Bayes factor hypothesis tests for ranks and researchers

van Doorn, J.B.

Publication date
2021

Document Version
Final published version

Bayes Factor Hypothesis Tests

For Ranks and Researchers

Johnny van Doorn

Bayes Factor Hypothesis Tests
For Ranks and Researchers

Johnny Boy van Doorn

Cover design: Viktor Beekman, @viktordepictor

This thesis was typeset using LaTeX and the thesis style design is based on the style design developed by Martijn Wieling (www.martijnwieling.nl).

Bayes Factor Hypothesis Tests
For Ranks and Researchers

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 Agnietenkapel
op donderdag 17 juni 2021, te 16:00 uur
door Johnny Boy van Doorn
geboren te Eindhoven

Promotiecommissie

Promotor:	Prof. dr. E.M. Wagenmakers	Universiteit van Amsterdam
Copromotores:	Dr. M. Marsman	Universiteit van Amsterdam
	Dr. A. Ly	Universiteit van Amsterdam
Overige leden:	Prof. dr. H.L.J. van der Maas	Universiteit van Amsterdam
	Dr. D. Matzke	Universiteit van Amsterdam
	Prof. dr. D. Borsboom	Universiteit van Amsterdam
	Prof. dr. I.G. Klugkist	Universiteit Utrecht
	Prof. dr. F. Tuerlinckx	Katholieke Universiteit Leuven
	Prof. dr. M.D. Lee	University of California Irvine

Faculteit der Maatschappij- en Gedragswetenschappen

Contents

1 Introduction 1
1.1 Two Friends Playing Mario Kart . . . 1
1.2 Chapter Outline . . . 6
1.2.1 Part I: For Researchers . . . 6
1.2.2 Part II: For Ranks . . . 7

I For Researchers 9

2 The JASP Guidelines for Conducting and Reporting a Bayesian Analysis 11
2.1 Introduction . . . 11
2.2 Stage 1: Planning the Analysis . . . 12
2.3 Stage 2: Executing the Analysis . . . 17
2.4 Stage 3: Interpreting the Results . . . 20
2.5 Stage 4: Reporting the Results . . . 25
2.6 Limitations and Challenges . . . 27
2.7 Concluding Comments . . . 28

3 Multiple Perspectives on Inference for two Simple Statistical Scenarios 31
3.1 Introduction . . . 31
3.2 Data Set I: Birth Defects and Cetirizine Exposure . . . 33
3.3 Data Set II: Amygdalar Activity and Perceived Stress . . . 40
3.4 Round-Table Discussion . . . 46
3.5 Concluding Comments . . . 50

4 An In-Class Demonstration of Bayesian Inference 53
4.1 Introduction . . . 53
4.2 The Tasting Experiment . . . 54
4.3 Theoretical Analysis . . . 55
4.4 Bayesian Inference with JASP . . . 57
4.5 Concluding Comments . . . 60

5 Strong Public Claims May Not Reflect Researchers' Private Convictions 63

6 Bayes Factors for Mixed Models 67
6.1 Introduction . . . 67
6.2 The Candidate Models . . . 70

6.2.1 The Model Comparisons 72
6.3 Examples 73
6.3.1 Example 1: The Effect of Aggregation 73
6.3.2 Example 2: The Effect of Measurement Error 76
6.3.3 Example 3: A Random Interaction Effect 80
6.4 Concluding Comments 84

II For Ranks 87

7 Bayesian Inference for Kendall's Rank Correlation Coefficient 89
7.1 Introduction 89
7.1.1 Kendall's τ 90
7.1.2 Modeling Test Statistics 91
7.1.3 Challenges 92
7.2 Methods 92
7.2.1 Defining T^* 92
7.2.2 Prior Distribution through Parametric Yoking 93
7.2.3 Posterior Distribution and Bayes Factor 94
7.2.4 Verifying the Asymptotic Normality of T^* 94
7.3 Results 95
7.3.1 Bayes Factor Behavior 95
7.3.2 Comparison to Pearson's ρ 95
7.3.3 Real Data Example 96
7.4 Concluding Comments 98

8 Bayesian Estimation of Kendall's τ Using a Latent Normal Approach 101
8.1 Introduction 101
8.2 Correction Using The Sample τ 102
8.3 Correction Using The Latent Normal Approach 103
8.3.1 Latent Normal Models 103
8.3.2 Posterior Distribution for the Latent Correlation 103
8.3.3 Relation to Kendall's τ 105
8.4 Results: Simulation Study 105
8.5 Concluding Comments 106

9 Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ 109
9.1 Introduction 109
9.2 General Methodology 111
9.2.1 Latent Normal Models 111
9.2.2 Posterior Distribution 112
9.2.3 Estimation and Testing 113
9.3 Case 1: Wilcoxon Rank Sum Test 113
9.3.1 Background 113
9.3.2 Sampling Algorithm 115
9.3.3 Simulation Study 116

9.3.4 Data Example 117
9.4 Case 2: Wilcoxon Signed Rank Test 119
9.4.1 Background 119
9.4.2 Sampling Algorithm 121
9.4.3 Simulation Study 123
9.4.4 Data Example 124
9.5 Case 3: Spearman's ρ_s 126
9.5.1 Background 126
9.5.2 Sampling Algorithm 127
9.5.3 Transforming Parameters 128
9.5.4 Simulation Study 128
9.5.5 Data Example 129
9.6 Concluding Comments 130

10 Using the Weighted Kendall's Distance to Analyze Rank Data in Psychology 133
10.1 Introduction 133
10.2 Kendall's Distance 134
10.2.1 Item Weights 135
10.2.2 Position Weights 136
10.2.3 Similarity Weights 137
10.2.4 Top-k Lists 137
10.3 Applications 138
10.3.1 Item Weights: Recall of Events on September 11 138
10.3.2 Position Weights: Month Preference 139
10.3.3 Similarity Weights: The Free Recall of Animals 143
10.3.4 Top-k: Expert Sporting Predictions 145
10.4 Concluding Comments 148

III Conclusion 149

11 Summary and Future Directions 151
11.1 Part I: For Researchers 151
11.1.1 Chapter Summaries 151
11.1.2 Discussion and Future Directions 153
11.2 Part II: For Ranks 154
11.2.1 Chapter Summaries 154
11.2.2 Discussion and Future Directions 156
11.3 General Conclusion 157

IV Appendices 159
References 161

A Online Supplementary Material 177

B Publications **181**
B.1 Under Review 181
B.2 Published 181

C Dutch summary **185**
C.1 Deel I: For Researchers 185
C.2 Deel II: For Ranks 187

D Acknowledgments **189**

Chapter 1

Introduction

1.1 Two Friends Playing Mario Kart

One faithful afternoon in June, 2016, rather than enjoying the sunny outside, me and my good friend Sjoerd were playing a video game called Mario Kart. In this game, 12 cartoon racers battle for first place in various outlandish racing circuits, with enticing names such as “Peach Beach” and “Rainbow Road”. We had been playing this game for a while already, and had generally been enjoying it. However, we could not agree as to who was the better racer of the two. Luckily, I had some experience in statistics, and proposed to gather some data in order to settle the dispute. In the year that followed, we recorded our respective end positions (i.e., *ranks*) for a total of 332 races. Table 1.1 lists the results of eight races, to illustrate the structure of our data, where we refer to ourselves as Player 1 and Player 2. Because there were also 10 computer controlled racers, the rank for each race could vary between 1 (i.e., finishing first) and 12 (i.e., finishing last). Our primary interest was to figure out who finished higher than the other, regardless of whether a computer controlled character had beat us both.

Unfortunately, gathering the data did not settle the dispute by itself. A statistical analysis was needed that could test two competing hypotheses. As a starting point, we wanted to simply see whether there was a difference in our skill levels or not, and considered two hypotheses:

$$\mathcal{H}_0 : \text{We are equally skilled.}$$

$$\mathcal{H}_1 : \text{There is a difference in our skill levels.}$$

The most popular approach for comparing these hypotheses is to conduct a frequentist paired samples t-test. This is a procedure that takes the difference between the average value in one group of observations (i.e., the ranks of Player 1) and the average value in another group of observations (i.e., the ranks of Player 2). Then, it assesses whether this difference deviates substantially from 0 by means of a p-value: the probability of the observed difference, or an unobserved greater difference, *if* $\mathcal{H}_0$ is true. However, the fundamental flaws of this approach are twofold: it is based on the p-value and it is a t-test applied to rank data.

The flaws of the p-value are well-documented (e.g., Bayarri & Berger, 2004; R. L. Wasserstein & Lazar, 2016; Wagenmakers et al., 2017). For example, the p-value does not quantify evidence in favor of $\mathcal{H}_0$: there can only be absence of evidence for a difference in skill, rather than evidence of absence. An appealing

Table 1.1: The outcomes of eight Mario Kart races. In the first race, on the Maple Treeway circuit, Player 1 finished first, and Player 2 finished third. In the second race, on Toad Factory, Player 1 finished second, and Player 2 finished first. The end position in bold indicates which of the two players won that specific round. Since we were only interested in beating each other, the outcome on Mario Circuit, where Player 1 finished fourth and Player 2 finished fifth, is therefore counted as a win for Player 1.

Circuit	Player 1	Player 2
Maple Treeway	**1**	3
Toad Factory	2	**1**
DK Mountain	**1**	2
Shy Guy Beach	**3**	7
Delphino Square	2	**1**
Peach Beach	9	**7**
Mario Circuit	4	5
Bowser's Castle	3	**1**

alternative to the p-value is the Bayes factor, which directly pits the two hypotheses against each other and directly quantifies how likely the data are under one hypothesis, compared to the other hypothesis. It therefore allows evidence in favor of either hypothesis under consideration.

The inappropriateness of the t-test in this scenario comes from two assumptions made by this test. First, the t-test assumes the data to be normally distributed. Because the data at hand are integers, and are often equal to 1, 2, and 3 (we generally manged to beat the computer controlled racers), the data are not continuous and follow a skewed distribution. Especially if we would have few observations, violations of the normality assumption can be critical to the validity of the test. Second, the t-test assumes the data to be measured on the ratio or interval level, rather than the ordinal level. The ratio measurement level implies that a value of 2 is twice as high as a value of 1. For these data this would imply that, when Player 1 finishes third and Player 2 finishes first (see Bowser's Castle in Table 1.1), Player 2 is judged to be three times as skilled as Player 1. Although the outcome of this race is an indication that Player 2 is more skilled than Player 1, it is an extremely strong statement to speak of "three times as skilled". The interval measurement level implies that the difference in skill between the first place and second place is as big as the difference in skill between the ninth and tenth place. While this statement is not as strong as the ratio measurement level, the rank data alone cannot distinguish between a linear relationship of skill and rank, or an exponential relationship of skill and rank.

The equivalent of the t-test that does not make such strong assumptions, is the Wilcoxon signed rank test. Instead of comparing the average ranks of Player 1 to the average ranks of Player 2, this test considers whether the ranks of Player 1 are more often higher, or more often lower than the ranks of Player 2. Additionally, it also considers the magnitude of the differences in ranks. Thus, the most appropriate test that we could conduct to compare $\mathcal{H}_0$ and $\mathcal{H}_1$, is the Bayesian

Wilcoxon signed rank test.

The Bayesian framework is centered on the notion of knowledge updating. Before Sjoerd and I had met each other (played any game of Mario kart), it was very hard to tell whether one of us would be the better player, and to what extent that player would be better. In other words, there was little prior knowledge about hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, and about the size of the difference in our skill level, if there was any. After 332 games, however, there were many data points with which to update that prior knowledge, in order to form posterior knowledge. Humans experience this process of knowledge updating every day, and do this naturally. Statisticians, on the other hand, have the need to formalize this procedure. The formalization of updating knowledge with ordinal data is quite challenging, because rank data typically lack a likelihood function, which is required to compute the probability of observing the data, given a hypothesized value for our difference in ability. Since the likelihood function is a fundamental part of Bayesian inference, this has frustrated the development of Bayesian inference for rank data. In the hope of one day proving to be better than Sjoerd at Mario Kart, I centered my research around this development.

It would not only be Mario Kart players that would gain from the development of Bayesian hypothesis tests for ranks. In psychological science, rank data are ubiquitous. The most common form of rank data is the Likert scale (Likert, 1932). An example of a 5-point Likert scale is the common “strongly agree”, “agree”, “neutral”, “disagree”, “strongly disagree”, where respondents then indicate to what extent they agree or disagree with a certain statement, or series of statements. Similarly, it can be used to measure respondents’ levels of emotion, pain, level of well-being or depression, or degree of exhibiting a certain behavior. For all of these measurements, treating these observations on an ordinal measurement scale is again of crucial importance: if person A “strongly agrees”, person B “agrees”, and person C is “neutral”, it absolutely need not be the case that person A agrees twice as much with the statement as person B (i.e., ratio measurement level), or that person B’s level of agreement is exactly in the middle of the levels of agreement of Person A and person C (i.e., interval measurement level).

Rank-based tests only consider the ordinal information in the data. So far, the discussion here has focused on applying these tests to rank data that occur due to the nature of the measuring instrument. However, these tests can also be applied to continuous measurements. While considering only the ordinal information of continuous measurements discards some information, it yields significant advantages in terms of robust inference. Specifically, rank-based tests are robust to the presence of outliers, monotonic transformations of the data (e.g., the log-transform), and non-normality of the data. Each of these three components can lead to arbitrary decision-making in the analytic process (e.g., “Is this data point an outlier, and should I remove it?”, “Should I consider the raw, or the log-transformed response times?”, or “Are my data normally distributed?”). How these decisions are handled can heavily influence the outcome of a non-rank-based test. In contrast, the rank-based test omits these arbitrary decisions altogether and makes for a more straightforward analysis procedure.

