $\mathcal{X} = \{x_1, x_2, x_3, \dots\}$, we take λ to be the counting measure. Consequently, the probability of observing the event $X \in A$ where $A = \{a = x_1, x_2, \dots, b = x_k\}$ is calculated by summing the pmf at each outcome, that is, $P_\theta(X \in A) = \sum_{x=a}^{x=b} p_\theta(x)$. Thus, we represent \mathcal{P}_Θ as the set $\mathcal{P}_\Theta = \{p_\theta(x) : \theta \in \Theta, P_\theta(x) = \int_{-\infty}^x p_\theta(y)dy \text{ for all } x \in \mathcal{X}\}$ in function space. With this representation of \mathcal{P}_Θ in function space, the parameterisation is now essentially the functional relationship f that pushes each θ in Θ to a pdf $p_\theta(x)$. If we choose f to be regular, we can then also transfer additional topological structure from Θ to \mathcal{P}_Θ .

Definition 13.E.1 (Regular parametric model). We call the model \mathcal{P}_Θ a *regular parametric model*, if the parameterisation $\theta \mapsto p_\theta(x) = f(x | \theta)$, that is, the functional relationship f , satisfies the following conditions

- (i) its domain Θ is an open subset of \mathbb{R}^d ,
- (ii) at each possible true value $\theta^* \in \Theta$, the spherical representation $\theta \mapsto m_\theta(x) = 2\sqrt{p_\theta(x)} = 2\sqrt{f(x | \theta)}$ is so-called Fréchet differentiable in $L_2(\lambda)$. The tangent function, i.e., the ‘‘derivative’’ in function space, at $m_{\theta^*}(x)$ is then given by

$$\frac{dm_\theta(x)}{d\theta} d\theta = \frac{1}{2}(\theta - \theta^*)^T \dot{l}(x | \theta^*) m_{\theta^*}(x), \quad (13.E.1)$$

where $\dot{l}(x | \theta^*)$ is a d -dimensional vector of score functions in $L_2(P_{\theta^*})$,

- (iii) the Fisher information matrix $I_X(\theta)$ is non-singular,
- (iv) the map $\theta \mapsto \dot{l}(x | \theta)m_\theta(x)$ is continuous from Θ to $L_2^d(\lambda)$.

Note that (ii) allows us to generalise the geometrical concepts discussed in Appendix 13.C.3 to more general random variables X . \diamond

We provide some intuition. Condition (i) implies that Θ inherits the topological structure of \mathbb{R}^d . In particular, we have an inner product on \mathbb{R}^d that allows us to project vectors onto each other, a norm that allows us to measure the length of a vector, and the Euclidean metric that allows us to measure the distance between two vectors by taking the square root of the sums of squares, that is, $\|\theta^* - \theta\|_2 = \sqrt{\sum_{i=1}^d (\theta_i^* - \theta_i)^2}$. For $d = 1$ this norm is just the absolute value, which is why we previously denoted this as $|\theta^* - \theta|$.

Condition (ii) implies that the measurement of distances in \mathbb{R}^d generalises to the measurement of distance in function space $L_2(\lambda)$. Intuitively, we perceive functions as vectors and say that a function h is a member of $L_2(\lambda)$, if it has a finite norm (length), i.e., $\|h(x)\|_{L_2(\lambda)} < \infty$, meaning

$$\|h(x)\|_{L_2(\lambda)} = \begin{cases} \sqrt{\int_{\mathcal{X}} [h(x)]^2 dx} & \text{if } X \text{ takes on outcomes on } \mathbb{R}, \\ \sqrt{\sum_{x \in \mathcal{X}} [h(x)]^2} & \text{if } X \text{ is discrete.} \end{cases} \quad (13.E.2)$$

As visualised in the main text, by considering $\mathcal{M}_\Theta = \{m_\theta(x) = \sqrt{p_\theta(x)} \mid p_\theta \in \mathcal{P}_\theta\}$ we relate Θ to a subset of the sphere with radius two in the function space $L_2(\lambda)$. In particular, Section 13.4 showed that whenever the parameter is one-dimensional, thus, a line, that the resulting collection \mathcal{M}_Θ also defines a line in model space. Similarly, Appendix 13.C.3 showed that whenever the parameter space is a subset of $[0, 1] \times [0, 1]$ that the resulting \mathcal{M}_Θ also forms a plain.

Fréchet differentiability at θ^* is formalised as

$$\frac{\|m_\theta(x) - m_{\theta^*}(x) - \frac{1}{2}(\theta - \theta^*)^T \dot{l}(x \mid \theta^*) m_{\theta^*}(x)\|_{L_2(\lambda)}}{\|\theta - \theta^*\|_2} \rightarrow 0. \quad (13.E.3)$$

This implies that the linear term $\frac{1}{2}(\theta - \theta^*)^T \dot{l}(x \mid \theta^*) m_{\theta^*}(x)$ is a good approximation to the “error” $m_\theta(x) - m_{\theta^*}(x)$ in the model \mathcal{M}_Θ , whenever θ is close to θ^* given that the score functions $\dot{l}(x \mid \theta^*)$ do not blow up. More specifically, this means that each component of $\dot{l}(x \mid \theta^*)$ has a finite norm. We say that the component $\frac{\partial}{\partial \theta_i} l(x \mid \theta^*)$ is in $L_2(P_{\theta^*})$, if $\|\frac{\partial}{\partial \theta_i} l(x \mid \theta^*)\|_{L_2(P_{\theta^*})} < \infty$, meaning

$$\left\| \frac{\partial}{\partial \theta_i} l(x \mid \theta^*) \right\|_{L_2(P_{\theta^*})} = \begin{cases} \sqrt{\int_{x \in \mathcal{X}} \left(\frac{\partial}{\partial \theta_i} l(x \mid \theta^*) \right)^2 p_{\theta^*}(x) dx} & \text{if } X \text{ is continuous,} \\ \sqrt{\sum_{x \in \mathcal{X}} \left(\frac{\partial}{\partial \theta_i} l(x \mid \theta^*) \right)^2 p_{\theta^*}(x)} & \text{if } X \text{ is discrete.} \end{cases} \quad (13.E.4)$$

This condition is visualised in Fig. 13.12 and Fig. 13.13 by tangent vectors with finite lengths. Under P_{θ^*} , each component $i = 1, \dots, d$ of the tangent vector is expected to be zero, that is,

$$\begin{cases} \int_{x \in \mathcal{X}} \frac{\partial}{\partial \theta_i} l(x \mid \theta^*) p_{\theta^*}(x) = 0 & \text{if } X \text{ is continuous,} \\ \sum_{x \in \mathcal{X}} \frac{\partial}{\partial \theta_i} l(x \mid \theta^*) p_{\theta^*}(x) = 0 & \text{if } X \text{ is discrete.} \end{cases} \quad (13.E.5)$$

This condition follows from the chain rule applied to the logarithm and an exchange of the order of integration with respect to x , and derivation with respect to θ_i , as

$$\int_{x \in \mathcal{X}} \frac{\partial}{\partial \theta_i} l(x | \theta^*) p_{\theta^*}(x) dx = \int_{x \in \mathcal{X}} \frac{\partial}{\partial \theta_i} p_{\theta^*}(x) dx = \frac{\partial}{\partial \theta_i} \int_{x \in \mathcal{X}} p_{\theta^*}(x) dx = \frac{\partial}{\partial \theta_i} 1 = 0. \quad (13.E.6)$$

Note that if $\int \frac{\partial}{\partial \theta_i} p_{\theta^*}(x) dx > 0$, then a small change at θ^* will lead to a function $p_{\theta^* + d\theta}(x)$ that does not integrate to one and, therefore, not a pdf.

Condition (iii) implies that the model does not collapse to a lower dimension. For instance, when the parameter space is a plain the resulting model \mathcal{M}_Θ cannot be line. Lastly, condition (iv) implies that the tangent functions change smoothly as we move from $m_{\theta^*}(x)$ to $m_\theta(x)$ on the sphere in $L_2(\lambda)$, where θ is a parameter value in the neighbourhood of θ^* .

The following conditions are stronger, thus, less general, but avoid Fréchet differentiability and are typically easier to check.

Lemma 13.E.1. *Let $\Theta \subset \mathbb{R}^d$ be open. At each possible true value $\theta^* \in \Theta$, we assume that $p_\theta(x)$ is continuously differentiable in θ for λ -almost all x with tangent vector $\dot{p}_{\theta^*}(x)$. We define the score function at x as*

$$\dot{l}(x | \theta^*) = \frac{\dot{p}_{\theta^*}(x)}{p_{\theta^*}(x)} 1_{[p_{\theta^*} > 0]}(x), \quad (13.E.7)$$

where $1_{[p_{\theta^*} > 0]}(x)$ is the indicator function

$$1_{[p_{\theta^*} > 0]}(x) = \begin{cases} 1 & \text{for all } x \text{ such that } p_{\theta^*}(x) > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (13.E.8)$$

The parameterisation $\theta \mapsto P_\theta$ is regular, if the norm of the score vector Eq. (13.E.7) is finite in quadratic mean, that is, $\dot{l}(X | \theta^*) \in L_2(P_{\theta^*})$, and if the corresponding Fisher information matrix based on the score functions Eq. (13.E.7) is non-singular and continuous in θ . \diamond

There are many better sources than the current manuscript on this topic that are mathematically much more rigorous and better written. For instance, Bickel et al. (1993) give a proof of the lemma above and many more beautiful, but sometimes rather (agonisingly) technically challenging, results. For a more accessible, but no less elegant, exposition of the theory we highly recommend van der Vaart (1998).

Part VI

Conclusion

Chapter 14

Discussion and Future Directions

In this dissertation we advocate the use of Bayes factors in empirical research to replace or complement standard null hypothesis tests based on p -values. These Bayes factors were specifically designed to quantify the evidence for or against the existence of an effect. This was done by comparing two models with the same distributional assumptions, where the alternative model is an extension of the null model by incorporating one extra parameter. Furthermore, instead of returning a decision to “reject” or “not reject”, a Bayes factor $\text{BF}_{10}(d)$ returns a non-negative number that represents the evidence within the observed data for the model that includes the effect. The returned number can be seen as a refinement of the binary decision with $\text{BF}_{10}(d) = \infty$ and $\text{BF}_{10}(d) = 0$ corresponding to definite rejection and acceptance of the null, respectively. Moreover, the Bayes factor allows its users to forgo the binary decision and acknowledge uncertainty, so that the evidence can be updated continually in light of new data, directly and easily. For empirical scientists to be able to use these Bayes factors we implemented them in *Jeffreys’s Amazing Statistics Program*, JASP, which is freely available and open-source (url: <https://jasp-stats.org>).

In Chapter 8 we showed how easy it is do a Bayesian reanalysis of published results in JASP. Most of the discussion centred on how Bayes factors quantify evidence from data already observed, but future research should also focus on how the already observed data can be used for follow-up experiments. This idea of generalising past observations to future data underlies the replication Bayes factor discussed in Chapter 9. Comprehensive knowledge updating requires that the data come from the same population, which is why we emphasised the role of openness and transparency in Chapter 7. By making research materials and data available, future researchers can then conduct a direct replication and build upon previous work. In some cases, however, a replication on the same population is not possible. For correlation and t -test Bayes factors we can nonetheless do meaningful inference by relocating the data, while for more complicated settings such as ANOVAs and contingency tables this is still work in progress.

When no previous data are available we recommend the use of default Jeffreys’s Bayes factors that are constructed from priors that adhere to the general criteria

for Bayesian model choice (Bayarri et al., 2012). One goal of this dissertation was to explain, apply, and extend these general criteria to scenarios common to empirical scientists. The first extension was to Pearson’s correlation, Chapter 2, resulting in an analytic Bayes factor, which was further extended to Kendall’s τ in Chapter 4. By modelling the test statistic, more specifically, approximating the sampling distribution of the test statistic with its asymptotic normal equivalent, a Bayes factor was derived that leads to interpretable results and is fast to compute. In future research we plan to apply the general procedure based on the asymptotic normal approximation and parametric yoking to other scenarios. The use of the normal approximation to the true sampling distribution, however, is not as principled as we wanted it to be and led to Bayes factors that provide less evidence for the alternative, whenever τ is far from zero. This motivated us to consider different approaches and the latent normal approach in van Doorn et al. (2017) in particular. Future research should further explore the relationship between Kendall’s τ and certain copula families as this will provide insights in statistical research on dependency.

The calculations used for the analytic posteriors for Pearson’s ρ also led to the informed t -test in Chapter 5. This work can easily be adapted to linear regression and is worth exploring further.

The first analytic result of Chapter 11 was used to construct a limit-consistent Bayes factor for the two-sample Poisson problem in Chapter 6. In future research we will use the posterior for the odds ratio to formulate a Bayesian test for two proportions, the homogeneity of the odds ratio and the test for independence in multiple 2-by-2 tables. Chapter 6 also described our attempt to extend Jeffreys’s principles of testing to problems that deal with discrete random variables based on the desideratum of limit-consistency. Further research should also focus on the relationship between predictive matching and limit-consistency, as the latter criterion might provide a fruitful technique to generalise Jeffreys’s ideas on testing to other settings.

Jeffreys’s principles to construct Bayes factors, however, requires one of the parameters to be perceived as the test-relevant one and the others as nuisance. This might be difficult for high-dimensional problems, but can be done for location-scale problems and the variable selection problem in particular. The multiplicity introduced can then be tackled by the method discussed by Scott and Berger (2006, 2010), which have yet to be incorporated in JASP. Furthermore, Jeffreys’s construction also requires that we choose the distribution form of the models, which increases the hazard of model misspecification. Model misspecification can have dramatic effects on Bayesian methods as was shown by Grünwald and van Ommen (2014). Fortunately, Grünwald (2017) and colleagues also developed a framework for safe Bayesian inference and methods to detect model misspecification. Further research in this area is necessary and on the way. One goal, therefore, is to extend Jeffreys’s principled Bayes factors to nonparametric models, which in itself comes with additional challenges of tractability and once again multiplicity.

To control for multiplicity with the Bayes factors described here, we recommended that researchers preregister their hypotheses and the tests they perform. The reason for this is that testing is a confirmatory tool of inference concerned with model uncertainty and that this differs from an estimation problem. Esti-

mation and exploration, however, should not be undervalued as they allow for the construction of theories and models, which can subsequently be tested. Models are always simplified description of reality and can always be improved upon.

Bayesian methods can help discover and improve models. For instance, by Bayesian model averaging, or by exploring the posterior of fitted models using so-called plausible values to give insights to how a hierarchical model should be formulated (e.g., Ly et al., 2017a; Marsman, 2014; Marsman et al., 2016b). For instance, in Ly et al. (2017a) we used plausible values to generalise the finding of Forstmann et al. (2008) based on $n = 19$ participants to the general population. Key to this generalisation was the acknowledgement of uncertainty via the posteriors and the mixing of the analytical posteriors developed in Chapter 10. When the posterior is not analytic, one can use the bridge sampler instead, see Chapter 12. The mixing of posteriors in Ly et al. (2017a), however, implies that the posterior, and the marginal likelihood in particular, can be evaluated quickly. Hence, to further make Bayesian methods accessible to empirical scientist, we need to make these sampling methods more efficient. Lastly, Chapter 13 provides some insights in the nature of statistical models and provides the empirical scientists with regularity conditions that allow them to formulate models in which the standard methods are (asymptotically) valid.

We hope to have made a convincing case for the use of Bayesian methods in the empirical sciences, and the Bayes factor in particular when it comes to testing. Our advocacy for Bayesian methods in psychology is, in essence, a call to adopt a principled method of learning. This call is neither new nor controversial, as Bayesian methods have been adopted in fields such as econometrics, statistics and computer science with great success.

References

- Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. Dover Publications. 108, 157, 160
- Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., and Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, 32:1376–1402. 195
- Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., and Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience Psychology and Economics*, 4:95–110. 173, 201
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 231
- Aldrich, J. (2005). The statistical education of Harold Jeffreys. *International Statistical Review*, 73(3):289–307. 225
- Amari, S.-I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., and Rao, C. R. (1987). *Differential geometry in statistical inference*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 10. Institute of Mathematical Statistics, Hayward, CA. 253
- Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., and Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, 41:521–543. 37
- Andrews, M. and Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, 66:1–7. 171
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21. 37, 128
- Atkinson, C. and Mitchell, A. (1981). Rao’s distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365. 253
- Bailey, W. N. (1964). *Generalized hypergeometric series*. Cambridge Tracts in Mathematics and Mathematical Physics, No. 32. Stechert-Hafner, Inc., New York. 162
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533:452–454. 4, 301

- Balasubramanian, V. (1996). A geometric formulation of Occam's razor for inference of parametric distributions. *arXiv preprint adap-org/9601001*. 232, 241
- Banterle, M., Grazian, C., Lee, A., and Robert, C. P. (2015). Accelerating Metropolis-Hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*. 243
- Barber, T. X. (1976). *Pitfalls in Human Research: Ten Pivotal Points*. Pergamon Press Inc., New York. 117
- Bark, R., Dieckmann, S., Bogerts, B., and Northoff, G. (2005). Deficit in decision making in catatonic schizophrenia: An exploratory study. *Psychiatry Research*, 134:131–141. 194
- Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 18, 103
- Batchelder, W. H. and Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, 87:375–397. 213
- Batchelder, W. H. and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6:57–86. 232
- Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G., Bertin, D., Fulks, W. B., Harvey, A. R., Thomsen Jr, D. L., Weber, M. A., Whitney, E. L., and Stampfel, R. (1953). *Higher transcendental functions*, volume 2. McGraw-Hill New York. 157
- Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G., Bertin, D., Fulks, W. B., Harvey, A. R., Thomsen Jr, D. L., Weber, M. A., Whitney, E. L., and Stampfel, R. (1954). *Tables of integral transforms*, volume 1. McGraw-Hill. 86, 152, 157
- Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trabajos de estadística y de investigación operativa*, 32(3):18–31. 150, 151
- Bayarri, M. J., Benjamin, D. J., Berger, J. O., and Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103. 116, 171
- Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142. 58
- Bayarri, M. J. and Berger, J. O. (2013). Hypothesis testing and model uncertainty. In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A., editors, *Bayesian Theory and Applications*, pages 361–394. Oxford University Press, Oxford. 103
- Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577. 25, 42, 47, 83, 103, 222, 268
- Bayarri, M. J. and Mayoral, A. M. (2002). Bayesian analysis and design for comparison of effect-sizes. *Journal of Statistical Planning and Inference*, 103:225–243. 130
- Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50:7–15. 174, 193, 194, 205

- Bechara, A., Damasio, H., Damasio, A. R., and Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, 19:5473–5481. 194
- Bechara, A., Damasio, H., Tranel, D., and Anderson, S. W. (1998). Dissociation of working memory from decision making within the human prefrontal cortex. *Journal of Neuroscience*, 18:428–437. 194
- Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275:1293–1295. 194
- Bechara, A., Tranel, D., and Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, 123:2189–2202. 194
- Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22:245–268. 171, 173
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Verlag. 37, 43, 242
- Berger, J. O. (2006). Bayes factors. In Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, vol. 1 (2nd ed.), pages 378–386. Wiley, Hoboken, NJ. 10, 127, 131, 173
- Berger, J. O., Bernardo, J. M., and Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, 10(1):189–221. 150
- Berger, J. O. and Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics: Vol. 4*, pages 29–72. Springer Verlag, New York. 82
- Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, pages 317–335. 120
- Berger, J. O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59:3 – 15. 172
- Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207. 47
- Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 307–321. 45, 103, 256
- Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics*, pages 963–982. 73, 150, 151
- Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). Institute of Mathematical Statistics, Hayward (CA). 116
- 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. 14
- Bernardo, J. M. (2015). An overall prior for the five-parameter normal distribution. Paper presented at the 11th International Workshop on Objective Bayes Methodology dedicated to Susie Bayarri. 149
- Bickel, P. J. (2006). Regularization in statistics. *Test*, 15(2):271–344. With discussion by Li, Tsybakov, van de Geer, Yu, Valdés, Rivero, Fan, van der Vaart, and a rejoinder by the author. 64

REFERENCES

- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore. 219, 263
- Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237. 46, 119, 242
- Blair, R. J. R., Colledge, E., and Mitchell, D. G. V. (2001). Somatic markers and response reversal: Is there orbitofrontal cortex dysfunction in boys with psychopathic tendencies? *Journal of Abnormal Child Psychology*, 29:499–511. 194
- Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A. J., Brown, S. D., and Forstmann, B. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, 66:115–133. 34
- Bolt, B. (1982). The constitution of the core: seismological evidence. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 306(1492):11–20. 11
- Borgwardt, K. M. and Ghahramani, Z. (2009). Bayesian two-sample tests. *arXiv preprint arXiv:0906.4032*. 64, 71
- Brown, C. E. (1998). Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences*, pages 155–157. Springer. 193
- Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, pages 101–117. 221, 249
- Burbea, J. (1984). Informative geometry of probability spaces. Technical report, DTIC Document. 253
- Burbea, J. and Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4):575–596. 253
- Burbea, J. and Rao, C. R. (1984). Differential metrics in probability spaces. *Probability and mathematical statistics*, 3(2):241–258. 253
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer Verlag, New York. 231
- Busemeyer, J. R. and Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, 14:253–262. 174, 193, 195, 196, 197, 198, 199, 201, 203, 204, 205
- Butler, R. W. and Wood, A. T. A. (2002). Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30(4):1155–1177. 110
- Calin-Jageman, R. J. and Caldwell, T. L. (2014). Replication of the superstition and performance study by Damisch, Stoberock, and Mussweiler (2010). *Social Psychology*, 45:239–245. 137
- Campbell, L. L. (1965). A coding theorem and Rényi’s entropy. *Information and Control*, 8(4):423–429. 260
- Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–484. 174, 206