About halfway through my PhD, my colleagues and I developed a Bayesian

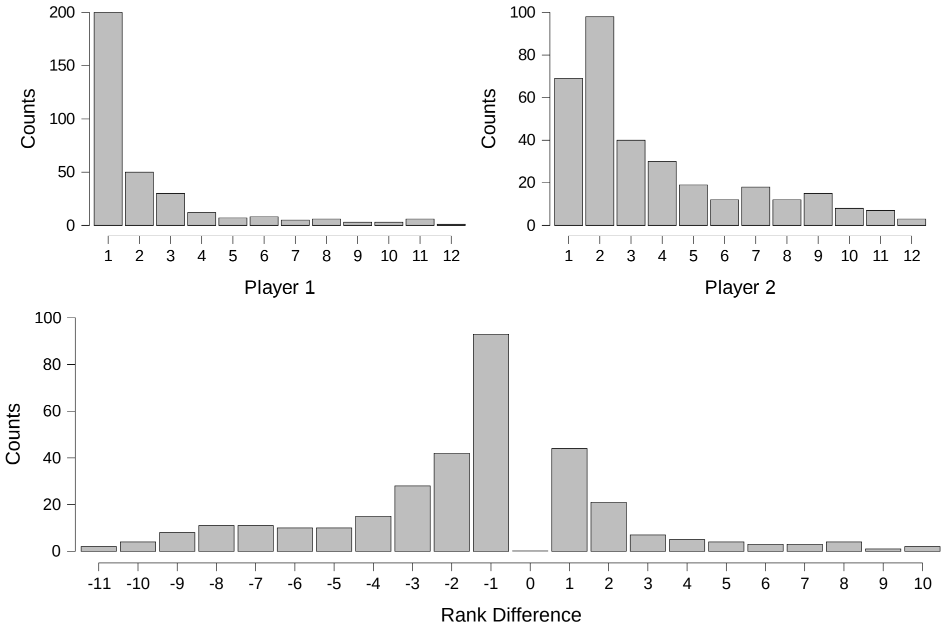


Figure 1.1: The top row presents the observed ranks of all 332 races for both Player 1 and Player 2. Since the observations for both players are paired, the bottom plot shows the difference between the ranks for each race. For instance, if Player 1 finished second, and Player 2 finished third, the difference is $2 - 3 = -1$. As illustrated by the bottom plot, Player 1 finished before Player 2 in the majority of the races. In other words, a negative difference in ranks is an indication of greater skill of Player 1. Since it is not possible to finish a race simultaneously (i.e., ties), there is no observed difference of 0.

framework for several rank-based hypothesis tests. One of these, the Wilcoxon signed rank test, was the exact test needed to analyze the Mario Kart data. The data are summarized in Figure 1.1. The plots in the top row display the frequencies of our respective ranks and the bottom row displays the difference in rank, for each race. A negative difference here is indicative of Player 1 having greater skill, and a positive difference here is indicative of Player 2 having greater skill. Thus, if there are more negative differences than positive differences (i.e., Player 1 defeated Player 2 in the majority of the races), there is evidence that Player 1 is better, and vice versa. Additionally, the magnitude of the differences can be considered. For example, if Player 1 wins five games, and Player 2 gets second place every time (i.e., five differences scores of -1), this is less diagnostic for a skill difference than when Player 1 wins five games, and Player 2 gets twelfth place every time (i.e., five difference scores of -11). The Wilcoxon signed rank test is precisely based on these two considerations.

With the test defined, the hypotheses can be formulated more specifically, and posit a *direction* of the effect:

$\mathcal{H}_-$: Player 1 is better
(i.e., there are more and greater *negative* differences)

$\mathcal{H}_+$: Player 2 is better
(i.e., there are more and greater *positive* differences)

$\mathcal{H}_0$: Players 1 and 2 are equally good
(i.e., the positive and negative differences are equal in frequency and magnitude)

Using the method outlined in Chapter 9, a Bayes factor can be obtained that quantifies the support of one of the hypotheses under consideration over another. Figure 1.2 shows the prior and posterior distribution for δ, which is a standardized measure of the differences between the ranks. Additionally, it includes the Bayes factor comparing $\mathcal{H}_-$ to $\mathcal{H}_0$, and indicates overwhelming evidence that Player 1 is better than Player 2 at Mario Kart: the observed data are $9,950,000$ times more likely under the hypothesis that Player 1 is better, than under the hypothesis that we are both equally good.

Since I am Player 1, I considered my research to be a great success. However, I realized that none of my academic peers were as impressed with the result as I was. I quickly came to the conclusion that the reason for this lack of enthusiasm must be a lack of understanding and popularity of Bayesian inference. Since Bayesian methods are relatively new, many researchers struggle to understand the central concepts inherent to these methods, do not know how to properly conduct and report Bayesian analyses, or are simply unfamiliar with the framework. The other half of my research was therefore dedicated to improving this situation by writing various tutorial-style articles on how to best conduct and teach Bayesian inference.

I strongly believe that the value of a statistical method stands or falls on the successful and correct application of the method. It is therefore of critical importance for the improvement of psychological science to not only develop a prudent

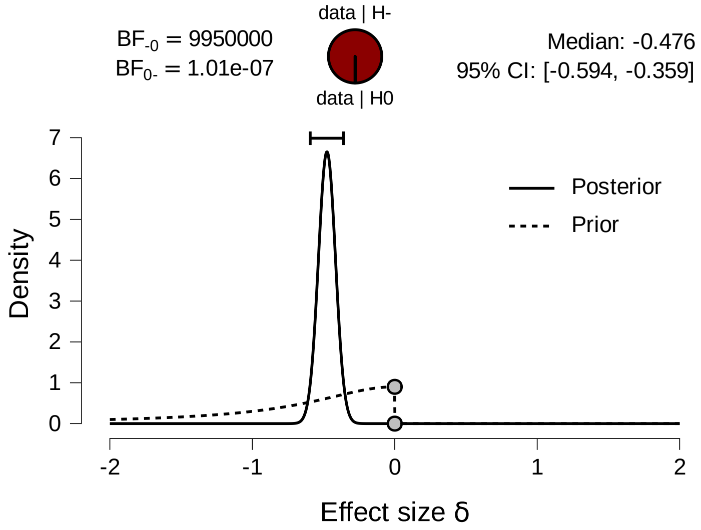


Figure 1.2: Bayesian Wilcoxon signed rank test for the difference parameter δ. The probability wheel at the top illustrates the ratio of the evidence in favor of the two hypotheses under consideration. The Bayes factor indicates that the data $9,950,000$ times more likely under $\mathcal{H}_-$ than under $\mathcal{H}_0$.

statistical method, but to also properly document and explain the method, and to develop easy-to-use tools for applying the method. The combination of part I (Bayes Factor Hypothesis Tests for Researchers) and part II (Bayes Factor Hypothesis Tests for Ranks) of this thesis reflect this belief, and will enable any interested researcher to learn about rank-based Bayesian inference, as well as properly understand, conduct, and report the Bayesian tests introduced in this thesis.

1.2 Chapter Outline

1.2.1 Part I: For Researchers

The first part of the dissertation is concerned with guiding researchers into the realm of Bayesian inference, with an emphasis on Bayes factor hypothesis testing. Instead of introducing new methods, this part focuses on demonstrating the proper use of existing statistical tools.

Chapter 2 introduces a set of guiding principles for researchers looking to apply the Bayesian framework. The guidelines are divided into four main parts of statistical inference: planning, conducting, interpreting, and reporting an analysis. The main focus of this chapter is promoting understanding and transparency in psychological science, such that any scientific finding can be easily contextualized and reproduced. The chapter uses a running example to demonstrate the practical relevance of each guideline as they are introduced. While the guidelines are aimed at Bayesian inference, many principles discussed here extend beyond the Bayesian framework.

Chapter 3 describes a qualitative study where four teams of statisticians were asked to analyze two relatively simple data sets (concerning an association between two continuous variables, and a cross table). The goal of this study was to demonstrate how, even for a simple scenario, there are many ways to approach a data set and corresponding research question. The round table discussion at the end of the chapter provides insight in how the different teams arrived at their specific analysis choices, and highlights how there is never a single, unanimous approach to statistical inference. The main question presented here, is whether it is problematic, or beneficial, when different teams choose different approaches, while arriving at qualitatively similar conclusions.

Chapter 4 provides an interactive and informative experiment that can be used to teach Bayesian inference at the beginner level. The experiment is based on Ronald Fisher's experiment "A lady tasting tea" (Fisher, 1935), and the chapter describes how a similar experiment can be conducted in real-time in a classroom setting. This educational exercise familiarizes participants with the core ideas of Bayesian inference, such as the prior and posterior distribution, likelihood, and Bayes factor.

Chapter 5 discusses how applied researchers reason about the claims made in their empirical papers. The chapter presents results from a questionnaire sent to lead authors of empirical articles in the journal *Nature Human Behavior*. Most respondents in the questionnaire only reported a modest increase of the plausibility of the main claim in their article, as a result of their data. The aim of this study was to gauge whether there would be a gap between the seeming certainty with which empirical articles are written, and the (semi)private convictions of the researchers themselves.

Chapter 6 aims to facilitate a much needed discussion and set of guiding principles about how to properly conduct Bayesian model comparison of mixed effects models (also called hierarchical models). This family of models offers great versatility and applicability in psychological science, in particular when combined with the benefits of Bayesian hypothesis testing. Unfortunately, there are several modeling choices to overcome, which are not well documented in the literature. Using three data examples, this chapter outlines modeling questions related to choice of alternative and null model, prior distribution, aggregation of the data, and measurement error in the context of mixed effects models.

1.2.2 Part II: For Ranks

The second part of the dissertation is concerned with Bayesian inference for rank data.

Chapter 7 introduces Bayesian inference for rank correlation Kendall's τ. By using the asymptotic likelihood of the test statistic, rather than the likelihood of the rank data, the problematic lack of a likelihood is omitted (Yuan & Johnson, 2008). A method for prior elicitation, "parametric yoking", is introduced to create a default prior distribution for Kendall's τ based on the Bayesian framework for Pearson's ρ. The result is a posterior distribution for estimation and a Bayes factor for hypothesis testing.

Chapter 8 introduces a general data augmentation framework for Bayesian inference for rank data and applies this to Kendall's τ. In this data augmentation framework, rank data are seen as impoverished measures of an underlying (i.e., latent) continuous scale. For the Mario Kart example, this means that the observed ranks are treated as ordinal manifestations of our latent ability. Using MCMC-sampling in combination with the ordinal information in the data, this latent scale can be approximated, and the uncertainty with which ranks are reflecting the latent construct is adequately accounted for. A mixture of parametric test can then be performed on these latent values. This method yields identical, or more accurate results for Kendall's τ as in Chapter 7. The behavior of this method is illustrated by a simulation study and data application.

Chapter 9 applies the data augmentation framework from Chapter 8 to other rank-based tests to yield Bayesian inference for Spearman's ρ, the Wilcoxon rank sum test, and the Wilcoxon signed rank test. By demonstrating the application of the framework to other tests, the generalizability of the data augmentation framework is underscored. The behavior of each test is illustrated by a simulation study and data application.

Chapter 10 outlines how Kendall's distance, the unstandardized version of Kendall's τ, is a highly versatile tool in psychological modeling that can be used to express the (dis)similarity between two participants' responses, or between a participants' responses and the ground truth. In contrast to the previous chapters, this chapter does not outline a statistical test or estimation procedure, but a summarizing statistic that can be used for further analysis, such as multidimensional scaling. The chapter describes the basic measure and four extensions that increase its applicability. With the measure in hand, it is then applied to four real-world examples to demonstrate its use.

Part I

For Researchers

Chapter 2

The JASP Guidelines for Conducting and Reporting a Bayesian Analysis

Abstract

Despite the increasing popularity of Bayesian inference in empirical research, few practical guidelines provide detailed recommendations for how to apply Bayesian procedures and interpret the results. Here we offer specific guidelines for four different stages of Bayesian statistical reasoning in a research setting: *planning* the analysis, *executing* the analysis, *interpreting* the results, and *reporting* the results. The guidelines for each stage are illustrated with a running example. Although the guidelines are geared toward analyses performed with the open-source statistical software JASP, most guidelines extend to Bayesian inference in general.

2.1 Introduction

In recent years Bayesian inference has become increasingly popular, both in statistical science and in applied fields such as psychology, biology, and econometrics (e.g., Vandekerckhove et al., 2018; Andrews & Baguley, 2013). For the pragmatic researcher, the adoption of the Bayesian framework brings several advantages over the standard framework of frequentist null-hypothesis significance testing (NHST), including (1) the ability to obtain evidence in favor of the null hypothesis and discriminate between "absence of evidence" and "evidence of absence" (Dienes, 2014; Keysers et al., 2020); (2) the ability to take into account prior knowledge to construct a more informative test (Lee & Vanpaemel, 2018; Gronau et al., 2018); and (3) the ability to monitor the evidence as the data accumulate (Rouder, 2014). However, the relative novelty of conducting Bayesian analyses in applied fields means that there are no detailed reporting standards,

This chapter is published as van Doorn, J.B., van den Bergh, D., Boehm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharsky, S., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.–J. (in press). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. Also available as *PsyArXiv preprint*: https://psyarxiv.com/yqxfr and this in turn may frustrate the broader adoption and proper interpretation of the Bayesian framework.

Several recent statistical guidelines include information on Bayesian inference, but these guidelines are either minimalist (The BaSiS group, 2001; Appelbaum et al., 2018), focus only on relatively complex statistical tests (Depaoli & van de Schoot, 2017), are too specific to a certain field (Spiegelhalter et al., 2000; Sung et al., 2005), or do not cover the full inferential process (Jarosz & Wiley, 2014). The current chapter aims to provide a general overview of the different stages of the Bayesian reasoning process in a research setting. Specifically, we focus on guidelines for analyses conducted in JASP (JASP Team, 2020; `jasp-stats.org`), although these guidelines can be generalized to other software packages for Bayesian inference. JASP is an open-source statistical software program with a graphical user interface that features both Bayesian and frequentist versions of common tools such as the *t*-test, the ANOVA, and regression analysis (e.g., Marsman & Wagenmakers, 2017; Wagenmakers, Love, et al., 2018a).

We discuss four stages of analysis: planning, executing, interpreting, and reporting. These stages and their individual components are summarized in Table 2.1 at the end of this chapter. In order to provide a concrete illustration of the guidelines for each of the four stages, each section features a data set reported by Frisby & Clatworthy (1975). This data set concerns the time it took two groups of participants to see a figure hidden in a stereogram – one group received advance visual information about the scene (i.e., the VV condition), whereas the other group did not (i.e., the NV condition).[1] Three additional examples (mixed ANOVA, correlation analysis, and a *t*-test with an informed prior) are provided in an online appendix at `https://osf.io/nw49j/`. Throughout the paper, we present three boxes that provide additional technical discussion. These boxes, while not strictly necessary, may prove useful to readers interested in greater detail.

2.2 Stage 1: Planning the Analysis

Specifying the goal of the analysis. We recommend that researchers carefully consider their goal, that is, the research question that they wish to answer, prior to the study (Jeffreys, 1939). When the goal is to ascertain the presence or absence of an effect, we recommend a Bayes factor hypothesis test (see Box 1). The Bayes factor compares the predictive performance of two hypotheses. This underscores an important point: in the Bayes factor testing framework, hypotheses cannot be evaluated until they are embedded in fully specified models with a prior distribution and likelihood (i.e., in such a way that they make quantitative predictions about the data). Thus, when we refer to the predictive performance of a hypothesis, we implicitly refer to the accuracy of the predictions made by the model that encompasses the hypothesis (Etz et al., 2018).

[1] The variables are participant number, the time (in seconds) each participant needed to see the hidden figure (i.e., fuse time), experimental condition (VV = with visual information, NV = without visual information), and the log-transformed fuse time.

When the goal is to determine the size of the effect, under the assumption that it is present, we recommend to plot the posterior distribution or summarize it by a credible interval (see Box 2). Testing and estimation are not mutually exclusive and may be used in sequence; for instance, one may first use a test to ascertain that the effect exists, and then continue to estimate the size of the effect.

Box 1. Hypothesis testing. The principled approach to Bayesian hypothesis testing is by means of the Bayes factor (e.g., Wrinch & Jeffreys, 1921; Etz & Wagenmakers, 2017; Jeffreys, 1939; Ly et al., 2016). The Bayes factor quantifies the relative predictive performance of two rival hypotheses, and it is the degree to which the data demand a change in beliefs concerning the hypotheses' relative plausibility (see Equation 2.1). Specifically, the first term in Equation 2.1 corresponds to the prior odds, that is, the relative plausibility of the rival hypotheses before seeing the data. The second term, the Bayes factor, indicates the evidence provided by the data. The third term, the posterior odds, indicates the relative plausibility of the rival hypotheses after having seen the data.

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{p(D \mid \mathcal{H}_1)}{p(D \mid \mathcal{H}_0)}}_{\text{Bayes factor}_{10}} = \underbrace{\frac{p(\mathcal{H}_1 \mid D)}{p(\mathcal{H}_0 \mid D)}}_{\text{Posterior odds}} \tag{2.1}$$

The subscript in the Bayes factor notation indicates which hypothesis is supported by the data. BF_{10} indicates the Bayes factor in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$, whereas BF_{01} indicates the Bayes factor in favor of $\mathcal{H}_0$ over $\mathcal{H}_1$. Specifically, $\text{BF}_{10} = 1/\text{BF}_{01}$. Larger values of BF_{10} indicate more support for $\mathcal{H}_1$. Bayes factors range from 0 to ∞, and a Bayes factor of 1 indicates that both hypotheses predicted the data equally well. This principle is further illustrated in Figure 2.4.

Box 2. Parameter estimation. For Bayesian parameter estimation, interest centers on the posterior distribution of the model parameters. The posterior distribution reflects the relative plausibility of the parameter values after prior knowledge has been updated by means of the data. Specifically, we start the estimation procedure by assigning the model parameters a prior distribution that reflects the relative plausibility of each parameter value before seeing the data. The information in the data is then used to update the prior distribution to the posterior distribution. Parameter values that predicted the data relatively well receive a boost in plausibility, whereas parameter values that predicted the data relatively poorly suffer a decline (Wagenmakers et al., 2016). Equation 2.2 illustrates this principle. The first term indicates the prior beliefs about the values of parameter θ. The second term is the updating factor: for each value of θ, the quality of its prediction is compared to the average quality of the predictions over all values of θ. The third term indicates the posterior beliefs about θ.

$$\underbrace{p(\theta)}_{\substack{\text{Prior belief}\\ \text{about } \theta}} \times \underbrace{\frac{\overbrace{p(\text{data} \mid \theta)}^{\substack{\text{Predictive adequacy}\\ \text{of specific } \theta}}}{p(\text{data})}}_{\substack{\text{Average predictive}\\ \text{adequacy across all } \theta's}} = \underbrace{p(\theta \mid \text{data})}_{\substack{\text{Posterior belief}\\ \text{about } \theta}} . \tag{2.2}$$

The posterior distribution can be plotted or summarized by an x% credible interval. An x% credible interval contains x% of the posterior mass. Two popular ways of creating a credible interval are the highest density credible interval, which is the narrowest interval containing the specified mass, and the central credible interval, which is created by cutting off $\frac{100-x}{2}$% from each of the tails of the posterior distribution.

Specifying the statistical model. The functional form of the model (i.e., the likelihood; Etz, 2018) is guided by the nature of the data and the research question. For instance, if interest centers on the association between two variables, one may specify a bivariate normal model in order to conduct inference on Pearson's correlation parameter ρ. The statistical model also determines which assumptions ought to be satisfied by the data. For instance, the statistical model might assume the dependent variable to be normally distributed. Violations of assumptions may be addressed at different points in the analysis, such as the data preprocessing steps discussed below, or by planning to conduct robust inferential procedures as a contingency plan.

The next step in model specification is to determine the sidedness of the procedure. For hypothesis testing, this means deciding whether the procedure is one-sided (i.e., the alternative hypothesis dictates a specific direction of the population effect) or two-sided (i.e., the alternative hypothesis dictates that the effect can be either positive or negative). The choice of one-sided versus two-sided depends on the research question at hand and this choice should be theoretically justified prior to the study. For hypothesis testing it is usually the case that the alternative hypothesis posits a specific direction. In Bayesian hypothesis testing, a one-sided hypothesis yields a more diagnostic test than a two-sided alternative (e.g., Wetzels et al., 2009; Jeffreys, 1961, p.283).[2]

For parameter estimation, we recommend to always use the two-sided model instead of the one-sided model: when a positive one-sided model is specified but the observed effect turns out to be negative, all of the posterior mass will

[2] A one-sided alternative hypothesis makes a more risky prediction than a two-sided hypothesis. Consequently, if the data are in line with the one-sided prediction, the one-sided alternative hypothesis is rewarded with a greater gain in plausibility compared to the two-sided alternative hypothesis; if the data oppose the one-sided prediction, the one-sided alternative hypothesis is penalized with a greater loss in plausibility compared to the two-sided alternative hypothesis.

nevertheless remain on the positive values, falsely suggesting the presence of a small positive effect.

The next step in model specification concerns the type and spread of the prior distribution, including its justification. For the most common statistical models (e.g., correlations, *t*-tests, and ANOVA), certain "default" prior distributions are available that can be used in cases where prior knowledge is absent, vague, or difficult to elicit (for more information, see Ly et al., 2016). These priors are default options in JASP. In cases where prior information is present, different "informed" prior distributions may be specified. However, the more the informed priors deviate from the default priors, the stronger becomes the need for a justification (see the informed *t*-test example in the online appendix at https://osf.io/ybszx/). Additionally, the robustness of the result to different prior distributions can be explored and included in the report. This is an important type of robustness check because the choice of prior can sometimes impact our inferences, such as in experiments with small sample sizes or missing data. In JASP, Bayes factor robustness plots show the Bayes factor for a wide range of prior distributions, allowing researchers to quickly examine the extent to which their conclusions depend on their prior specification. An example of such a plot is given later in Figure 2.7.

Specifying data preprocessing steps. Dependent on the goal of the analysis and the statistical model, different data preprocessing steps might be taken. For instance, if the statistical model assumes normally distributed data, a transformation to normality (e.g., the logarithmic transformation) might be considered (e.g., Draper & Cox, 1969). Other points to consider at this stage are when and how outliers may be identified and accounted for, which variables are to be analyzed, and whether further transformation or combination of data are necessary. These decisions can be somewhat arbitrary, and yet may exert a large influence on the results (Wicherts et al., 2016). In order to assess the degree to which the conclusions are robust to arbitrary modeling decisions, it is advisable to conduct a multiverse analysis (Steegen et al., 2016). Preferably, the multiverse analysis is specified at study onset. A multiverse analysis can easily be conducted in JASP, but doing so is not the goal of the current paper.

Specifying the sampling plan. As may be expected from a framework for the continual updating of knowledge, Bayesian inference allows researchers to monitor evidence as the data come in, and stop whenever they like, for any reason whatsoever. Thus, strictly speaking there is no Bayesian need to pre-specify sample size at all (e.g., Berger & Wolpert, 1988). Nevertheless, Bayesians are free to specify a sampling plan if they so desire; for instance, one may commit to stop data collection as soon as $\text{BF}_{10} \geq 10$ or $\text{BF}_{01} \geq 10$. This approach can also be combined with a maximum sample size (N), where data collection stops when either the maximum N or the desired Bayes factor is obtained, whichever comes first (for examples see Matzke et al., 2015; Wagenmakers et al., 2015).

In order to examine what sampling plans are feasible, researchers can conduct a *Bayes factor design analysis* (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019), a method that shows the predicted outcomes for different designs and sampling plans. Of course, when the study is observational and the data are available 'en bloc', the sampling plan becomes irrelevant in the planning stage.

Stereogram Example

First, we consider the research goal, which was to determine if participants who receive advance visual information exhibit a shorter fuse time (Frisby & Clatworthy, 1975). A Bayes factor hypothesis test can be used to quantify the evidence that the data provide for and against the hypothesis that an effect is present. Should this test reveal support in favor of the presence of the effect, then we have grounds for a follow-up analysis in which the size of the effect is estimated.

Second, we specify the statistical model. The study focus is on the difference in performance between two between-subjects conditions, suggesting a two-sample *t*-test on the fuse times is appropriate. The main measure of the study is a reaction time variable, which can for various reasons be non-normally distributed (Lo & Andrews, 2015; but see Schramm & Rouder, 2019). If our data show signs of non-normality we will conduct two alternatives: a *t*-test on the log-transformed fuse time data and a non-parametric *t*-test (i.e., the Mann-Whitney U test), which is robust to non-normality and unaffected by the log-transformation of the fuse times.

For hypothesis testing, we compare the null hypothesis (i.e., advance visual information has no effect on fuse times) to a one-sided alternative hypothesis (i.e., advance visual information *shortens* the fuse times), in line with the directional nature of the original research question. The rival hypotheses are thus $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_+ : \delta > 0$, where δ is the standardized effect size (i.e., the population version of Cohen's d), $\mathcal{H}_0$ denotes the null hypothesis, and $\mathcal{H}_+$ denotes the one-sided alternative hypothesis (note the '+' in the subscript). For parameter estimation (under the assumption that the effect exists) we use the two-sided *t*-test model and plot the posterior distribution of δ. This distribution can also be summarized by a 95% central credible interval.

We complete the model specification by assigning prior distributions to the model parameters. Since we have only little prior knowledge about the topic, we select a default prior option for the two-sample *t*-test, that is, a Cauchy distribution[3] with spread r set to $1/\sqrt{2}$. Since we specified a one-sided alternative hypothesis, the prior distribution is truncated at zero, such that only positive effect size values are allowed. The robustness of the Bayes factor to this prior specification can be easily assessed in JASP by means of a Bayes factor robustness plot.

Since the data are already available, we do not have to specify a sampling plan. The original data set has a total sample size of 103, from which 25 participants were eliminated due to failing an initial stereo-acuity test, leaving 78 participants (43 in the NV condition and 35 in the VV condition). The data are available online at https://osf.io/5vjyt/.

[3] The fat-tailed Cauchy distribution is a popular default choice because it fulfills particular desiderata, see Jeffreys, 1961; Liang et al., 2008; Ly et al., 2016; Rouder et al., 2009 for details.

2.3 Stage 2: Executing the Analysis

Before executing the primary analysis and interpreting the outcome, it is important to confirm that the intended analyses are appropriate and the models are not grossly misspecified for the data at hand. In other words, it is strongly recommended to examine the validity of the model assumptions (e.g., normally distributed residuals or equal variances across groups). Such assumptions may be checked by plotting the data, inspecting summary statistics, or conducting formal assumption tests (but see Tijmstra, 2018).