- Cavedini, P., Riboldi, G., D'Annucci, A., Belotti, P., Cisima, M., and Bellodi, L. (2002a). Decision-making heterogeneity in obsessive-compulsive disorder: Ventromedial prefrontal cortex function predicts different treatment outcomes. *Neuropsychologia*, 40:205–211. 194
- Cavedini, P., Riboldi, G., Keller, R., D'Annucci, A., and Bellodi, L. (2002b). Frontal lobe dysfunction in pathological gambling patients. *Biological Psychiatry*, 51:334–341. 194
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49:609–610. 99, 117
- Chandramouli, S. and Shiffrin, R. M. (2016). Extending Bayesian induction. *Journal of Mathematical Psychology*, 72:38–42. 41, 51, 52, 54, 57
- Chechile, R. A. (1973). *The Relative Storage and Retrieval Losses in Short-Term Memory as a Function of the Similarity and Amount of Information Processing in the Interpolated Task*. PhD thesis, University of Pittsburgh. 213
- Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media. 173
- Chen, Y. and Hanson, T. E. (2014). Bayesian nonparametric k-sample tests for censored and uncensored data. *Computational Statistics & Data Analysis*, 71:335–346. 71
- Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *The Annals of Mathematical Statistics*, pages 972–994. 72
- Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281. 207
- Colonius, H. (2016). An invitation to coupling and copulas: With applications to multisensory modeling. *Journal of Mathematical Psychology*, 74:2–10. 74
- Consonni, G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23(3):332–353. 45
- Cook, A. (1990). Sir Harold Jeffreys. 2 April 1891–18 March 1989. *Biographical Memoirs of Fellows of the Royal Society*, 36:302–333. 11
- Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons. 257
- Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39. 256
- Cramér, H. (1946). Methods of mathematical statistics. *Princeton University Press*, 23. 249
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25:7–29. 124
- Cvitković, M., Smith, A.-S., and Pande, J. (2017). Asymptotic expansions of the hypergeometric function with two large parameters-application to the partition function of a lattice gas in a field of traps. *Journal of Physics A: Mathematical and Theoretical*, 50(26). 112
- Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., and Stout, J. C. (2015). An improved cognitive model of the Iowa and Soochow gambling tasks with regard

- to model fitting performance and tests of parameter consistency. *Frontiers in Psychology*, 6:229. 195
- Dai, X., Wertenbroch, K., and Brendl, C. M. (2008). The value heuristic in judgments of relative frequency. *Psychological Science*, 19:18–19. 139, 140
- Damisch, L., Stoerrock, B., and Mussweiler, T. (2010). Keep your fingers crossed! How superstition improves performance. *Psychological Science*, 21:1014–1020. 137, 138
- Dass, S. C. and Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *Journal of statistical planning and inference*, 119(1):143–152. 65
- Dawid, A. P. (1977). Further comments on some comments on a paper by Bradley Efron. *The Annals of Statistics*, 5(6):1249. 253
- Dawid, A. P. (2011). Posterior model probabilities. In Gabbay, D. M., Bandyopadhyay, P. S., Forster, M. R., Thagard, P., and Woods, J., editors, *Handbook of the Philosophy of Science*, volume 7, pages 607–630. Elsevier, North-Holland. 222, 242
- Dawid, A. P. and Lauritzen, S. L. (2001). Compatible prior distributions. *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pages 109–118. 45
- Dawid, A. P., Stone, M., and Zidek, J. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 189–233. 151
- de Rooij, S. and Grünwald, P. D. (2011). Luckiness and regret in Minimum Description Length inference. In Gabbay, D. M., Bandyopadhyay, P. S., Forster, M. R., Thagard, P., and Woods, J., editors, *Handbook of the Philosophy of Science*, volume 7, pages 865–900. Elsevier, North-Holland. 231, 232
- DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915. 174, 183, 184, 205
- Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, pages 204 – 223. 193, 203, 206
- Dickey, J. M. and Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, 41:214–226. 22, 70, 98, 133, 193, 203, 206
- Didelot, X., Everitt, R. G., Johansen, A. M., and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76. 172
- Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375. 48
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781. 132
- Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669. 64
- Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Allen, J., Banks, J., Baranski, E., Bernstein, M., Bonfiglio, D., Boucher, L., Brown, E., Budiman, N.,

- Cairo, A., Capaldi, C., Chartier, C., Chung, J., Cicero, D., Coleman, J., Conway, J., Davis, W., Devos, T., Fletcher, M., German, K., Grahe, J., Hermann, A., Hicks, J., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D., Joy-Gaba, J., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R., Lucas, R., Lustgraaf, C., Martin, D., Menon, M., Metzger, M., Moloney, J., Morse, P., Prislin, R., Razza, T., Re, D., Rule, N., Sacco, D., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R., Summerville, A., Tskhay, K., van Allen, Z., Vaughn, L., Walker, R., Weinberg, A., Wilson, J., Wirth, J., Wortman, J., and Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82. 4, 129, 301
- Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, 63:400–402. 120
- Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242. 15, 46, 120
- Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242. With a discussion by C. R. Rao, Don A. Pierce, D. R. Cox, D. V. Lindley, Lucien LeCam, J. K. Ghosh, J. Pfanzagl, Niels Keiding, A. Philip Dawid, Jim Reeds and with a reply by the author. 253
- Etz, A. and Vandekerckhove, J. (2017). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*. Manuscript accepted for publication. 132, 133
- Etz, A. and Wagenmakers, E.-J. (2017). J. B. S. Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science*, 32(2):313–329. 43, 47, 83, 102, 124, 131, 173, 225
- Ferguson, T. S., Genest, C., and Hallin, M. (2000). Kendall’s tau for serial dependence. *Canadian Journal of Statistics*, 28:587–604. 74, 75
- Festinger, L. and Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2):203. 123, 125, 126
- Feynman, R. (1998). *The Meaning of it All: Thoughts of a Citizen–Scientist*. Perseus Books, Reading, MA. 117
- Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160. 218
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521. 149
- Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770. 149, 216
- Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32. 149
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368. 149, 218, 247, 248

REFERENCES

- Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725. 218
- Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., and Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, 105(45):17538–17542. 269
- Fraser, D. A. (1961). On fiducial inference. *The Annals of Mathematical Statistics*, pages 661–676. 151
- Fréchet, M. (1943). Sur l’extension de certaines évaluations statistiques au cas de petits échantillons. *Revue de l’Institut International de Statistique*, pages 182–205. 249
- Frieden, B. R. (2004). *Science from Fisher information: A unification*. Cambridge University Press. 242
- Friston, K. J. and Penny, W. (2011). Post hoc Bayesian model selection. *NeuroImage*, 56:2089–2099. 49
- Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7:143–167. 174, 182, 186, 187, 192, 193, 205
- Galton, F. (1907). Vox populi. *Nature*, 75:450–451. 87
- Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton, FL. 37, 172, 174, 178
- Geisser, S. (1980). The contributions of Sir Harold Jeffreys to Bayesian inference. In Zellner, A. and Kadane, Joseph, B., editors, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 13–20. Amsterdam: North-Holland. 11, 37
- Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514. 182, 183
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall/CRC, London, 3rd edition. 73
- Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185. 172, 206
- Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368. 74
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1997). Non-informative priors via sieves and packing numbers. In *Advances in statistical decision theory and applications*, pages 119–132. Springer. 64, 243, 259
- Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531. 64
- Ghosal, S., Lember, J., and Van Der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89. 64
- Ghosh, J. K. (1985). Efficiency of estimates—part I. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 310–325. 248

- Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science*, pages 187–202. 103
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton (FL). 37
- Gino, F. and Wiltermuth, S. S. (2014). Evil genius? How dishonesty can lead to greater creativity. *Psychological Science*, 4:973–981. 21, 22, 23, 24
- Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214. 243
- Goldacre, B. (2009). *Bad Science*. Fourth Estate, London. 117
- Golden, R. M. (1995). Making correct statistical inferences using the wrong probability model. *Journal of Mathematical Psychology*, 39:3–20. 242
- Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, 44(1):153–170. 242
- Good, I. J. (1980). The contributions of Jeffreys to Bayesian statistics. In Zellner, A., editor, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 21–34. North-Holland Publishing Company, Amsterdam, The Netherlands. 11
- Gradshteyn, I. S. and Ryzhik, I. M. (2007). *Table of Integrals, Series, and Products*. Academic Press, 7 edition. 32, 38, 109, 153, 155, 156, 162, 166, 167
- Grazian, C. and Liseo, B. (2014). Approximate integrated likelihood via ABC methods. *arXiv preprint arXiv:1403.0387*. 243
- Grazian, C. and Robert, C. P. (2015). Jeffreys' priors for mixture estimation. In *Bayesian Statistics from Methods to Models and Applications*, pages 37–48. Springer. 51, 243
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. 174, 206
- Greiner, R. (1909). Über das Fehlersystem der Kollektivmasslehre. *Zeitschrift für Mathematik und Physik*, pages 121–158. 73
- Griffin, H. (1958). Graphic computation of tau as a coefficient of disarray. *Journal of the American Statistical Association*, 53:441–447. 70
- Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2017a). Informed Bayesian *t*-tests. *arXiv preprint arXiv:1704.02479*. 126, 128, 137, 222
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017b). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97. 243
- Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., and Wagenmakers, E.-J. (2017c). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, 2:123–138. 140
- Gronau, Q. F. and Wagenmakers, E.-J. (2017). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*. Manuscript accepted for publication. 98, 132
- Grünwald, P. (2016). Safe probability. *arXiv preprint arXiv:1604.01785*. 249

REFERENCES

- Grünwald, P. and van Ommen, T. (2014). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *arXiv preprint arXiv:1412.3730*. 249, 268
- Grünwald, P. D. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, ILLC and University of Amsterdam. 241
- Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge, MA. 231, 232
- Grünwald, P. D. (2017). Safe probability. *Journal of Statistical Planning and Inference*. Manuscript accepted for publication. 268
- Grünwald, P. D. and Mehta, N. A. (2016). Fast rates with unbounded losses. *arXiv preprint arXiv:1605.00252*. 242
- Grünwald, P. D., Myung, I. J., and Pitt, M. A., editors (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press, Cambridge, MA. 231
- Gunel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, 61:545–557. 138, 139
- Haight, F. A. (1967). *Handbook of the Poisson distribution*. Wiley. 102
- Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14(4):323–330. 219, 248
- Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194. 242
- Hald, A. (2008). *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935*. Springer Science & Business Media. 225
- Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28:55–61. 82
- Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*. London: Methuen. 176
- Hannig, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101(473):254–269. 151
- Harms, C. (2016). A Bayes factor for replications of ANOVA results. *arXiv preprint*. 130
- Heck, D. W., Moshagen, M., and Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, 60:29–34. 234
- Heck, D. W., Wagenmakers, E.-J., and Morey, R. D. (2015). Testing order constraints: Qualitative differences between Bayes factors and normalized maximum likelihood. *Statistics & Probability Letters*, 105:157–162. 50
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325. 74
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, pages 382–401. 82, 173

- Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, 10(2):297–320. 64, 71
- Hotelling, H. and Pabs, M. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, 7:29–43. 72
- Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 46, pages 281–284. Cambridge University Press. 256
- Huzurbazar, V. S. (1956). Sufficient statistics and orthogonal parameters. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 17(3):217–220. 256
- Huzurbazar, V. S. (1991). Sir Harold Jeffreys: Recollections of a student. *Chance*, 4(2):18–21. 11
- Inagaki, N. (1970). On the limiting distribution of a sequence of estimators with uniformity property. *Annals of the Institute of Statistical Mathematics*, 22(1):1–13. 219, 248
- Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17:295–311. 179
- Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., and Wagenmakers, E.-J. (2017). Default “Gunel and Dickey” Bayes factors for contingency tables. *Behavior Research Methods*, 49:638–652. 138
- JASP Team (2017). JASP (Version 0.8.2)[Computer software]. 82, 118, 124, 300
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge. 11
- Jeffreys, H. (1924). *The Earth, Its Origin, History and Physical Constitution*. Cambridge University Press. 11
- Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press. 11
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, 31(2):203–222. 16, 102, 150
- Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 165:161–198. 135
- Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press, Oxford, UK, 1st edition. 102, 112
- Jeffreys, H. (1942). On the significance tests for the introduction of new functions to represent measures. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 180(982):256–268. 83
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461. 18, 43, 59, 62, 103, 226
- Jeffreys, H. (1948). *Theory of Probability*. Oxford University Press, Oxford, UK, 2nd edition. 24, 82, 87, 112, 126
- Jeffreys, H. (1955). The present position in probability theory. *The British Journal for the Philosophy of Science*, 5:275–289. 12
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, UK, 3rd edition. 3, 12, 13, 14, 15, 18, 20, 24, 25, 27, 33, 34, 35, 36, 37, 39, 43, 45,

- 47, 49, 69, 70, 73, 76, 82, 83, 84, 99, 102, 103, 105, 124, 127, 131, 135, 150, 151, 173, 186, 222, 225, 256
- Jeffreys, H. (1973). *Scientific Inference*. Cambridge University Press, Cambridge, UK, 3 edition. 12, 13, 14
- Jeffreys, H. (1980). Some general points in probability theory. In Zellner, A. and Kadane, Joseph, B., editors, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 451–453. Amsterdam: North-Holland. 13, 14, 37, 49
- Jeffreys, H. and Jeffreys, B. S. (1946). *Methods of Mathematical Physics*. Cambridge University Press, Cambridge, UK. 11
- Johnson, V. E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):689–701. 69, 71, 79
- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110:19313–19317. 15, 120
- Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170. 51
- Kamary, K., Eun, J., and Robert, Christian, P. (2016). Non-informative reparameterisations for location-scale mixtures. *arXiv preprint arXiv:1601.01178*. 51
- Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. *arXiv preprint arXiv:1412.2044*. 48, 49, 50, 51
- Kary, A., Taylor, R., and Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, 72:210–219. 49
- Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, 4(3):188–234. 228, 232
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795. 70, 83, 102, 124, 131, 172, 186
- Kass, R. E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–144. 256
- Kass, R. E. and Vos, P. W. (2011). *Geometrical foundations of asymptotic inference*, volume 908. John Wiley & Sons. 253
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370. 103
- Kelman, H. C. (1953). Attitude change as a function of response restriction. *Human Relations*, 6(3):185–214. 125
- Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93. 69, 72
- Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Oxford University Press, New York. 69, 73, 74, 75

- Keuken, M. C., Ly, A., Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A., Brown, S. D., and Forstmann, B. U. (2017). Corrigendum to “a purely confirmatory replication study of structural brain-behavior correlations”[cortex 66 (2015) 115–133]. *Cortex*, 93:229–233. 34
- Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., and Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLoS Biology*, 14:e1002456. 118
- Klaassen, C. A. J. and Lenstra, A. J. (2003). Vanishing Fisher information. *Acta Applicandae Mathematicae*, 78(1):193–200. 232
- Klauer, K. C. and Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, 55(6):430–450. 232
- Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381. 119
- Kleijn, B. J. K. and Zhao, Y. Y. (2017). Criteria for posterior consistency. *arXiv preprint arXiv:1308.1263*. 64, 243
- Klein, R., Ratliff, K., Vianello, M., Adams, Jr., R. B., Bahník, v., Bernstein, M., Bocian, K., Brandt, M., Brooks, B., Brumbaugh, C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J., Hovermale, J., Hunt, S., Huntsinger, J., IJzerman, H., John, M.-S., Joy-Gaba, J., Barry Kappes, H., Krueger, L., Kurtz, J., Levitan, C., Mallett, R., Morris, W., Nelson, A., Nier, J., Packard, G., Pilati, R., Rutchick, A., Schmidt, K., Skorinko, J., Smith, R., Steiner, T., Storbeck, J., Van Swol, L., Thompson, D., van 't Veer, A., Vaughn, L., Vranka, M., Wichman, A., Woodzicka, J., and Nosek, B. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45:142–152. 4, 129, 301
- Klugkist, I., Laudy, O., and Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, 10(4):477. 23
- Korostelev, A. P. and Korosteleva, O. (2011). *Mathematical statistics: Asymptotic minimax theory*, volume 119. American Mathematical Society. 242
- Kotz, S., Kozubowski, T. J., and Podgorski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Springer, New York. 221
- Kraft, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master’s thesis, Massachusetts Institute of Technology. 257
- Krishnamoorthy, K. and Thomson, J. (2004). A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1):23–35. 102
- Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354:110–114. 132, 133, 134, 135, 136, 143, 144, 145
- Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53:814–861. 69, 70, 73

REFERENCES

- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. 259
- Labadi, L. A., Masuadi, E., and Zarepour, M. (2014). Two-sample Bayesian nonparametric goodness-of-fit test. *arXiv preprint arXiv:1411.3427*. 64, 71
- LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828. 219, 248
- LeCam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review/Revue Internationale de Statistique*, 58(2):153–171. 218
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15:1–15. 172
- Lee, M. D., Lodewyckx, T., and Wagenmakers, E.-J. (2015). Three Bayesian analyses of memory deficits in patients with dissociative identity disorder. In Raaijmakers, J. R., Criss, A., Goldstone, R., Nosofsky, R., and Steyvers, M., editors, *Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin*, pages 189–200. Psychology Press. 49
- Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge. 10, 222
- Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer Science & Business Media. 225
- Levelt, W. J., Drenth, P., and Noort, E. (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. 4, 301
- Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655. 10, 131, 173
- Li, Y. and Clyde, M. A. (2015). Mixtures of g-priors in generalized linear models. *arXiv preprint arXiv:1503.06913*. 222
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481). 24, 25, 85, 222
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44:187–192. 15, 18, 103
- Lindley, D. V. (1965). *Introduction to Probability & Statistics from a Bayesian Viewpoint. Part 2. Inference*. Cambridge University Press, Cambridge. 150, 151
- Lindley, D. V. (1972). *Bayesian Statistics, a Review*. SIAM, Philadelphia (PA). 87, 133
- Lindley, D. V. (1977). The distinction between inference and decision. *Synthese*, 36:51–58. 43
- Lindley, D. V. (1980). Jeffreys’s contribution to modern statistical thought. In Zellner, A., editor, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 35–39. North-Holland Publishing Company, Amsterdam, The Netherlands. 11
- Lindley, D. V. (1985). *Making Decisions*. Wiley, London, 2 edition. 118, 119
- Lindley, D. V. (1991). Sir Harold Jeffreys. *Chance*, 4(2):10–14. 21, 11
- Lindley, D. V. (1997). Some comments on Bayes factors. *Journal of Statistical Planning and Inference*, 61(1):181–189. 63
- Liu, C. C. and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52:362–375. 36, 120

- López, J. L. and Pagola, P. J. (2011). A systematic “saddle point near a pole” asymptotic method with application to the Gauss hypergeometric function. *Studies in Applied Mathematics*, 127(1):24–37. 112
- Luce, R. (1959). *Individual choice behavior*. Wiley, New York. 196
- Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., and Matzke, D. (2017a). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In Moustafa, A. A., editor, *Computational Models of Brain and Behavior*, pages 467–480. John Wiley & Sons. 269
- Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2017b). Replication Bayes factors from evidence updating. *Manuscript in preparation*. 47, 119, 127, 242
- Ly, A., Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2017c). A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80:40–55. 16, 18, 44, 52, 62, 63, 103, 138, 151, 175
- Ly, A., Marsman, M., and Wagenmakers, E.-J. (2017d). Analytic posteriors for Pearson’s correlation coefficient. *Statistica Neerlandica*. Manuscript accepted for publication. 86, 222
- Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q. F., and Wagenmakers, E.-J. (2017e). Bayesian reanalyses from summary statistics and the strength of statistical evidence. *Manuscript submitted for publication*. 133, 222
- Ly, A., Verhagen, A. J., and Wagenmakers, E.-J. (2016a). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32. 42, 43, 45, 73, 76, 83, 84, 90, 99, 102, 112, 117, 126, 131, 173, 222, 225, 256
- Ly, A., Verhagen, A. J., and Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72:43–55. 112, 117, 126, 131, 173, 222, 225, 256
- Marin, J.-M. and Robert, C. P. (2010). On resolving the Savage–Dickey paradox. *Electronic Journal of Statistics*, 4:643–654. 18, 22, 203
- Marin, J.-M. and Robert, C. P. (2014). *Bayesian essentials with R*. Springer. 48
- Marsman, M. (2014). *Plausible values in statistical inference*. PhD thesis, Universiteit Twente. 269
- Marsman, M., Ly, A., and Wagenmakers, E.-J. (2016a). Four requirements for an acceptable research program. *Basic and Applied Social Psychology*, 38(6):308–312. 140, 242
- Marsman, M., Maris, G., Bechger, T., and Glas, C. (2016b). What can we learn from plausible values? *Psychometrika*, 81(2):274–289. 269
- Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., and Wagenmakers, E.-J. (2017). A Bayesian bird’s eye view of ‘Replications of important results in social psychology’. *Royal Society Open Science*, 4:160426. 82, 130
- Marsman, M. and Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*, 77(3):529–539. 120
- Martino, D. J., Bucay, D., Butman, J. T., and Allegri, R. F. (2007). Neuropsychological frontal impairments and negative symptoms in schizophrenia. *Psychiatry Research*, 152:121–128. 194