A powerful demonstration of the dangers of failing to check the assumptions is provided by Anscombe's quartet (Anscombe, 1973; see Figure 2.1). The quartet consists of four fictitious data sets of equal size that each have the same observed Pearson's product moment correlation r, and therefore lead to the same inferential result both in a frequentist and a Bayesian framework. However, visual inspection of the scatterplots immediately reveals that three of the four data sets are not suitable for a linear correlation analysis, and the statistical inference for these three data sets is meaningless or even misleading. This example highlights the adage that conducting a Bayesian analysis does not safeguard against general statistical malpractice – the Bayesian framework is as vulnerable to violations of assumptions as its frequentist counterpart. In cases where assumptions are violated, an ordinal or non-parametric test can be used, and the parametric results should be interpreted with caution.

Once the quality of the data has been confirmed, the planned analyses can be carried out. JASP offers a graphical user interface for both frequentist and Bayesian analyses. JASP 0.10.2 features the following Bayesian analyses: the binomial test, the chi-square test, the multinomial test, the t-test (one-sample, paired sample, two-sample, Wilcoxon rank sum, and Wilcoxon signed-rank tests), A/B tests, ANOVA, ANCOVA, repeated measures ANOVA, correlations (Pearson's ρ and Kendall's τ), linear regression, and log-linear regression. After loading the data into JASP, the desired analysis can be conducted by dragging and dropping variables into the appropriate boxes; tick marks can be used to select the desired output.

The resulting output (i.e., figures and tables) can be annotated and saved as a `.jasp` file. Output can then be shared with peers, with or without the real data in the `.jasp` file; if the real data are added, reviewers can easily reproduce the analyses, conduct alternative analyses, or insert comments.

Stereogram Example

In order to check for violations of the assumptions of the t-test, the top row of Figure 2.2 shows boxplots and Q-Q plots of the dependent variable fuse time, split by condition. Visual inspection of the boxplots suggests that the variances of the fuse times may not be equal (observed standard deviations of the NV and VV groups are 8.085 and 4.802, respectively), suggesting the equal variance assumption may be unlikely to hold. There also appear to be a number of potential outliers in both groups. Moreover, the Q-Q plots show that the normality as-

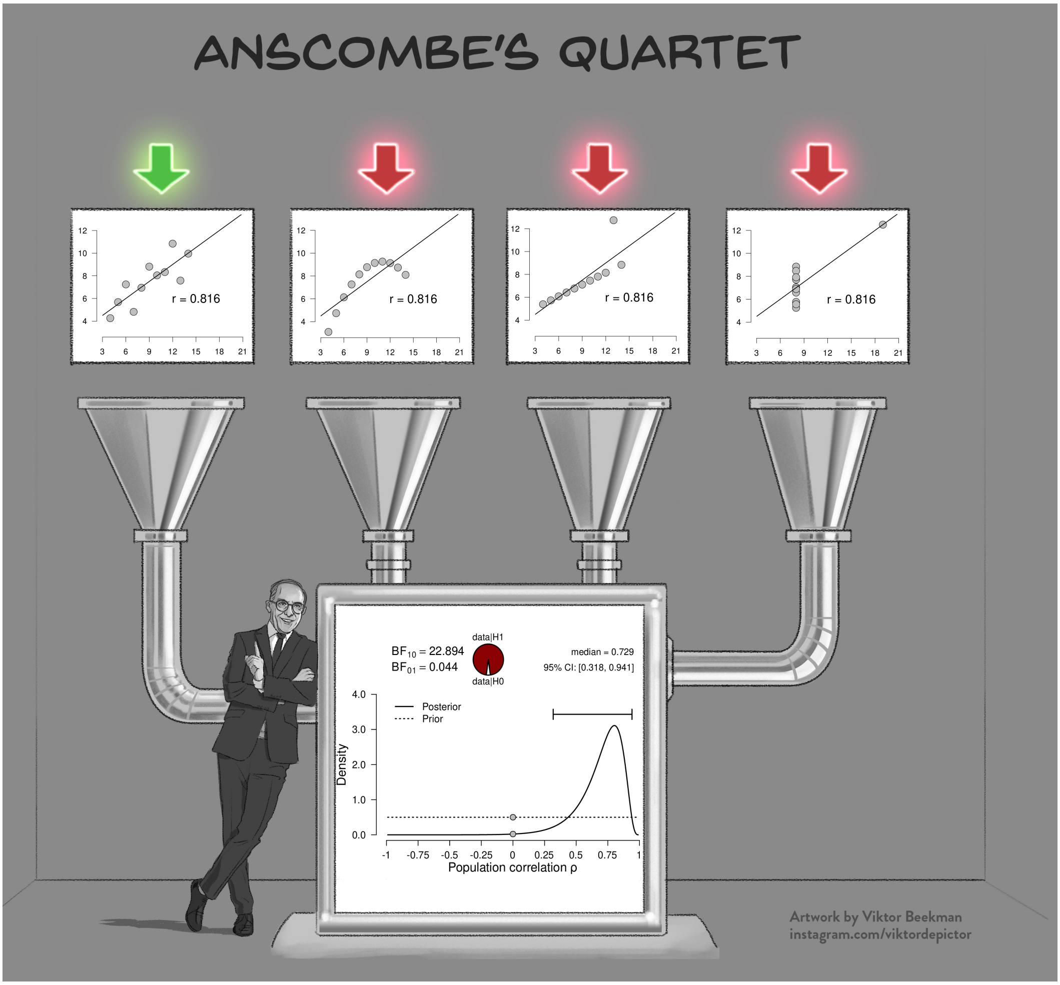
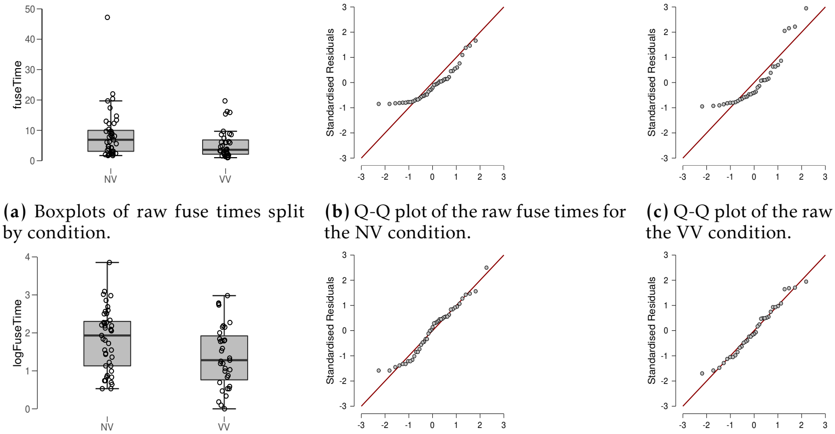


Figure 2.1: Model misspecification is also a problem for Bayesian analyses. The four scatterplots on top show Anscombe's quartet (Anscombe, 1973); the bottom panel shows the corresponding inference, which is identical for all four scatter plots. Except for the leftmost scatterplot, all data violate the assumptions of the linear correlation analysis in important ways.

sumption of the t-test is untenable here. Thus, in line with our analysis plan we will apply the log-transformation to the fuse times. The standard deviations of the log-transformed fuse times in the groups are roughly equal (observed standard deviations are 0.814 and 0.818 in the NV and the VV group, respectively); the Q-Q plots in the bottom row of Figure 2.2 also look acceptable for both groups and there are no apparent outliers. However, it seems prudent to assess the robustness of the result by also conducting the Bayesian Mann-Whitney U test (van Doorn et al., 2020) on the fuse times.



(a) Boxplots of raw fuse times split by condition.

(b) Q-Q plot of the raw fuse times for the NV condition.

(c) Q-Q plot of the raw fuse times for the VV condition.

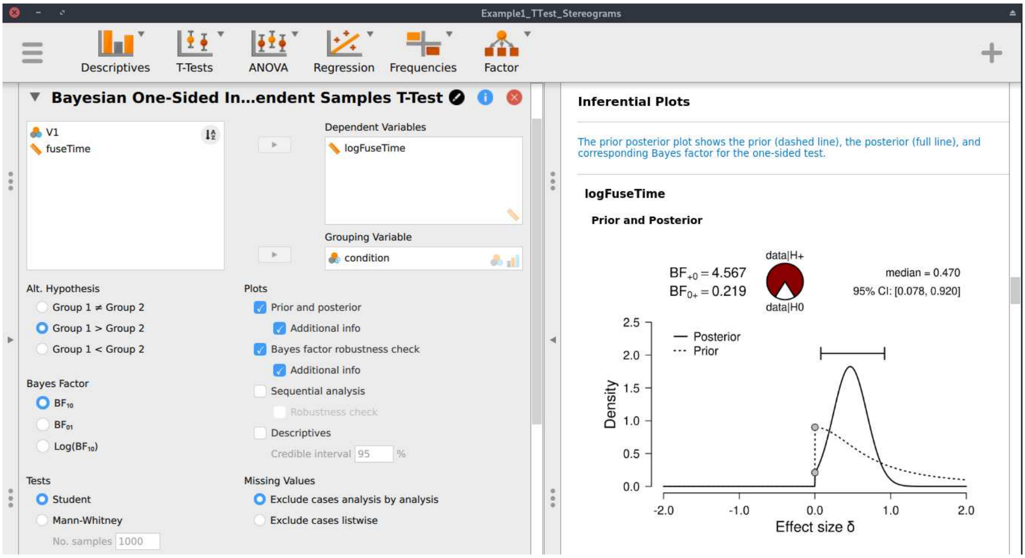
(d) Boxplots of log fuse times split by condition.

(e) Q-Q plot of the log fuse times for the NV condition

(f) Q-Q plot of the log fuse times for the VV condition.

Figure 2.2: Descriptive plots allow a visual assessment of the assumptions of the *t*-test for the stereogram data. The top row shows descriptive plots for the raw fuse times, and the bottom row shows descriptive plots for the log-transformed fuse times. The left column shows boxplots, including the jittered data points, for each of the experimental conditions. The middle and right columns show Q-Q plots of the dependent variable, split by experimental condition. Here we see that the log-transformed dependent variable is more appropriate for the *t*-test, due to its distribution and absence of outliers. Figures from JASP.

Following the assumption check we proceed to execute the analyses in JASP. For hypothesis testing, we obtain a Bayes factor using the one-sided Bayesian two-sample *t*-test. Figure 2.3 shows the JASP user interface for this procedure. For parameter estimation, we obtain a posterior distribution and credible interval, using the two-sided Bayesian two-sample *t*-test. The relevant boxes for the various plots were ticked, and an annotated `.jasp` file was created with all of the relevant analyses: the one-sided Bayes factor hypothesis tests, the robustness check, the posterior distribution from the two-sided analysis, and the one-sided results of the Bayesian Mann-Whitney U test. The `.jasp` file can be found at `https://osf.io/nw49j/`. The next section outlines how these results are to be interpreted.



Bayesian One-Sided In...endent Samples T-Test

V1
fuseTime

Dependent Variables
logFuseTime

Grouping Variable
condition

Alt. Hypothesis
○ Group 1 ≠ Group 2
● Group 1 > Group 2
○ Group 1 < Group 2

Bayes Factor
● BF_{10}
○ BF_{01}
○ $Log(BF_{10})$

Tests
● Student
○ Mann-Whitney
No. samples 1000

Plots
☑ Prior and posterior
☑ Additional info
☑ Bayes factor robustness check
☑ Additional info
□ Sequential analysis
□ Robustness check
□ Descriptives
Credible interval 95 %

Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise

Inferential Plots

The prior posterior plot shows the prior (dashed line), the posterior (full line), and corresponding Bayes factor for the one-sided test.

logFuseTime

Prior and Posterior

$BF_{+0} = 4.567$
$BF_{0+} = 0.219$

median = 0.470
95% CI: [0.078, 0.920]

Figure 2.3: JASP menu for the Bayesian two-sample *t*-test. The left input panel offers the analysis options, including the specification of the alternative hypothesis and the selection of plots. The right output panel shows the corresponding analysis output. The prior and posterior plot is explained in more detail in Figure 2.6b. The input panel specifies the one-sided analysis for hypothesis testing; a two-sided analysis for estimation can be obtained by selecting "Group 1 ≠ Group 2" under "Alt. Hypothesis".

2.4 Stage 3: Interpreting the Results

With the analysis outcome in hand we are ready to draw conclusions. We first discuss the scenario of hypothesis testing, where the goal typically is to conclude whether an effect is present or absent. Then, we discuss the scenario of parameter estimation, where the goal is to estimate the size of the population effect, assuming it is present. When both hypothesis testing and estimation procedures have been planned and executed, there is no predetermined order for their interpretation. One may adhere to the adage "only estimate something when there is something to be estimated" (Wagenmakers, Marsman, et al., 2018) and first test whether an effect is present, and then estimate its size (assuming the test provided sufficiently strong evidence against the null), or one may first estimate the magnitude of an effect, and then quantify the degree to which this magnitude warrants a shift in plausibility away from or toward the null hypothesis (but see Box 3).