REFERENCES

- Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *CHI 2017 Conference Proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems*. 128
- Matzke, D., Dolan, C. V., Batchelder, W. H., and Wagenmakers, E.-J. (2015a). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80:205–235. 172
- Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., and Wagenmakers, E.-J. (2015b). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, 144:e1–e15. 144
- Matzke, D. and Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, 16(5):798–817. 172
- McMillan, B. (1956). Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116. 257
- Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3):220–232. 22
- Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11:552–586. 187, 207
- Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6(4):831–860. 5, 171, 173, 174, 187, 188, 205, 302
- Mira, A. and Nicholls, G. (2004). Bridge estimation of the probability density at a point. *Statistica Sinica*, 14:603–612. 207
- Mitchell, A. F. (1962). Sufficient statistics and orthogonal parameters. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 58, pages 326–337. Cambridge University Press. 256
- Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Coker Morey, C., Newman, D. P., Schönbrodt, F. D., Vanpaemel, W., Wagenmakers, E.-J., and Zwaan, R. A. (2016). The peer reviewers’ openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, 3:150547. 118
- Morey, R. D. and Rouder, J. N. (2015). BayesFactor 0.9.11-1. Comprehensive R Archive Network. 82, 87, 126
- Morey, R. D. and Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, 92:121–124. 22, 91
- Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4. 91
- Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, 67(1):153–171. 33
- Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, 72:104–115. 33

- Mulder, J., Hoijtink, H., and Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, 140(4):887–906. 22
- Mulder, J. and Wagenmakers, E.-J. (2016). Editors’ introduction to the special issue “Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments”. *Journal of Mathematical Psychology*, 72:1–5. 172
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47:90–100. 16, 214, 218
- Myung, I. J., Balasubramanian, V., and Pitt, M. (2000a). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, 97(21):11170–11175. 232
- Myung, I. J., Forster, M. R., and Browne, M. W. (2000b). Guest editors’ introduction: Special issue on model selection. *Journal of Mathematical Psychology*, 44:1–2. 172
- Myung, I. J., Forster, M. R., and Browne, M. W. (2000c). Model selection [Special issue]. *Journal of Mathematical Psychology*, 44(1–2). 37, 231, 232
- Myung, I. J. and Navarro, D. J. (2005). Information matrix. *Encyclopedia of Statistics in Behavioral Science*. 214
- Myung, I. J., Navarro, D. J., and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50:167–179. 231, 232
- Myung, I. J. and Pitt, M. A. (1997). Applying Occam’s razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1):79–95. 10, 37, 124
- Myung, I. J. and Pitt, M. A. (2016). Model comparison in psychology. In Wixted, J. and Wagenmakers, E.-J., editors, *The Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience (Fourth Edition)*, volume 5: Methodology. John Wiley & Sons, New York, NY. Manuscript accepted for publication. 232
- Navarro, D. J., Griffiths, T. L., Steyvers, M., and Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50:101–122. 173
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139. 179
- Nelsen, R. (2006). *An Introduction to Copulas*. Springer-Verlag New York, second edition. 74
- Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48. 178, 182, 183
- Ng, H., Gu, K., and Tang, M. (2007). A comparative study of tests for the difference of two Poisson means. *Computational Statistics & Data Analysis*, 51(6):3085–3099. 102
- Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, 5:241–301. 124
- Noether, G. E. (1955). On a theorem of Pitman. *Annals of Mathematical Statistics*, 26:64–68. 72

REFERENCES

- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.-J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348:1422–1425. 118
- Nosek, B. A. and Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45:137–141. 4, 117, 129, 301
- Ntzoufras, I. (2009). Bayesian model and variable evaluation. In *Bayesian modeling using WinBUGS*, pages 389–433. John Wiley & Sons. 172, 174, 178, 180
- Nuijten, M. B., Hartgerink, C. H., Assen, M. A., Epskamp, S., and Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48(4):1205–1226. 81
- Oberhettinger, F. (1972). Hypergeometric functions. In Abramowitz, M. and Stegun, I. A., editors, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, pages 555–566. New York: Dover Publications. 38, 160
- O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference* (2nd ed.). Arnold, London. 10, 131
- Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W., editors (2010). *NIST handbook of mathematical functions*. U.S. Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge University Press, Cambridge. 161, 163
- Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6):657–660. 117
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. 4, 129, 139, 301
- Overstall, A. M. and Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54:3269–3288. 184, 186, 187, 189, 209
- Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95:135–143. 179
- Pajor, A. (2016). Estimating the Marginal Likelihood Using the Arithmetic Mean Identity. *Bayesian Analysis*, pages 1–27. 178
- Pashler, H. and Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7:528–530. 4, 117, 129, 301
- Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482. 117
- Peirce, C. S. (1883). A theory of probable inference. In Peirce, C. S., editor, *Studies in Logic*, pages 126–181. Little & Brown, Boston. 117
- Pericchi, L. R., Liu, G., and Torres, D. (2008). Objective Bayes factors for informative hypotheses: “Completing” the informative hypothesis and “splitting” the

- Bayes factor. In Hoijtink, H., Klugkist, I., and Boelen, P. A., editors, *Bayesian Evaluation of Informative Hypotheses*, pages 131–154. Springer Verlag, New York. 23
- Pitt, M., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3):472–491. 172, 231, 241
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–8. Vienna. 201
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11. 193
- Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1:969–979. 171
- Przyborowski, J. and Wilenski, H. (1940). Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder. *Biometrika*, 31(3-4):313–323. 102, 104
- Ractliffe, J. F. (1964). The significance of the difference between two Poisson variables: An experimental investigation. *Applied Statistics*, pages 84–86. 102
- Raftery, A. E. (1995). Bayesian model selection in social research. In Marsden, P. V., editor, *Sociological Methodology*, pages 111–196. Blackwells, Cambridge. 231
- Raftery, A. E. and Banfield, J. D. (1991). Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics (Bayesian image restoration, with two applications in spatial statistics)–(Discussion). *Annals of the Institute of Statistical Mathematics*, 43:32–43. 176
- Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Kegan Paul, London. 12
- Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Springer Texts in Statistics. Wiley, New York. 71
- Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91. 249, 253
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85:59–108. 213
- Ray, K. and Schmidt-Hieber, J. (2016). Minimax theory for a class of nonlinear statistical inverse problems. *Inverse Problems*, 32(6):065003. 243
- Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561. 260
- Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64–99. New York. 195
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47. 232, 233, 240

REFERENCES

- Rivoirard, V. and Rousseau, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523. 242
- Robert, C. P. (1993). A note on Jeffreys–Lindley paradox. *Statistica Sinica*, 3(2):601–608. 43
- Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media. 43
- Robert, C. P. (2014). On the Jeffreys–Lindley paradox. *Philosophy of Science*, 81(2):216–232. 43
- Robert, C. P. (2015). The Metropolis–Hastings algorithm. *arXiv preprint arXiv:1504.01896*. 50, 180
- Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72:33–37. 41, 42, 48, 50, 51, 173, 222
- Robert, C. P., Chopin, N., and Rousseau, J. (2009). Harold Jeffreys’s Theory of Probability revisited. *Statistical Science*, pages 141–172. 11, 29, 34, 37, 42, 83, 150, 225
- Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review*, 107:358–367. 232
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21:301–308. 82, 132
- Rouder, J. N. and Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12:573–604. 172, 173
- Rouder, J. N., Lu, J., Morey, R. D., Sun, D., and Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General*, 137:370–389. 173
- Rouder, J. N., Lu, J., Speckman, P., Sun, D., and Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12:195–223. 172, 173
- Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., and Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, 72:621–642. 172
- Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., and Wagenmakers, E.-J. (2016a). Is there a free lunch in inference? *Topics in Cognitive Science*, 8:520–547. 99, 117, 124
- Rouder, J. N., Morey, R. D., and Wagenmakers, E.-J. (2016b). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, 2:1–12. 99
- 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(2):225–237. 10, 21, 24, 82, 84, 90, 99, 118
- Rozeboom, W. W. (1960). The fallacy of the null–hypothesis significance test. *Psychological Bulletin*, 57:416–428. 42, 119
- Rudin, W. (1991). *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition. 228
- Rutherford, E., Geiger, H., and Bateman, H. (1910). LXXVI. the probability variations in the distribution of α particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20(118):698–707. 102

- Salomond, J.-B. (2013). Bayesian testing for embedded hypotheses with application to shape constraints. *arXiv preprint arXiv:1303.6466*. 64
- Salomond, J.-B. (2014). Adaptive Bayes test for monotonicity. In *The Contribution of Young Researchers to Bayesian Statistics*, pages 29–33. Springer. 64
- Savage, L. J. (1961). The foundations of statistics reconsidered. In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, pages 575–586. University of California Press, Berkely, CA. 82
- Scheibehenne, B., Gronau, Q. F., Jamil, T., and Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmacz, Williams, and Burkner. *Psychological Science*, 28(11):1698–1701. 140
- Scheibehenne, B., Jamil, T., and Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, 27(7):1043–1046. 115, 118, 130, 140
- Scheibehenne, B. and Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, 22:391–407. 173
- Schönbrodt, F. D. and Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*. Manuscript accepted for publication. 132
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464. 206, 231
- Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference*, 136(7):2144–2162. 268
- Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619. 43, 268
- Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71. 120
- Senn, S. (2009). Comment. *Statistical Science*, 24(2):185–186. 11
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. 79
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423. 258
- Shiffrin, R. M., Chandramouli, S., and Grünwald, P. D. (2016). Bayes factors, relations to minimum description length, and overlapping model classes. *Journal of Mathematical Psychology*, 72:56–77. 50, 63
- Shiffrin, R. M., Lee, M. D., Kim, W., and Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32:1248–1284. 172, 173
- Shiue, W.-K. and Bain, L. J. (1982). Experiment size and power comparisons for two-sample Poisson tests. *Applied Statistics*, 31(2):130–134. 102
- Sinharay, S. and Stern, H. S. (2005). An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 14:415–435. 192

REFERENCES

- Steegen, S., Dewitte, L., Tuerlinckx, F., and Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, 5:786. 87, 88, 89
- Steingroever, H., Pachur, T., Smíra, M., and Lee, M. D. (2016a). Bayesian techniques for analyzing group differences in the Iowa gambling task: A case study of intuitive and deliberate decision makers. *Psychonomic Bulletin & Review*. Manuscript accepted for publication. 194, 195
- Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., and Wagenmakers, E.-J. (2013a). Performance of healthy participants on the Iowa gambling task. *Psychological Assessment*, 25:180–193. 194
- Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2013b). A comparison of reinforcement-learning models for the Iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, 5:Article 2. 194, 195
- Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2013c). Validating the PVL-Delta model for the Iowa gambling task. *Frontiers in Psychology*, 4: 898. 194
- Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa gambling task. *Decision*, 1:161–183. 194, 195, 196
- Steingroever, H., Wetzels, R., and Wagenmakers, E.-J. (2016b). Bayes factors for reinforcement-learning models of the Iowa Gambling Task. *Decision*, 3(2):115. 49, 193, 194, 195, 197, 198, 199, 200
- Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10:681–690. 43
- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64(3):153–181. 213
- Stigler, S. (1973). Studies in the history of probability and statistics. XXXII Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60(3):439–445. 216
- Stigler, S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Belknap Press. 225
- Stigler, S. (2007). The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620. 149
- Stirzaker, D. (2000). Advice to hedgehogs, or, constants can vary. *The Mathematical Gazette*, 84(500):197–210. 102
- Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *The Annals of Statistics*, 25:1371–1470. 203
- Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777. 90, 91, 95
- Stulp, G., Buunk, A. P., Verhulst, S., and Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, 24(1):159–171. 30, 31, 32, 33
- Sun, D. and Berger, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8:525–562. 151

- Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, New York. 87
- Swirles, B. (1991). Harold Jeffreys: Some reminiscences. *Chance*, 4(2):22–23, 26.
- Temme, N. M. (2003). Large parameter cases of the Gauss hypergeometric function. *Journal of Computational and Applied Mathematics*, 153(1):441–462. 110, 112
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, pages 1701–1728. 156
- Trafimow, D. and Marks, M. (2015). Editorial. *Basic And Applied Social Psychology*, 37:1–2. 120
- Tribus, M. and McIrvine, E. C. (1971). Energy and information. *Scientific American*, 225(3):179–188. 258
- Turner, B. M., Sederberg, P. B., and McClelland, J. L. (2016). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*, 72:191–199. 49
- van der Pas, S. and Grünwald, P. D. (2014). Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in single parameter model selection. *arXiv preprint arXiv:1408.5724*. 242
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. 119, 228, 242, 248, 263
- van der Vaart, A. W. (2002). The statistical work of Lucien Le Cam. *Annals of Statistics*, pages 631–682. 248
- van Doorn, J., Ly, A., Marsman, M., and Wagenmakers, E.-J. (2017). Bayesian estimation of Kendall’s tau using a latent normal approach. *arXiv preprint arXiv:1703.01805*. 268
- van Erven, T., Grünwald, P., and De Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):361–417. 37, 242
- van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820. 260
- van Ommen, T., Koolen, W. M., Feenstra, T. E., and Grünwald, P. D. (2016). Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57. 249
- Vandekerckhove, J., Matzke, D., and Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In Busemeyer, J., Townsend, J., Wang, Z. J., and Eidels, A., editors, *Oxford Handbook of Computational and Mathematical Psychology*. Oxford: Oxford University Press. 179
- Vandekerckhove, J., Rouder, J. N., and Kruschke, J. K. (2017). Beyond the new statistics: Bayesian inference for psychology [special issue]. *In preparation for Psychonomic Bulletin & Review*. 124
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54:491–498. 36, 99

REFERENCES

- Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, 72:183–190. 171
- Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618. 203
- Verhagen, A. J., Levy, R., Millsap, R. E., and Fox, J.-P. (2015). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, 72:171–182. 171
- Verhagen, A. J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475. 47, 87, 115, 119, 127, 129, 130, 137, 138, 140
- Vul, E. and Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19:645–647. 87, 88
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, 14:779–804. 58
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupiáñez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Özdogru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiantsev, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., and Zwaan, R. A. (2016a). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11:917–928. 90, 91, 95
- Wagenmakers, E.-J., Grünwald, P. D., and Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50:149–166. 37, 121, 242
- Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., and Morey, R. D. (2014). Another statistical paradox, or why intervals cannot be used for model comparison. *Manuscript submitted for publication*. 36
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60:158–189. 22, 133, 193, 203
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derkx, K., Rouder, J. N., and Morey, R. D. (2017a). Bayesian statistical inference for psychological science. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*. Manuscript accepted for publication. 126, 132

- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., and Morey, R. D. (2017b). Bayesian statistical inference for psychological science. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*. Manuscript accepted for publication. 124, 132
- Wagenmakers, E.-J., Morey, R. D., and Lee, M. D. (2016b). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25:169–176. 116, 124, 126, 131
- Wagenmakers, E.-J., Verhagen, A. J., and Ly, A. (2016c). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, pages 413–426. 34, 115, 119, 130
- Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., and Morey, R. D. (2015a). A power fallacy. *Behavior Research Methods*, 47:913–917. 116
- Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., and Morey, R. D. (2017c). The need for Bayesian hypothesis testing in psychological science. In Lilienfeld, S. O. and Waldman, I., editors, *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, pages 123–138. John Wiley and Sons. 14, 116
- Wagenmakers, E.-J. and Waldorp, L. (2006a). Editors’ introduction. *Journal of Mathematical Psychology*, 50:99–100. 172
- Wagenmakers, E.-J. and Waldorp, L. (2006b). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, 50(2). 37, 232
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R., and van der Maas, H. L. J. (2015b). A skeptical eye on psi. In May, E. and Marwaha, S., editors, *Extrasensory Perception: Support, Skepticism, and Science*, pages 153–176. ABC-CLIO. 117, 135
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100:426–432. 43
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7:627–633. 117
- Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, pages 165–205. 243
- Waldorp, L. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *International Journal of Biomedical Imaging*, 2009:723912. 242
- Waldorp, L., Christoffels, I., and van de Ven, V. (2011). Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions. *NeuroImage*, 54(4):2695–2705. 242
- Waldorp, L., Huizenga, H., and Grasman, R. (2005). The Wald test and Cramér–Rao bound for misspecified models in electromagnetic source analysis. *IEEE Transactions on Signal Processing*, 53(9):3427–3435. 242
- Wang, L. and Meng, X.-L. (2016). Warp bridge sampling: The next generation. *arXiv preprint arXiv:1609.07690*. 187

REFERENCES

- Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer Science and Business Media, New York. 69
- Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. 124
- Wetzels, R., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2010a). An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis*, 54:2094–2102. 63, 193, 203
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6:291–298. 15, 22, 82, 95, 96, 97, 120
- Wetzels, R., Raaijmakers, J. G., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, 16(4):752–760. 24, 84, 99
- Wetzels, R., Tutschkow, D., Dolan, C., van der Sluis, S., Dutilh, G., and Wagenmakers, E.-J. (2016). A Bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, 72:200–209. 171
- Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., and Wagenmakers, E.-J. (2010b). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, 54:14–27. 173, 201
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25. 238, 242
- Whittaker, E. (1902). On the functions associated with the parabolic cylinder in harmonic analysis. *Proceedings of the London Mathematical Society*, 1(1):417–427. 157
- Wijsman, R. (1973). On the attainment of the Cramér–Rao lower bound. *The Annals of Statistics*, 1(3):538–542. 249
- Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence*, 15:223–228. 76, 78
- Witte, E. H. and Zenker, F. (2016). Reconstructing recent work on macro-social stress as a research program. *Basic and Applied Social Psychology*, 38(6):301–307. 4, 115, 116, 117, 118, 119, 120, 121, 301
- Worthy, D. A. and Maddox, W. T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W. K. Estes. *Journal of Mathematical Psychology*, 59:41–49. 195
- Worthy, D. A., Pang, B., and Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, 4. 195
- Wrinch, D. and Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, 38:715–731. 49, 225
- Wrinch, D. and Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42:369–390. 13, 49, 83, 226
- Wrinch, D. and Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 45:368–375. 49, 226

- Wu, H., Myung, I. J., and Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, 17:275–286. 232
- Yang, G. L. (1999). A conversation with Lucien Le Cam. *Statistical Science*, pages 223–241. 248
- Yang, G. L. and Le Cam, L. (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, Berlin. 46, 242
- Yuan, Y. and Johnson, V. E. (2008). Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, 18:1185–1200. 71, 72
- Zellner, A. (1980). Introduction. In Zellner, A., editor, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 1–10. North-Holland Publishing Company, Amsterdam, The Netherlands. 11
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243. 24
- Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, Jose, M., DeGroot, Morris, H., Lindley, Dennis, V., and Smith, Adrian, F., editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, volume 1, pages 585–603. Springer. 24, 25
- Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, B. M. (2017). Making replications mainstream. *Behavioral and Brain Sciences*, pages 1–50. 129

Nederlandse Samenvatting

De Bayesiaanse hypothese toetsen die wij ontwikkeld hebben zijn bedoeld om empirische onderzoekers te helpen met (i) het kwantificeren van evidentie voor of tegen een hypothese, en (ii) het leren en construeren van (statistische) modellen en theorieën op basis van geobserveerde data.

Een statistisch model geeft een versimpelde beschrijving van de werkelijkheid met een mathematische relatie $f(d|\theta)$ tussen observaties, de data, d en *parameters* θ . Zo kan d bijvoorbeeld refereren naar bloeddrukmetingen voor en na een behandeling van een steekproef van patiënten, θ refereert dan naar de effectgrootte, en f is in de meeste gevallen een normale verdeling om er rekening mee te houden dat de metingen slechts een steekproef zijn uit een grotere populatie patiënten.

Om te toetsen of de behandeling effectief is vergelijken we het *nul model* \mathcal{M}_0 , het statistisch model waarbij de effectgrootte op nul wordt gezet $\theta = 0$, met het alternatief model \mathcal{M}_1 , het model waarin de effectgrootte elke reële waarde kan aannemen.

De *a priori plausibiliteit* van de effectiviteit hangt af van welke behandeling de patiënt krijgt voorgeschreven. De *a priori* kans dat de behandeling effectief is, is relatief hoog, zeg, $P(\mathcal{M}_1) = 0.9$ en $P(\mathcal{M}_0) = 0.1$, wanneer de patiënten worden voorgeschreven om pillen in te nemen met een actieve stof ontwikkeld om bloeddruk te verlagen. Een ander, maar equivalente, manier om deze *a priori* model kansen te beschrijven is met behulp van de *a priori model kansverhouding*, in dit geval, negen staat tot één, dus $\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} = 9$. Op basis van de geobserveerde data kunnen we de *a priori* model kansverhouding bijwerken tot *a posteriori* model kansverhouding $\frac{P(\mathcal{M}_1|d)}{P(\mathcal{M}_0|d)}$ met behulp van de *regel van Bayes* en leidt tot de volgende cruciale vergelijking:

$$\frac{P(\mathcal{M}_1|d)}{P(\mathcal{M}_0|d)} = \underbrace{\frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_0)}}_{\text{BF}_{10}(d)} \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}, \quad (.0.1)$$

waar $P(\mathcal{M}_i|d)$ refereert naar de *a posteriori model kans* van \mathcal{M}_i gegeven de observaties, en $p(d|\mathcal{M}_i)$ refereert naar de marginale waarschijnlijkheid van model \mathcal{M}_i . De term $\text{BF}_{10}(d)$ is de zogeheten factor van Bayes, oftewel, *Bayes factor*,

en beschrijft hoe de a priori model kansverhouding wordt bijgewerkt tot de a posteriori model kansverhouding gegeven de observaties d .