If the goal of the analysis is hypothesis testing, we recommend using the Bayes factor. As described in Box 1, the Bayes factor quantifies the relative predictive performance of two rival hypotheses (Wagenmakers et al., 2016; see Box 1). Importantly, the Bayes factor is a *relative* metric of the hypotheses' predictive quality. For instance, if $BF_{10} = 5$, this means that the data are 5 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. However, a Bayes factor in favor of $\mathcal{H}_1$ does not mean

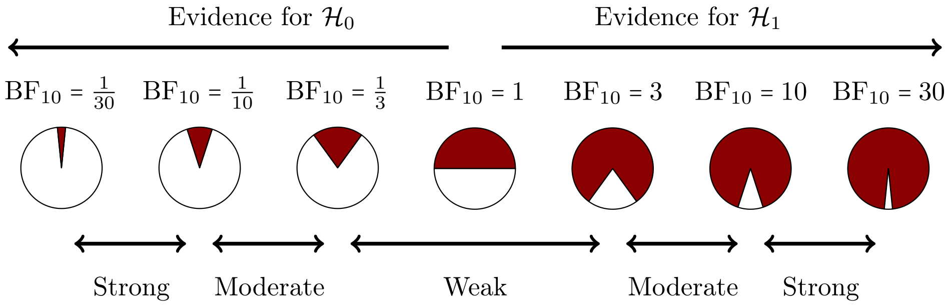

Evidence for $\mathcal{H}_0$
Evidence for $\mathcal{H}_1$
$\text{BF}_{10} = \frac{1}{30}$
$\text{BF}_{10} = \frac{1}{10}$
$\text{BF}_{10} = \frac{1}{3}$
$\text{BF}_{10} = 1$
$\text{BF}_{10} = 3$
$\text{BF}_{10} = 10$
$\text{BF}_{10} = 30$
Strong
Moderate
Weak
Moderate
Strong

Figure 2.4: A graphical representation of a Bayes factor classification table. As the Bayes factor deviates from 1, which indicates equal support for $\mathcal{H}_0$ and $\mathcal{H}_1$, more support is gained for either $\mathcal{H}_0$ or $\mathcal{H}_1$. Bayes factors between 1 and 3 are considered to be weak, Bayes factors between 3 and 10 are considered moderate, and Bayes factors greater than 10 are considered strong evidence. The Bayes factors are also represented as probability wheels, where the ratio of white (i.e., support for $\mathcal{H}_0$) to red (i.e., support for $\mathcal{H}_1$) surface is a function of the Bayes factor. The probability wheels further underscore the continuous scale of evidence that Bayes factors represent. These classifications are heuristic and should not be misused as an absolute rule for all-or-nothing conclusions.

that $\mathcal{H}_1$ predicts the data well. As Figure 2.1 illustrates, $\mathcal{H}_1$ provides a dreadful account of three out of four data sets, yet is still supported relative to $\mathcal{H}_0$.

There can be no hard Bayes factor bound (other than zero and infinity) for accepting or rejecting a hypothesis wholesale, but there have been some attempts to classify the strength of evidence that different Bayes factors provide (e.g., Jeffreys, 1939; Kass & Raftery, 1995). One such classification scheme is shown in Figure 2.4. Several magnitudes of the Bayes factor are visualized as a probability wheel, where the proportion of red to white is determined by the degree of evidence in favor of $\mathcal{H}_0$ and $\mathcal{H}_1$.[4] In line with Jeffreys, a Bayes factor between 1 and 3 is considered weak evidence, a Bayes factor between 3 and 10 is considered moderate evidence, and a Bayes factor greater than 10 is considered strong evidence. Note that these classifications should only be used as general rules of thumb to facilitate communication and interpretation of evidential strength. Indeed, one of the merits of the Bayes factor is that it offers an assessment of evidence on a continuous scale.

When the goal of the analysis is parameter estimation, the posterior distribution is key (see Box 2). The posterior distribution is often summarized by a location parameter (point estimate) and uncertainty measure (interval estimate). For point estimation, the posterior median (reported by JASP), mean, or mode can be reported, although these do not contain any information about the uncertainty of the estimate. In order to capture the uncertainty of the estimate, an x% credible interval can be reported. The credible interval $[L, U]$ has a x% probability that the true parameter lies in the interval that ranges from L to U (an interpretation that is often wrongly attributed to frequentist confidence intervals, see Morey et

[4] Specifically, the proportion of red is the posterior probability of $\mathcal{H}_1$ under a prior probability of 0.5; for a more detailed explanation and a cartoon see https://tinyurl.com/ydhfndxa

al., 2016). For example, if we obtain a 95% credible interval of $[-1, 0.5]$ for effect size δ, we can be 95% certain that the true value of δ lies between -1 and 0.5, assuming that the alternative hypothesis we specify is true. In case one does not want to make this assumption, one can present the *unconditional* posterior distribution instead. For more discussion on this point, see Box 3.

Box 3. Conditional vs. Unconditional Inference. A widely accepted view on statistical inference is neatly summarized by Fisher (1925), who states that "it is a useful preliminary before making a statistical estimate ... to test if there is anything to justify estimation at all" (p. 300; see also J. Haaf et al., 2019). In the Bayesian framework, this stance naturally leads to posterior distributions *conditional* on $\mathcal{H}_1$, which ignores the possibility that the null value could be true. Generally, when we say "prior distribution" or "posterior distribution" we are following convention and referring to such conditional distributions. However, only presenting conditional posterior distributions can potentially be misleading in cases where the null hypothesis remains relatively plausible after seeing the data. A general benefit of Bayesian analysis is that one can compute an *unconditional* posterior distribution for the parameter using model averaging (e.g., Hinne et al., 2020; Clyde et al., 2011). An unconditional posterior distribution for a parameter accounts for both the uncertainty about the parameter within any one model and the uncertainty about the model itself, providing an estimate of the parameter that is a compromise between the candidate models (for more details see Hoeting et al., 1999). In the case of a *t*-test, which features only the null and the alternative hypothesis, the unconditional posterior consists of a mixture between a spike under $\mathcal{H}_0$ and a bell-shaped posterior distribution under $\mathcal{H}_1$ (Rouder et al., 2018; van den Bergh et al., 2019). Figure 2.5 illustrates this approach for the stereogram example.

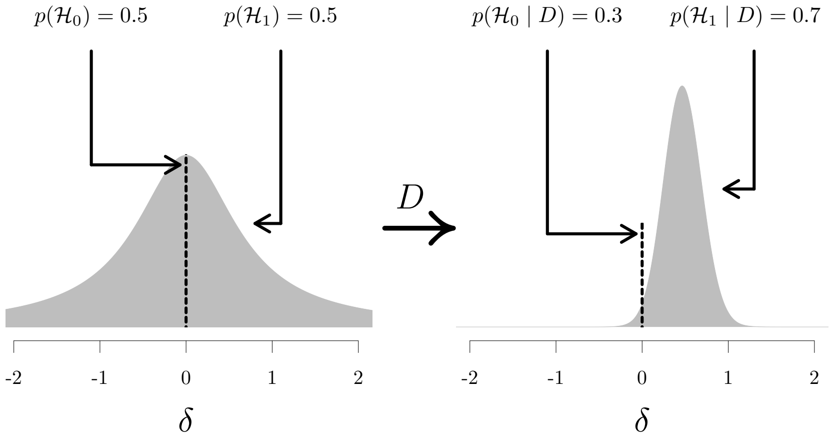


Figure 2.5: Updating the unconditional prior distribution to the unconditional posterior distribution for the stereogram example. The left panel shows the unconditional prior distribution, which is a mixture between the prior distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$. The prior distribution under $\mathcal{H}_0$ is a spike at the null value, indicated by the dotted line; the prior distribution under $\mathcal{H}_1$ is a Cauchy distribution, indicated by the gray mass. The mixture proportion is determined by the prior model probabilities $p(\mathcal{H}_0)$ and $p(\mathcal{H}_1)$. The right panel shows the unconditional posterior distribution, after updating the prior distribution with the data D. This distribution is a mixture between the posterior distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$., where the mixture proportion is determined by the posterior model probabilities $p(\mathcal{H}_0 \mid D)$ and $p(\mathcal{H}_1 \mid D)$. Since $p(\mathcal{H}_1 \mid D) = 0.7$ (i.e., the data provide support for $\mathcal{H}_1$ over $\mathcal{H}_0$), about 70% of the unconditional posterior mass is comprised of the posterior mass under $\mathcal{H}_1$, indicated by the gray mass. Thus, the unconditional posterior distribution provides information about plausible values for δ, while taking into account the uncertainty of $\mathcal{H}_1$ being true. In both panels, the dotted line and gray mass have been rescaled such that the height of the dotted line and the highest point of the gray mass reflect the prior (left) and posterior (right) model probabilities.

Common Pitfalls in Interpreting Bayesian Results

Bayesian veterans sometimes argue that Bayesian concepts are intuitive and easier to grasp than frequentist concepts. However, in our experience there exist persistent misinterpretations of Bayesian results. Here we list five:

- The Bayes factor does not equal the posterior odds; in fact, the posterior odds are equal to the Bayes factor multiplied by the prior odds (see also Equation 2.1). These prior odds reflect the relative plausibility of the rival hypotheses before seeing the data (e.g., 50/50 when both hypotheses are equally plausible, or 80/20 when one hypothesis is deemed to be 4 times more plausible than the other). For instance, a proponent and a skeptic may differ greatly in their assessment of the prior plausibility of a hypothesis; their prior odds differ, and, consequently, so will their posterior odds. However, as the Bayes factor is the updating factor from prior odds to posterior odds, proponent and skeptic ought to change their beliefs to the same degree (assuming they agree on the model specification, including the parameter prior distributions).

- Prior model probabilities (i.e., prior odds) and parameter prior distributions play different conceptual roles.[5] The former concerns prior beliefs about the hypotheses, for instance that both $\mathcal{H}_0$ and $\mathcal{H}_1$ are equally plausible a priori. The latter concerns prior beliefs about the model parameters within a model, for instance that all values of Pearson's ρ are equally likely a priori (i.e., a uniform prior distribution on the correlation parameter). Prior model probabilities and parameter prior distributions can be combined to one unconditional prior distribution as described in Box 3 and Figure 2.5.

- The Bayes factor and credible interval have different purposes and can yield different conclusions. Specifically, the typical credible interval for an effect size is conditional on $\mathcal{H}_1$ being true and quantifies the strength of an effect, assuming it is present (but see Box 3); in contrast, the Bayes factor quantifies evidence for the presence or absence of an effect. A common misconception is to conduct a "hypothesis test" by inspecting only credible intervals. Berger (2006, p. 383) remarks: "[...] Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a hypothesis)."

- The strength of evidence in the data is easy to overstate: a Bayes factor of 3 provides some support for one hypothesis over another, but should not warrant the confident all-or-none acceptance of that hypothesis.

- The results of an analysis always depend on the questions that were asked.[6] For instance, choosing a one-sided analysis over a two-sided analysis will impact both the Bayes factor and the posterior distribution. For an illustration of this, see Figure 2.6 for a comparison between one-sided and a two-sided results.

In order to avoid these and other pitfalls, we recommend that researchers who are doubtful about the correct interpretation of their Bayesian results solicit expert advice (for instance through the JASP forum at http://forum.cogsci.nl).

[5] This confusion does not arise for the rarely reported unconditional distributions (see Box 3).

[6] This is known as Jeffreys's platitude: "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" (Jeffreys, 1939, p.vi).

Stereogram Example

For hypothesis testing, the results of the one-sided t-test are presented in Figure 2.6a. The resulting BF_{+0} is 4.567, indicating moderate evidence in favor of $\mathcal{H}_+$: the data are approximately 4.6 times more likely under $\mathcal{H}_+$ than under $\mathcal{H}_0$. To assess the robustness of this result, we also planned a Mann-Whitney U test. The resulting BF+0 is 5.191, qualitatively similar to the Bayes factor from the parametric test. Additionally, we could have specified a multiverse analysis where data exclusion criteria (i.e., exclusion vs. no exclusion), the type of test (i..e, Mann-Whitney U vs. t-test), and data transformations (i.e., log-transformed vs. raw fuse times) are varied. Typically in multiverse analyses these three decisions would be crossed, resulting in at least eight different analyses. However, in our case some of these analyses are implausible or redundant. First, because the Mann-Whitney U test is unaffected by the log transformation, the log-transformed and raw fuse times yield the same results. Second, due to the multiple assumption violations, the t-test model for raw fuse times is severely misspecified and hence we do not trust the validity of its result. Third, we do not know which observations were excluded by Frisby & Clatworthy (1975). Consequently, only two of these eight analyses are relevant. Furthermore, a more comprehensive multiverse analysis could also consider the Bayes factors from two-sided tests (i.e., BF10 = 2:323 for the t-test and BF10 = 2:557 for the Mann-Whitney U test). However, these tests are not in line with the theory under consideration, as they answer a different theoretical question (see “Specifying the statistical model” in the Planning section).

For parameter estimation, the results of the two-sided t-test are presented in Figure 2.6b. The 95% central credible interval for δ is relatively wide, ranging from 0.046 to 0.904: this means that, under the assumption that the effect exists and given the model we specified, we can be 95% certain that the true value of δ lies between 0.046 to 0.904. In conclusion, there is moderate evidence for the presence of an effect, and large uncertainty about its size.

2.5 Stage 4: Reporting the Results

For increased transparency, and to allow a skeptical assessment of the statistical claims, we recommend to present an elaborate analysis report including relevant tables, figures, assumption checks, and background information. The extent to which this needs to be done in the manuscript itself depends on context. Ideally, an annotated `.jasp` file is created that presents the full results and analysis settings. The resulting file can then be uploaded to the Open Science Framework (OSF; https://osf.io), where it can be viewed by collaborators and peers, even without having JASP installed. Note that the `.jasp` file retains the settings that were used to create the reported output. Analyses not conducted in JASP should mimic such transparency, for instance through uploading an R-script. In this section, we list several desiderata for reporting, both for hypothesis testing and parameter estimation. What to include in the report depends on the goal of the analysis, regardless of whether the result is conclusive or not.