De Bayes factor is makkelijk interpreteerbaar: $\text{BF}_{10}(d) = 7$ indiceert dat de observaties 7 keer zo waarschijnlijk zijn onder \mathcal{M}_1 als onder \mathcal{M}_0 , en $\text{BF}_{10}(d) = .2$ indiceert dat de observaties 5 keer zo waarschijnlijk zijn onder \mathcal{M}_0 als onder \mathcal{M}_1 . Gegeven de observaties is de Bayes factor $\text{BF}_{10}(d)$ altijd een niet-negatief getal en hoe hoger (lager) dit getal, hoe meer (minder) evidentie er is voor \mathcal{M}_1 ten opzichte van \mathcal{M}_0 . Op een gelijksoortige manier, wanneer ook de activiteitsniveau van de patiënten is gemeten, kunnen we toetsen of de behandeling mensen moe maakt. Op deze manier krijgen we geleidelijk meer inzicht in de effecten van de behandeling op de populatie van patiënten.

De Bayes factor is een ratio van de marginale waarschijnlijkheid $p(d | \mathcal{M}_i)$ die aangeeft hoe goed het model op de geobserveerde data past. Deze marginale waarschijnlijkheid wordt berekend door de relatie $f_i(d | \theta)$ van model \mathcal{M}_i gegeven de observaties d te evalueren op elke mogelijke parameter waarde θ en te middelen ten opzichte van een *a priori verdeling* $\pi_i(\theta)$:

$$p(d | \mathcal{M}_i) = \int f_i(d | \theta) \pi_i(\theta) d\theta. \quad (.0.2)$$

Gegeven twee modellen, dus, de relaties $f_1(d | \theta)$ en $f_0(d | \theta)$, is het de taak van de statisticus om twee a priori delingen, namelijk, $\pi_0(\theta)$ en $\pi_1(\theta)$ te kiezen om daarmee een Bayes factor te construeren. Voor een gebruiksvriendelijke Bayes factor moet de statisticus er ook voor zorgen dat deze uit te rekenen is voor elke data set d . In dit proefschrift beschreven we hoe men a priori delingen voor Bayes factoren moet selecteren en berekenen. De resulterende Bayes factoren zijn of worden nog geïmplementeerd in het gratis software-pakket vernoemd naar Harold Jeffreys, *Jeffreys's Amazing Statistics Program*, JASP, (url: <https://jasp-stats.org/>, JASP Team, 2017).

Deel I. De onderliggende principes van de Bayes factor

Het eerste gedeelte van dit proefschrift richtte zich op de filosofie, de motivering en de constructie van zogeheten *Jeffreys's Bayes factoren*.

In Hoofdstuk 2 bespraken we de onderliggende principes van de Bayes factor, hoe deze te interpreteren, en gaven we een beschrijving van de algemene constructie waarmee Jeffreys a priori delingen selecteerde voor Bayes factoren. In deze constructie is het van belang om een Bayes factor te ontwerpen dat *predictief geijkt* en *informatie consistent* is. Een predictief geikte Bayes factor is één wanneer de steekproefgrootte te klein en daarom ambigu is, terwijl een informatie consistente Bayes factor oneindig is wanneer de observaties overweldigend wijzen naar het bestaan van een effect. De constructie waarmee Jeffreys Bayes factoren ontwerpt is ontleend uit hoe hij zijn Bayesiaanse *t*-toets opzet. Deze constructie hebben we gebruikt om een Jeffreys's Bayes factor af te leiden voor de product-moment correlatiecoëfficiënt van Pearson. De resulterende Bayes factor is analytisch en makkelijk te gebruiken.

In Hoofdstuk 3 reageren wij op twee discussie artikelen op ons werk over Harold Jeffreys. In dit hoofdstuk belichtten wij de zogeheten *Jeffreys-Lindley-Bartlett*

paradox, het verschil tussen inferentie en besluitvorming, en het verschil tussen schatten en toetsen toe.

Deel II. Bayes factoren voor veelgebruikte statistische analyses

Het tweede deel van dit proefschrift richtte zich op Bayes factoren die wij geconstrueerd hebben voor bepaalde veelgebruikte statistische analyses.

In Hoofdstuk 4 zetten we een Bayesiaanse methode uiteen voor het schatten en toetsen van de rangcorrelatiecoëfficiënt τ van Kendall. Voor deze methode modelleerden we de toets statistiek die we daarna gecombineerd hebben met het analytische resultaat voor de correlatiecoëfficiënt van Pearson.

In Hoofdstuk 5 hebben we de afleiding van het analytische resultaat voor de correlatiecoëfficiënt van Pearson gebruikt om een geïnformeerde Bayesiaanse t -toets te construeren. De klasse van a priori verdelingen die wij hiervoor gebruikten is een veralgemenisering van de verdelingen die Harold Jeffreys aandroeg voor dit probleem, maar laat de locatie en schaal op de a priori verdeling van de effectgrootte vrij. Hierdoor kunnen onderzoekers wanneer ze substantiële voorkennis hebben deze gebruiken in hun t -toetsen.

In Hoofdstuk 6 introduceerden we *limiet-consistentie* als een desideratum voor het selecteren van a priori verdelingen voor twee-steekproef toetsen. Voor dit desideratum bekijken we de hypothetische scenario waarin de dataverzameling voor een proces vroegtijdig wordt beëindigd, terwijl de dataverzameling van het tweede proces voor een onbepaalde tijd doorgaat. In dergelijke gevallen zou de Bayes factor moeten convergeren naar een eindige limiet. We constateren dat de Bayes factor die Jeffreys voorstelde voor het twee-steekproef Poisson probleem limiet-inconsistent is. Als oplossing generaliseren wij de Bayes factor van Jeffreys zodat deze wel limiet-consistent is.

Deel III. Wetenschappelijk kennis vergaren met Bayes factoren

Het derde deel van dit proefschrift richtte zich op het gebruik van Bayes factoren in de empirische wetenschappen als een instrument voor wetenschappelijk leren. In het bijzonder bespreken wij de rol van de Bayes factor in de “replicatie- en reproduceerbaarheidscrisis” (Baker, 2016, Levert et al., 2012, Pashler and Wagenmakers, 2012).

In Hoofdstuk 7 bespraken wij kort hoe psychologen zich hebben ingezet om de reproduceerbaarheid van het veld te vergroten met grootschalige replicatie initiatieven, zoals het “Reproducibility Project: Psychology” (Open Science Collaboration, 2015), de speciale replicatie editie van *Social Psychology* (Nosek and Lakens, 2014) en de vele ManyLabs experimenten (Ebersole et al., 2016; Klein et al., 2014). Dit hoofdstuk is een commentaar op het werk van Witte and Zenker (2016). Zij beweren dat een “ander” gebruik van standaard statistische methoden op basis van p -waarden een oplossing is voor de replicatie- en reproduceerbaarheids crisis. Ons standpunt is dat deze crisis veel omvattender is dan een discussie over de statistische methoden. Wij pleiten er namelijk voor om confirmatieve studies te preregistreren. Door te preregistreren worden termen beter gedefinieerd en vermijdt men het probleem van achteraf kanskapitalisatie. Daarnaast vinden

wij dat wetenschap open en transparant moet zijn waarbij onzekerheid gerapporteerd wordt, omdat dit een betere en eerlijke beeld geeft van het wetenschappelijke proces.

In Hoofdstuk 8 beschreven wij het gemak waarmee men een Bayesiaanse heranalyse kan doen, zelfs wanneer de volledige dataset niet beschikbaar is. Dit is relevant voor onderzoekers die naast p -waarden ook een Bayes factor willen rapporteren. Een Bayesiaanse heranalyse is ook handig voor redacteuren, recensenten, lezers en verslaggevers, omdat zij in één oogopslag de evidentie kunnen bepalen in gerapporteerde statistieken. Daarnaast demonstreren we hoe gevoelig de evidentie is voor veranderingen in de a priori verdelen door middel van een robuustheidsanalyse. De Bayesiaanse heranalyse leidt ook tot een a posteriori verdeling waaruit men kan concluderen welke gebieden in de parameterruimte plausibeler worden nadat we de observaties in ogenschouw nemen. Als laatste bespraken wij hoe de a posteriori verdeling gebruikt kan worden als voorkennis in een vervolgstudie.

In Hoofdstuk 9 bespraken wij een algemene methode om de evidentie te extraheren uit de observaties van een directe replicatiepoging gegeven de observaties van een oorspronkelijke studie. Deze algemene methode is ontworpen om onderzoekers te helpen modellen te bouwen en kennis te vergaren uit een groeiend aantal replicatiestudies.

Deel IV. Analytische resultaten

Het vierde deel van dit proefschrift richtte zich op verschillende analytische resultaten die zijn gebruikt voor de constructie van de Bayesiaanse toetsen in dit proefschrift.

In Hoofdstuk 10 leidden wij de analytische a posteriori verdeling af voor een grote klasse van a priori verdelen op de product-moment correlatiecoëfficiënt van Pearson. Dit resultaat is gebruikt voor de analytische Bayes factor in Hoofdstuk 2 en de afleiding vormt de basis van Hoofdstuk 4 en 5.

In Hoofdstuk 11 leidden wij analytische a posteriori verdelingen af voor modellen met discrete data. Het eerste resultaat is gebruikt in Hoofdstuk 6 om een limiet-consistente Bayes factor te construeren voor het twee-steekproef Poisson probleem. Dit resultaat kan ook gebruikt worden om een robuustheidsanalyse te definiëren voor een binomiaal toets. Daarnaast bevat dit hoofdstuk analytische uitdrukkingen voor de eenzijdige binomiaal Bayes factoren. Het laatste resultaat is een analytische uitdrukking voor de ratio van kansverhoudingen in een 2-keer-2 contingentie tabel.

Deel V. Twee handleidingen

Het vijfde en laatste deel van dit proefschrift richtte zich op hulpmiddelen bij het construeren van Bayes factoren en biedt verdieping in mathematische statistisch modelleren.

In Hoofdstuk 12 legden wij uit hoe *bridge sampling* (Meng and Wong, 1996) gebruikt kan worden om uitkomsten van een MCMC-procedure te transformeren in een schatting van de marginale waarschijnlijkheid van een model. De bridge

sampler is relevant voor complexe modellen met hiërarchische structuren die lastig te beschrijven zijn met standaard wiskundige functies.

In Hoofdstuk 13 gaven we een algemene beschrijving van mathematische statistiek en de rol van Fisher informatie voor statistische modellen in het bijzonder. In het frequentistische paradigma werd uiteengezet hoe men hypothese toetsen en betrouwbaarheidsintervallen kan construeren door Fisher informatie te combineren met maximale waarschijnlijkheidsschatters. In het Bayesiaanse paradigma werd uiteengezet hoe men een standaard *a priori* verdeling kan construeren uit Fisher informatie. In het minimale beschrijvingslengte paradigma werd uiteengezet hoe Fisher informatie gebruikt wordt om de mate van model complexiteit te beschrijven. De resultaten hangen af van bepaalde regulariteitscondities die gegeven zijn in de appendix. Wanneer men modellen construeert die voldoen aan deze condities zullen de standaard statistische methoden (asymptotisch) geldig zijn.

Acknowledgements — Dankwoord

This dissertation would not have been completed without the help and support of many people. First and foremost, I would like to thank my supervisor Eric-Jan Wagenmakers, co-supervisor Maarten Marsman, the members of my defence committee: Jim Berger, Michael Lee, Peter Grünwald, Han van der Maas, Lourens Waldorp, and Raoul Grasman, as well the European Research Council who provided financial support for this project.

EJ, thank you for the interesting discussions, your guidance, generosity, and your optimistic demeanour with which you lead an ever expanding lab. With such a stimulating and inspiring surrounding, it was easy for me to grow as an independent researcher. I am proud to call myself a member of your lab, and I am grateful for the opportunity to have led the JASP team whenever you were too busy.

Marsman, thank you for being the unanticipated voice of reason, for calming me down at the right moment, and for being a great example of perseverance and intelligence. Dora, your cheerful presence as well as your role as the glue that binds the lab together made it easy for me to adapt to the group. It is a great honour to have you and Sacha as my paranympths. Sacha, thank you for being a great friend, for dragging me to the gym, and the best King's days. Helen, for introducing new phrases into the Dutch language, and for honouring me with a star for good behaviour. Udo, for your display of strength by doing both a PhD and mathematics bachelor at the same time. Johnny, for being so chilled out with all the pressure I put on you, and your ability to surpass yourself every time. Quentin, thank you for being my external memory, for making sense of my equations, and for building on them with great success. Etz, for being so expressive, and infectiously happy and free whenever you are around.

Bruno, thank you for being such a great down-to-earth lead programmer and for coping with my stupidity once every while. Frans, for being so patient with me, and for sharing your Gall and Gall card. Tim de Jong, thank you for your unbreakable (I tried) spirit, your snide remarks, and for totally exceeding all expectations. Akash, thank you for being such a miracle worker and for making my work as CTO so much easier. Erik-Jan, for your enthusiasm, realism, and for being so knowledgeable. Koen, for doing much of the leg work in our software

project, and for your drive for learning. Alexandra, for your creativity and for keeping Marsman at bay. Tim Draws, for the passion with which you make our work known to the general public. Don, for your courage by going out of your comfort zone to solve any problem. Louise, for always being so helpful and the best gossips of course. Michael, thank you for being a laksa regular, a friendly latent supervisor around which we plan our best activities.

Our lab could not thrive without the leadership of Han and Denny. I benefitted from interactions with members of their groups as well. Jonas, Lourens, Riet, Sharon, Claire, Gilles, Laura, Sacha, Udo, Johnny, Quentin, Adela, Raoul, Tahira, Pia, Don, Koen, Claudia, Joost, thanks for your active participation in my statistical learning reading seminar. It is a great pleasure to broaden my horizons with such a motivated group of inquisitive young researchers.

In addition to having a great group to work with at the Psychological Methods department, I was fortunate enough to have the opportunity to also explore other research interests in mathematical statistics, causality, and machine learning with other groups. Bas, Aad, Harry, Botond, Bartek, Frank, Paulo, Eduard, Richard, Paulo, Moritz, Shota, Joris, Johannes, Tim, Weimin, Jakob, Kolyan, JB, Stéphanie, Fengnan, Kevin, Suzanne, Dong, Gino, Nurzhan, Jarno, Jan, and Alice, thanks for giving me my mathematics fix at the Bayes Club, Statistics for Structures, or at the Van Dantzig seminar. It is an honour and a pleasure to (Bayes) club with you. Joris, Stephan, Philip, Sara, Patrick, Thijs, Mijung, Tom, and Tineke, thank you for letting me invade your weekly meetings, and for being highly correlated to all that I know about causality. Peter, Wouter, Nishant, Rianne, Tom, Judith, Allard, and Alice, thank you for welcoming in your group and teaching me about statistical learning.

My interest in research was mostly sparked by André Schram, Chris Klaassen, Henk Pijls, Ale Jan Homburg and Bas Kleijn, who mentored and inspired me during my time as a student. During my PhD I always felt the warm support from my friends, family and my dearest (the bestest) Gracia. The time I spent with friends such as Huub, Lai, Yaron, Fleur, Mark, Evan, Anannd, Vincent, Farhaz, Freek and Johan during this time were brief, but always well-appreciated. Words cannot express the amount of help I received from my grandfather, parents and sister, and I will be forever grateful for all the sacrifices they made on my behalf. They have always been a great example of diligence, hard work, and loving kindness. Gracia offered me love, care, support, great dinners, which kept on me on track to finish my PhD, and her presence makes everything better and so much more fun.

I am grateful to all co-authors who have contributed to the papers that form the basis of this dissertation. This thesis would have been impossible without your ideas, knowledge, and advice. It goes without saying that any remaining errors in this dissertation are all mine.

Publications

1. Wagenmakers, E.-J., Verhagen, A. J., **Ly, A.**, Bakker, M., Lee, M. D., Matzke, D., & Rouder, J. N. (2015). A power fallacy. *Behavior Research Methods*, 47, 913–917.
2. Love, J.P., Selker, R., Verhagen, A.J., Marsman, M., Gronau, Q.F., Jamil, T., Šmíra, M., Epskamp, S., Wild, A., **Ly, A.**, Matzke, M., Morey, R.D., Rouder, J.N. & Wagenmakers, E.-J. (2015). Software to sharpen your stats. *APS Observer*, 28, 27–29.
3. Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., **Ly, A.**, Verhagen, A. J., Selker, R., Sasiadek, A., Gronau, Q. F., Love, J., & Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology: Cognition*, 6, 494.
4. Love, J., Selker, R., Marsman, M., Jamil, T. Dropmann, D., Verhagen, A.J., **Ly, A.**, Gronau, Q.F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J.N., Morey, R.D., & Wagenmakers, E.-J. (2015). JASP – Graphical statistical software for common statistical designs. Manuscript accepted for publication.
5. Tierney, W., Schweinsberg, M., Jordan, J., Kennedy, D., Qureshi, I., Sommer, S. A., Thornley, N., Madan, N., Vianello, M., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C., du Plessis, C., Gronau, Q. F., Hafenbrack, A., Liao, E., **Ly, A.**, Marsman, M., Murase, T., Schaerer, M., Tworek, C., Wagenmakers, E.-J., Wong, L., Anderson, T., Bauman, C., Bedwell, W., Brescoll, V., Canavan, A., Chandler, J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P., Amell, A., Frick, S., Gamez-Djokic, M., Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B., Hartwich, N., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Legate, N., Luoma, T., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J., Sackett, A., Sellier, A.-L., Sokolova, T., Sowden, W., Storage, D., Sun, X.,

- van Bavel, J., Washburn, A., Wei, C., Wetter, E., Wilson, C., Darroux, S.-C., & Uhlmann, E. (2016). Data from a pre-publication independent replication initiative examining ten moral judgement effects. *Scientific Data*, 3, 160082.
6. Wagenmakers, E.-J., Verhagen, A. J., & **Ly, A.** (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48(2), 413–426.
 7. **Ly, A.**, Verhagen, A. 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.
 8. **Ly, A.**, Verhagen, A. J., & Wagenmakers, E.-J. (2016). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.
 9. Marsman, M., **Ly, A.**, & Wagenmakers, E.-J. (2016). Four requirements for an acceptable research program. *Basic and Applied Social Psychology*, 38(6), 308–312.
 10. van Doorn, J., **Ly, A.**, Marsman, M., & Wagenmakers, E.-J. (2016). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*.
 11. Dutilh, G., Vandekerckhove, J., **Ly, A.**, Matzke, D., Pedroni, A., Frey, R., Rieskamp, J., & Wagenmakers, E.-J. (2017). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics*, 79(3), 713–725.
 12. Wagenmakers, E.-J., Verhagen, A. J., **Ly, A.**, Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In Lilienfeld, S. O., & Waldman, I. (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, 123–138, John Wiley and Sons.
 13. **Ly, A.**, Marsman, M., & Wagenmakers, E.-J. (2017). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*.
 14. Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., **Ly, A.**, Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E.-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibecher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman,

- G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C., Darroux, S-C., & Uhlmann, E. L. (2017). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*. Manuscript accepted for publication.
15. Jamil, T., Marsman, M., **Ly, A.**, Morey, R. D., & Wagenmakers, E.-J. (2017). What are the odds' Modern relevance and Bayes factor solutions for MacAlister's problem from the 1881 Educational Times. *Educational and Psychological Measurement*, 77(5), 819–830.
 16. Jamil, T., **Ly, A.**, Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, 49(2), 638–652.
 17. Wagenmakers, E.-J., Marsman, M., Jamil, T., **Ly, A.**, Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*. Manuscript accepted for publication.
 18. Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., **Ly, A.**, Verhagen, A. J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., Rouder, J. N., & Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*. Manuscript accepted for publication.
 19. Keuken, M. C., **Ly, A.**, Boekel, W. E., Wagenmakers, E.-J. , Belay, L., Verhagen, A. J., Brown, S. D., & Forstmann, B. U. (2017). Corrigendum for: A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, 93, 229–233.
 20. **Ly, A.**, Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80, 40–55.
 21. Matzke, D., **Ly, A.**, Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, 3(1), 25.
 22. van Doorn, J., **Ly, A.**, Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian estimation of Kendall's tau using a latent normal approach. Manuscript submitted for publication.
 23. Gronau, Q. F., Sarafoglou, A., Matzke, D., **Ly, A.**, Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80–97.

24. Gronau, Q. F., **Ly, A.**, & Wagenmakers, E.-J. (2017). Informed Bayesian *t*-tests. Manuscript submitted for publication.
25. **Ly, A.**, Raj, A., Etz, A., Marsman, M., Gronau, Q. F., & Wagenmakers, E.-J. (2017). Bayesian reanalyses from summary statistics: A guide for academic consumers. Manuscript submitted for publication.
26. **Ly, A.**, Etz, A., Marsman, M., & Wagenmakers, E.-J. (2017). Replication Bayes factors from evidence updating. Manuscript submitted for publication.
27. **Ly, A.**, Raj, A., Marsman, M., & Wagenmakers, E.-J. (2017). A limit-consistent Bayes factor for testing the equality of two Poisson rates. Manuscript in preparation.
28. **JASP team** (2017). JASP (Version 0.8.4.0)[Computer software].
29. **Ly, A.**, Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (2017). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In Mousstafa, A. A., (Ed.), *Computational Models of Brain and Behavior*, 467–480.