In all cases, we recommend to provide a complete description of the prior specification (i.e., the type of distribution and its parameter values) and, especially for informed priors, to provide a justification for the choices that were made. When reporting a specific analysis, we advise to refer to the relevant background literature for details. In JASP, the relevant references for specific tests can be copied from the drop-down menus in the results panel.

When the goal of the analysis is hypothesis testing, it is key to outline which hypotheses are compared by clearly stating each hypothesis and including the corresponding subscript in the Bayes factor notation. Furthermore, we recommend to include, if available, the Bayes factor robustness check discussed in the section on planning (see Figure 2.7 for an example). This check provides an assessment of the robustness of the Bayes factor under different prior specifications: if the qualitative conclusions do not change across a range of different plausible prior distributions, this indicates that the analysis is relatively robust. If this plot is unavailable, the robustness of the Bayes factor can be checked manually by specifying several different prior distributions (see the mixed ANOVA analysis in the online appendix at https://osf.io/wae57/ for an example). When data come in sequentially, it may also be of interest to examine the sequential Bayes factor plot, which shows the evidential flow as a function of increasing sample size.

When the goal of the analysis is parameter estimation, it is important to present a plot of the posterior distribution, or report a summary, for instance through the median and a 95% credible interval. Ideally, the results of the analysis are reported both graphically and numerically. This means that, when possible, a plot is presented that includes the posterior distribution, prior distribution, Bayes factor, 95% credible interval, and posterior median.[7]

Numeric results can be presented either in a table or in the main text. If relevant, we recommend to report the results from both estimation and hypothesis test. For some analyses, the results are based on a numerical algorithm, such as Markov chain Monte Carlo (MCMC), which yields an error percentage. If applicable and available, the error percentage ought to be reported too, to indicate the numeric robustness of the result. Lower values of the error percentage indicate greater numerical stability of the result.[8] In order to increase numerical stability, JASP includes an option to increase the number of samples for MCMC sampling when applicable.

[7] The posterior median is popular because it is robust to skewed distributions and invariant under smooth transformations of parameters, although other measures of central tendency, such as the mode or the mean, are also in common use.

[8] We generally recommend error percentages below 20% as acceptable. A 20% change in the Bayes factor will result in one making the same qualitative conclusions. However, this threshold naturally increases with the magnitude of the Bayes factor. For instance, a Bayes factor of 10 with a 50% error percentage could be expected to fluctuate between 5 and 15 upon recomputation. This could be considered a large change. However, with a Bayes factor of 1000 a 50% reduction would still leave us with overwhelming evidence.

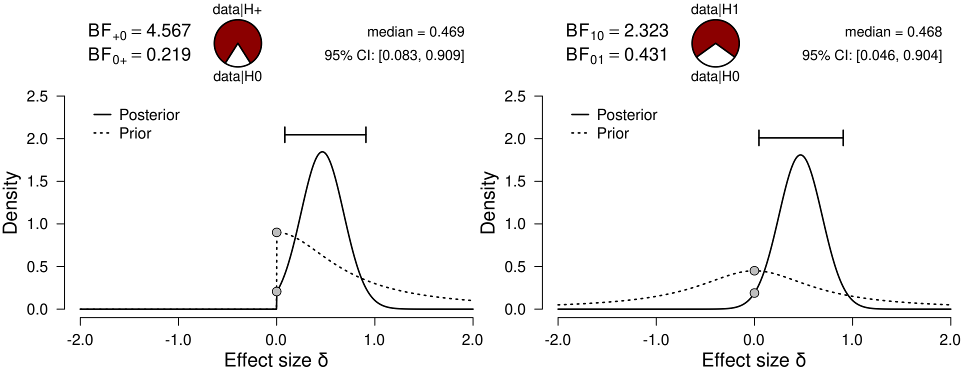
Stereogram Example

This is an example report of the stereograms *t*-test example:

> Here we summarize the results of the Bayesian analysis for the stereogram data. For this analysis we used the Bayesian *t*-test framework proposed by Jeffreys (1961, see also Rouder et al. 2009). We analyzed the data with JASP (JASP Team, 2020). An annotated `.jasp` file, including distribution plots, data, and input options, is available at `https://osf.io/25ekj/`. Due to model misspecification (i.e., non-normality, presence of outliers, and unequal variances), we applied a log-transformation to the fuse times. This remedied the misspecification. To assess the robustness of the results, we also applied a Mann-Whitney U test.
>
> First, we discuss the results for hypothesis testing. The null hypothesis postulates that there is no difference in log fuse time between the groups and therefore $\mathcal{H}_0 : \delta = 0$. The one-sided alternative hypothesis states that only positive values of δ are possible, and assigns more prior mass to values closer to 0 than extreme values. Specifically, δ was assigned a Cauchy prior distribution with $r = 1/\sqrt{2}$, truncated to allow only positive effect size values. Figure 2.6a shows that the Bayes factor indicates evidence for $\mathcal{H}_+$; specifically, $\mathrm{BF}_{+0} = 4.567$, which means that the data are approximately 4.5 times more likely to occur under $\mathcal{H}_+$ than under $\mathcal{H}_0$. This result indicates moderate evidence in favor of $\mathcal{H}_+$. The error percentage is $< 0.001\%$, which indicates great stability of the numerical algorithm that was used to obtain the result. The Mann-Whitney U test yielded a qualitatively similar result, BF_{+0} is 5.191. In order to asses the robustness of the Bayes factor to our prior specification, Figure 2.7 shows BF_{+0} as a function of the prior width r. Across a wide range of widths, the Bayes factor appears to be relatively stable, ranging from about 3 to 5.
>
> Second, we discuss the results for parameter estimation. Of interest is the posterior distribution of the standardized effect size δ (i.e., the population version of Cohen's d, the standardized difference in mean fuse times). For parameter estimation, δ was assigned a Cauchy prior distribution with $r = 1/\sqrt{2}$. Figure 2.6b shows that the median of the resulting posterior distribution for δ equals 0.47 with a central 95% credible interval for δ that ranges from 0.046 to 0.904. If the effect is assumed to exist, there remains substantial uncertainty about its size, with values close to 0 having the same posterior density as values close to 1.

2.6 Limitations and Challenges

The Bayesian toolkit for the empirical social scientist still has some limitations to overcome. First, for some frequentist analyses, the Bayesian counterpart has



(a) One-sided analysis for testing: $H_+ : \delta > 0$

(b) Two-sided analysis for estimation: $\mathcal{H}_1 : \delta \sim$ Cauchy

Figure 2.6: Bayesian two-sample *t*-test for the parameter *δ*. The probability wheel on top visualizes the evidence that the data provide for the two rival hypotheses. The two gray dots indicate the prior and posterior density at the test value (Dickey & Lientz, 1970; Wagenmakers et al., 2010). The median and the 95% central credible interval of the posterior distribution are shown in the top right corner. The left panel shows the one-sided procedure for hypothesis testing and the right panel shows the two-sided procedure for parameter estimation. Both figures from JASP.

not yet been developed or implemented in JASP. Secondly, some analyses in JASP currently provide only a Bayes factor, and not a visual representation of the posterior distributions, for instance due to the multidimensional parameter space of the model. Thirdly, some analyses in JASP are only available with a relatively limited set of prior distributions. However, these are not principled limitations and the software is actively being developed to overcome these limitations. When dealing with more complex models that go beyond the staple analyses such as *t*-tests, there exist a number of software packages that allow custom coding, such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017). Another option for Bayesian inference is to code the analyses in a programming language such as R (R Development Core Team, 2004) or Python (van Rossum, 1995). This requires a certain degree of programming ability, but grants the user more flexibility. Popular packages for conducting Bayesian analyses in R are the BayesFactor package (Morey & Rouder, 2018) and the brms package (Bürkner, 2017), among others (see https://cran.r-project.org/web/views/Bayesian.html for a more exhaustive list). For Python, a popular package for Bayesian analyses is PyMC3 (Salvatier et al., 2016). The practical guidelines provided in this paper can largely be generalized to the application of these software programs.

2.7 Concluding Comments

We have attempted to provide concise recommendations for planning, executing, interpreting, and reporting Bayesian analyses. These recommendations are summarized in Table 2.1. Our guidelines focused on the standard analyses that are currently featured in JASP. When going beyond these analyses, some of the dis-

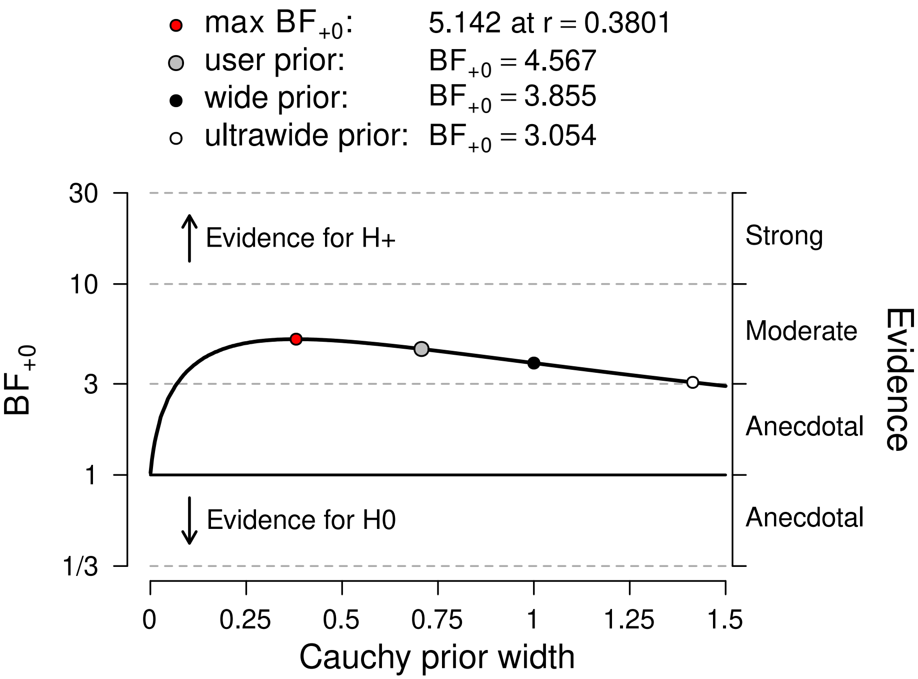


Figure 2.7: The Bayes factor robustness plot. The maximum BF_{+0} is attained when setting the prior width r to 0.38. The plot indicates BF_{+0} for the user specified prior ($r = 1/\sqrt{2}$), wide prior ($r = 1$), and ultrawide prior ($r = \sqrt{2}$). The evidence for the alternative hypothesis is relatively stable across a wide range of prior distributions, suggesting that the analysis is robust. However, the evidence in favor of $\mathcal{H}_+$ is not particularly strong and will not convince a skeptic.

cussed guidelines will be easier to implement than others. However, the general process of transparent, comprehensive, and careful statistical reporting extends to all Bayesian procedures and indeed to statistical analyses across the board.

Stage	Recommendation
Planning	Write the methods section in advance of data collection
	Distinguish between exploratory and confirmatory research
	Specify the goal; estimation, testing, or both
	If the goal is testing, decide on one-sided or two-sided procedure
	Choose a statistical model
	Determine which model checks will need to be performed
	Specify how to deal with possible model violations
	Choose a prior distribution
	Consider how to assess the impact of prior choices on the inferences
	Specify the sampling plan
	Consider a Bayes factor design analysis
	Preregister the analysis plan for increased transparency
Executing	Check the quality of the data (e.g., assumption checks)
	Annotate the JASP output
Interpreting	Beware of the common pitfalls
	Use the correct interpretation of Bayes factor and credible interval
	When in doubt, ask for advice (e.g., on the JASP forum)
Reporting	Mention the goal of the analysis
	Include a plot of the prior and posterior distribution, if available
	If testing, report the Bayes factor, including its subscripts
	If estimating, report the posterior median and x% credible interval
	Include which prior settings were used
	Justify the prior settings (particularly for informed priors for testing)
	Discuss the robustness of the result
	If relevant, report the results from both estimation and testing
	Refer to the statistical literature for details about the analyses used
	Consider a sequential analysis
	Report the results any multiverse analyses, if conducted
	Make the `.jasp` file and data available online

Table 2.1: A summary of the guidelines for the different stages of a Bayesian analysis, with a focus on analyses conducted in JASP. Note that the stages have a predetermined order, but the individual recommendations can be rearranged where necessary.

Chapter 3

Multiple Perspectives on Inference for two Simple Statistical Scenarios

Abstract

When data analysts operate within different statistical frameworks (e.g., frequentist versus Bayesian, emphasis on estimation versus emphasis on testing), how does this impact the qualitative conclusions that are drawn for real data? To study this question empirically we selected from the literature two simple scenarios –involving a comparison of two proportions and a Pearson correlation– and asked four teams of statisticians to provide a concise analysis and a qualitative interpretation of the outcome. The results showed considerable overall agreement; nevertheless, this agreement did not appear to diminish the intensity of the subsequent debate over which statistical framework is more appropriate to address the questions at hand.

3.1 Introduction

When analyzing a specific data set, statisticians usually operate within the confines of their preferred inferential paradigm. For instance, frequentist statisticians interested in hypothesis testing may report p-values, whereas those interested in estimation may seek to draw conclusions from confidence intervals. In the Bayesian realm, those who wish to test hypotheses may use Bayes factors and those who wish to estimate parameters may report credible intervals. And then there are likelihoodists, information-theorists, and machine-learners — there exists a diverse collection of statistical approaches, many of which are philosophically incompatible.

Moreover, proponents of the various camps regularly explain why their position is the most exalted, either in practical or theoretical terms. For instance, in a well-known article ‘Why Isn’t Everyone a Bayesian?’, Bradley Efron claimed that

This chapter is published as van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.–J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician, 73*, 328–339. Also available as *PsyArXiv preprint*: https://psyarxiv.com/ue5wb/

"The high ground of scientific objectivity has been seized by the frequentists" (Efron, 1986, p. 4), upon which Dennis Lindley replied that "Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong." (Lindley, 1986, p. 7). Similarly spirited debates occurred earlier, notably between Fisher and Jeffreys (e.g., Howie, 2002) and between Fisher and Neyman. Even today, the paradigmatic debates show no sign of stalling, neither in the published literature (e.g., Benjamin et al., 2018; McShane et al., 2019; R. L. Wasserstein & Lazar, 2016) nor on social media.

The question that concerns us here is purely pragmatic: 'does it matter?' In other words, will reasonable statistical analyses on the same data set, each conducted within their own paradigm, result in qualitatively similar conclusions (Berger, 2003)? One of the first to pose this question was Ronald Fisher. In a letter to Harold Jeffreys, dated March 29, 1934, Fisher proposed that "From the point of view of interesting the general scientific public, which really ought to be much more interested than it is in the problem of inductive inference, probably the most useful thing we could do would be to take one or more specific puzzles and show what our respective methods made of them" (Bennett, 1990, p. 156; see also Howie, 2002, p. 167). The two men then proceeded to construct somewhat idiosyncratic statistical 'puzzles' that the other found difficult to solve. Nevertheless, three years and several letters later, on May 18, 1937, Jeffreys stated that "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (Bennett, 1990, p. 162). Similarly, Edwards et al. (1963) suggested that well-conducted experiments often satisfy Berkson's *interocular traumatic test* – "you know what the data mean when the conclusion hits you between the eyes" (p. 217). Nevertheless, surprisingly little is known about the extent to which, in concrete scenarios, a data analyst's statistical plumage affects the inference.

Here we revisit Fisher's challenge. We invited four groups of statisticians to analyze two real data sets, report and interpret their results in about 300 words, and discuss these results and interpretations in a round-table discussion. The data sets are provided online at https://osf.io/hykmz/ and described below. In addition to providing an empirical answer to the question 'does it matter?', we hope to highlight how the same data set can give rise to rather different statistical treatments. In our opinion, this method variability ought to be acknowledged rather than ignored (for a complementary approach see Silberzahn et al., 2018).[1]

The selected data sets are straightforward: the first data set concerns a 2x2 contingency table, and the second concerns a correlation between two variables. The simplicity of the statistical scenarios is on purpose, as we hoped to facilitate a detailed discussion about assumptions and conclusions that could otherwise have remained hidden underneath an unavoidable layer of statistical sophistication. The full instructions for participation can be found online at https://osf.io/

[1] In contrast to the current approach, Silberzahn et al. (2018) used a relatively complex data set and did not emphasize the differences in interpretation caused by the adoption of dissimilar statistical paradigms.

dg9t7/.

3.2 Data Set I: Birth Defects and Cetirizine Exposure

Study summary

Cetirizine is a non-sedating long-acting antihistamine with some mast-cell stabilizing activity. It is used for the symptomatic relief of allergic conditions, such as rhinitis and urticaria, which are common in pregnant women. In the study of interest, Weber-Schoendorfer & Schaefer (2008) aimed to assess the safety of cetirizine during the first trimester of pregnancy when used. The pregnancy outcomes of a cetirizine group ($n = 181$) were compared to those of the control group ($n = 1685$; pregnant women who had been counseled during pregnancy about exposures known to be non-teratogenic). Due to the observational nature of the data, the allocation of participants to the groups was non-randomized. The main point of interest was the rate of birth defects.[2] Variables of the data set[3] are described in Table 3.1 and the data are presented in Table 3.2.

Cetirizine research question

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? In the next sections each of four data analysis teams will attempt to address this question.

Table 3.1: Variable names and their description.

Variable	Description
CetirizineExposure	Whether exposed to cetirizine
BirthDefect	Whether birth defects were detected
Counts	Count data for each cell

Table 3.2: Cetirizine exposure and birth defects.

	Birth Defects		
Cetirizine Exposure	No	Yes	Total
No	1588	97	1685
Yes	167	14	181
Total	1755	111	1866

[2] The original study focused on Cetirizine-induced differences in major birth defects, spontaneous abortions, and preterm deliveries. We decided to look at all birth defects, because the sample sizes were larger for this comparison and we deemed the data more interesting.

[3] The data set is made available on the OSF repository: https://osf.io/hykmz/.

Analysis and interpretation by Lakens and Hennig

Preamble

Frequentist statistics is based on idealised models of data generating processes. We cannot expect these models to be literally true in practice, but it is instructive to see whether data are consistent with such models, which is what hypothesis tests and confidence intervals allow us to examine. We do appreciate that automatic procedures involving for example fixed significance levels allow us to control error probabilities assuming the model, but given that the models do never hold precisely, and that there are often issues with measurement or selection effects, in most cases we think it is prudent to interpret outcomes in a coarse way rather than to read too much meaning into, say, differences between p-values of 0.047 and 0.062. We stick to quite elementary methodology in our analyses.

Analysis and software

We performed a Pearson's Chi-squared test with Yates' continuity correction to test for dependence between exposure of pregnant women exposed to cetirizine and birth defects using the `chisq.test` function in R software version 3.4.3 (R Development Core Team, 2004). However, because Weber-Schoendorfer & Schaefer (2008) wanted "to assess the safety of cetirizine during the first trimester of pregnancy", their actual research question is whether we can reject the presence of a meaningful effect. We therefore performed an equivalence test on proportions (J. J. Chen et al., 2000) as implemented in the TOSTtwo.prop function in the TOSTER package (Lakens, 2017).

Results and interpretation

The chi-squared test yielded $\chi^2(1, \text{N} = 1866) = 0.817$, $p = .366$, which suggests that the data are consistent with an independence model at any significance level in general use. The answer to the question whether the drug is safe depends on a smallest effect size of interest (when is the difference in birth defects considered too small to matter?). This choice, and the selection of equivalence bounds more generally, should always be justified by clinical and statistical considerations pertinent to the case at hand. In the absence of a discussion of this essential aspect of the study by the authors, and in order to show an example computation, we will test against a difference in proportions of 10%, which, although debatable, has been suggested as a sensible bound for some drugs (see Röhmel, 2001, for a discussion).

An equivalence test against a difference in proportions ($M_{dif} = 0.02$, 95% CI[$-0.02; 0.06$]) of 10% based on Fishers exact z-test was significant, $z = -3.88$, $p < 0.001$. This means that we can reject differences in proportions as large, or larger, than 10%, again at any significance level in general use.

Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects? Based on the current study, there is no evidence that cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Obviously this does not mean that cetirizine is safe; in fact the observed birth defect rate in the cetirizine group is about 2% higher than without exposure, which may or may not be explained by random variation. Is cetirizine during the first trimester of pregnancy 'safe'? If we accept a difference in the proportion of birth defects of 10%, and desire a 5% long run error rate, there is clear evidence that the drug is safe. However, we expect that a cost-benefit analysis would suggest proportions of 5% to be unacceptably high, which is in the 95% confidence interval and therefore well compatible with the data. Therefore, we would personally consider the current data inconclusive.

Analysis and interpretation by Morey and Homer

Fitting a classical logistic model with the binary birth defect outcome predicted from the cetirizine indicator confirmed the non-significant relationship ($p = 0.287$). The point estimate of the effect of taking cetirizine is to increase the odds of the birth defect by only 37%. At the baseline levels of birth defects in the non-cetirizine-exposed sample (approximately 6%), this would amount to about an extra two birth defects in every hundred pregnancies in the cetirizine-exposed group.

There are several problems with taking these data as evidence that cetirizine is safe. The first is the observational nature of the data. We have no way of knowing whether an apparent effect — or lack of effect — reflects confounds. Suppose, though, that we set this question aside and assess the evidence that birth defects are not more common in the cetirizine group. We can use a classical one-sided CI to determine the size of the differences we can rule out. We call the upper bound of the $100(1-\alpha)\%$ CI the "worst case" for that confidence coefficient. Figure 3.1 shows that at 95%, the worst case odds increase is for the cetirizine group is 124%. At 99.5%, the worst case increase is 195%. We can translate this into more a more intuitive metric of numbers of birth defects: at baseline rates of birth defects, these would amount to additional 6 and 10 birth defects per 100, respectively (Figure 3.2).

The large p-value of the initial significance test suggests we cannot rule out that cetirizine group has lower rates of birth defects; the one-sided test assuming a decrease in birth defects as the null would not yield a rejection except at high α levels. But also, the "worst case" analysis using the upper bound of the one-sided CI suggests we also cannot rule out a substantial *increase* in birth defects in the cetirizine group.

We are unsure whether cetirizine is safe, but it seems clear to us that these data do not provide much evidence of its relative safety, contrary to what Weber-Schoendorfer and Schaefer suggest.

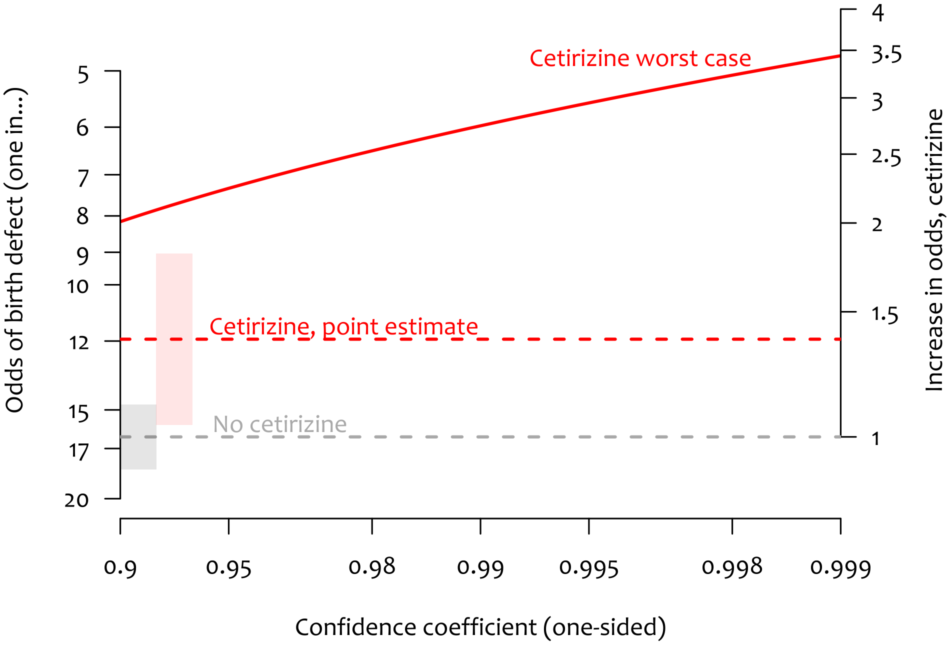


Figure 3.1: Estimates of the odds of a birth defect when no cetirizine (control) was taken during pregnancy and when cetirizine was taken. Horizontal dashed lines and shaded regions show point estimates and standard errors. The solid line labeled "Cetirizine worst case" shows the upper bound of the one-sided CI as a function of the confidence coefficient (x-axis). The right axis shows the estimated increase in odds of a birth defect for the cetirizine group compared to the control group.

Analysis and interpretation by Gronau, van Doorn, and Wagenmakers

We used the model proposed by Kass & Vaidyanathan (1992):

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= \beta - \frac{\psi}{2} \\
\log\left(\frac{p_2}{1-p_2}\right) &= \beta + \frac{\psi}{2} \\
y_1 &\sim \text{Binomial}(n_1, p_1) \\
y_2 &\sim \text{Binomial}(n_2, p_2).
\end{aligned} \tag{3.1}$$

Here, $y_1 = 97$, $n_1 = 1{,}685$, $y_2 = 14$, and $n_2 = 181$, p_1 is the probability of a birth defect in the control group, and p_2 is that probability in the cetirizine group. Probabilities p_1 and p_2 are functions of model parameters β and ψ. Nuisance parameter β corresponds to the grand mean of the log odds, whereas the test-relevant parameter ψ corresponds to the log odds ratio. We assigned β a standard normal prior and used a zero-centered normal prior with standard deviation σ for the log odds ratio ψ. Inference was conducted with Stan (Carpenter et al., 2017; Stan Development Team, 2016) and the `bridgesampling` package (Gronau

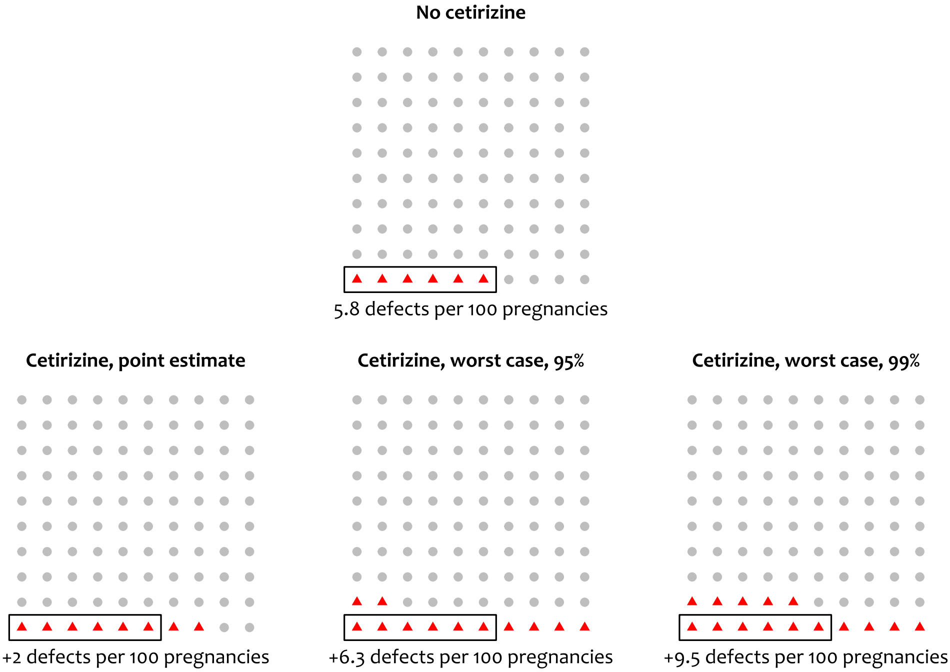


Figure 3.2: Frequency representations of the number of birth defects expected under various scenarios. Top: Expected frequency of birth defects when cetirizine was not taken (control). Bottom-left: Point estimate of the expected frequency of birth defects when cetirizine is taken. Bottom-middle (bottom-right): Upper bound of a one-sided 95% (99%) CI for the expected frequency of birth defects when cetirizine was taken. Because the analysis is intended to be comparative, in the bottom panels the no-cetirizine estimate was assumed to be the truth when calculating the increase in frequency.

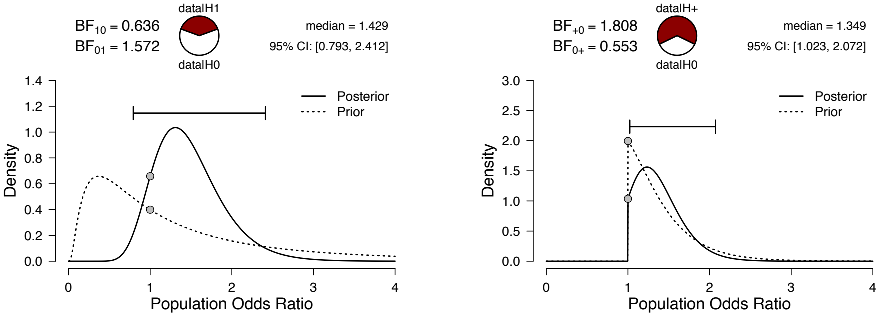
et al., 2020). For ease of interpretation, the results will be shown on the odds ratio scale.

Our first analysis focuses on estimation and uses $\sigma = 1$. The result, shown in the left panel of Figure 3.3, indicates that the posterior median equals 1.429, with a 95% credible interval ranging from 0.793 to 2.412. This credible interval is relatively wide, indicating substantial uncertainty about the true value of the odds ratio.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \psi = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$ we initially use $\sigma = 0.4$ (i.e., a mildly informative prior; Diamond & Kaul, 2004), truncated at zero to respect the fact that cetirizine exposure is hypothesized to cause a *higher* incidence of birth defects: $\mathcal{H}_+ : \psi \sim N^+(0, 0.4^2)$.

As can be seen from the right panel of Figure 3.3, the observed data are predicted about 1.8 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. According to Jeffreys (1961, Appendix B), this level of support is “not worth more than a bare mention”. To investigate the robustness of this result we explored a range of alternative prior choices for σ under $\mathcal{H}_+$, varying it from 0.01 to 2. The results of this sensitivity analysis are shown in Figure 3.4 and reveal that across a wide range of priors, the data never provide more than anecdotal support for one model over the other. When σ is selected post-hoc to maximize the support for $\mathcal{H}_+$ this yields $\text{BF}_{+0} = 1.84$, which, starting from a position of equipoise, raises the probability of $\mathcal{H}_+$ from 0.50 to about 0.65, leaving a posterior probability of 0.35 for $\mathcal{H}_0$.

In sum, based on this data set we cannot draw strong conclusions about whether or not cetirizine exposure during pregnancy is associated with a higher incidence of birth defects. Our analysis shows an 'absence of evidence', not 'evidence for absence'.



(**a**) Estimation results.

(**b**) Testing results.

Figure 3.3: Gronau, van Doorn, and Wagenmakers' Bayesian analysis of the cetirizine data set. The left panel shows the results of estimating the log odds ratio under $\mathcal{H}_1$ with a two-sided standard normal prior. For ease of interpretation, results are displayed on the odds ratio scale. The right panel shows the results of testing the one-sided alternative hypothesis $\mathcal{H}_+ : \psi \sim \mathcal{N}^+(0, 0.4^2)$ versus the null hypothesis $\mathcal{H}_0 : \psi = 0$. Figures inspired by JASP (jasp-stats.org).

Analysis and interpretation by Gelman

I summarized the data with a simple comparison: the proportion of birth defects is 0.06 in the control group and 0.08 in the cetirizine group. The difference is 0.02 with a standard error of 0.02. I got essentially the same result with a logistic regression predicting birth defect: the coefficient of cetirizine is 0.3 with a standard error of 0.3. I performed the analyses in R using rstanarm (code available at https://osf.io/nh4gc/).

I then looked up the article, "The safety of cetirizine during pregnancy: A prospective observational cohort study", by Weber-Schoendorfer & Schaefer (2008) and noticed some interesting things:

1. The published article gives $N = 196$ and 1686 for the two groups, not quite the same as the 181 and 1685 for the "all birth defects" category. I couldn't follow the exact reasoning.[4]

[4] Clarification: the original paper does not provide a rationale for why several participants were excluded from the analysis.

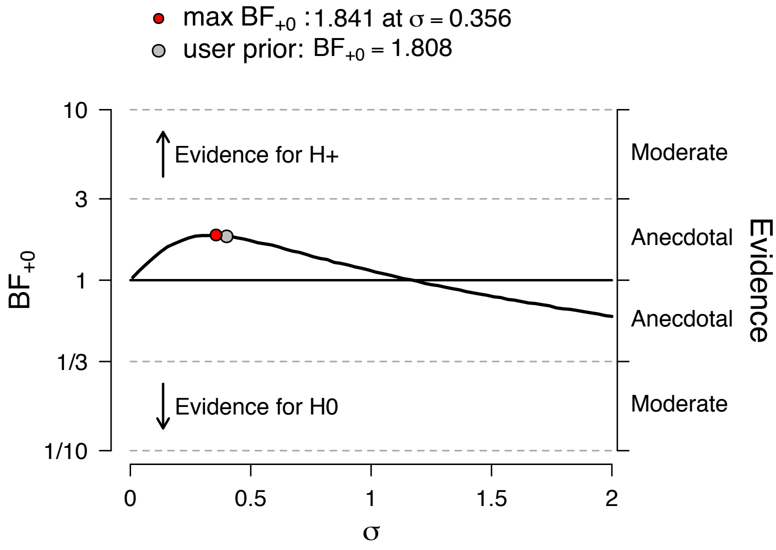


Figure 3.4: Sensitivity analysis for the Bayesian one-sided test. The Bayes factor BF_{+0} is a function of the prior standard deviation σ. Figure inspired by JASP.

2. The two groups differ in various background variables: most notably, the cetirizine group has a higher smoking rate (17% compared to 10%).
3. In the published article, the outcome of focus was “major birth defects”, not the “all birth defects” given for us to study.
4. The published article has a causal aim (as can be seen, for example, from the word “safety” in its title); our assignment is purely observational.

Now the question, “Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?” I have not read the literature on the topic. To understand how the data at hand address this question, I would like to think of the mapping from prior to posterior distribution. In this case, the prior would be the distribution of association with birth defects of all drugs of this sort. That is, imagine a population of drugs, $j = 1, 2, \ldots$, taken by pregnant women, and for each drug, define θ_j as the proportion of birth defects among women who took drug j, minus the proportion of birth defects in the general population. Just based on my general understanding (which could be wrong), I would expect this distribution to be more positive than negative and concentrated near zero: some drugs could be mildly protective against birth defects or associated with low-risk pregnancies, most would have small effects and not be strongly associated with low or high-risk pregnancies, and some could cause birth defects or be taken disproportionately by women with high-risk pregnancies. Supposing that the prior is concentrated within the range (−0.01, +0.01), the data would not add much information to this prior.

To answer, “Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?”, the key question would seem to be whether the drug is more or less likely to be taken by women at higher risk of birth defects. I’m guessing that maternal age is a big predictor here. In the reported study, average age of the exposed and control groups was the same, but I don’t know if that’s generally the case or if the designers of the study were purposely seeking a balanced comparison.

3.3 Data Set II: Amygdalar Activity and Perceived Stress

Study summary

In a recent study published in the *Lancet*, Tawakol et al. (2017) tested the hypothesis that perceived stress is positively associated with resting activity in the amygdala. In the second study reported in Tawakol et al. (2017), $n = 13$ individuals with an increased burden of chronic stress (i.e., a history of post-traumatic stress disorder or PTSD) were recruited from the community, completed a Perceived Stress Scale (i.e., the PSS-10; Cohen et al., 1983) and had their amygdalar activity measured. Variables of the data set[5] are described in Table 3.3, the raw data are presented in Table 3.4, and the data are visualized in Figure 3.5.

Amygdala research question

Do PTSD patients with high resting state amygdalar activity experience more stress? In the next sections each of four data analysis teams will attempt to address this question.

Table 3.3: Variable names and their description.

Variable	**Description**
Perceived stress scale	Participant score on the PSS
Amygdalar activity	Intensity of amygdalar resting state activity

Analysis and interpretation by Lakens and Hennig

Analysis and software

We calculated tests for uncorrelatedness based on both Pearson’s product-moment correlation and Spearman’s rank correlation using the `cor.test` func-

[5] The data set is made available on the OSF repository: https://osf.io/hykmz/.

Table 3.4: Raw data as extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

Perceived Stress Scale	**Amygdalar Activity**
12.0103	5.2418
32.0350	6.8601
22.0296	6.4402
20.0079	5.4620
24.0155	5.4439
24.0155	5.3349
24.0155	5.4216
26.0082	5.5176
28.0120	5.1615
21.9872	4.7114
21.9872	4.1844
20.0138	4.3079
16.0088	3.3015

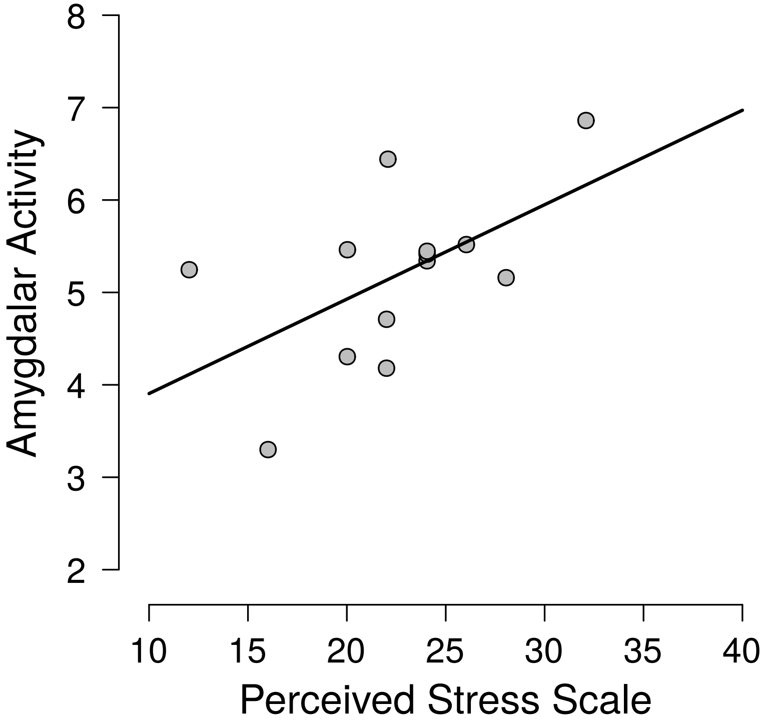


Figure 3.5: Scatter plot of amygdalar activity and perceived stress in 13 patients with PTSD. Data extracted from Figure 5 in Tawakol et al. (2017), with help of Jurgen Rusch, Philips Research Eindhoven.

tion in the stats package in R 3.4.3.

Results and interpretation

The Pearson correlation between perceived stress and resting activity in the amygdala is $r = 0.555$, and the corresponding test yields $p = 0.047$. Although this is just smaller than the conventional 5% level, we do not consider it as clear evidence for nonzero correlation. From the appendix it becomes clear that the reported correlations are exploratory: "Patients completed a battery of self-report measures that assessed variables that may correlate with PTSD symptom severity, including comorbid depressive and anxiety symptoms (MADRS, HAMA) and a well-validated questionnaire Perceived Stress Scale (PSS-10)." Therefore, corrections for multiple comparisons would be required to maintain a given significance level. The article does not provide us with sufficient information to determine the number of tests that were performed, but corrections for multiple comparisons would thus be in order. Consequently, the fairly large observed correlation and the borderline significant p-value can be interpreted as an indication that it may be worthwhile to investigate the issue with a larger sample size, but do not give conclusive evidence. Visual inspection of the data does not give any indication against the validity of using Pearson's correlation, but with $N = 13$ we do not have very strong information regarding the distributional shape. The analogous test based on Spearman's correlation yields $p = 0.062$, which given its weaker power is compatible with the qualitative interpretation we gave based on the Pearson correlation.

Do PTSD patients with high resting state amygdalar activity experience more stress? Based on the current study, we can not conclude that PTSD patients with high resting state amygdalar activity experience more stress. The single $p = 0.047$ is not low enough to indicate clear evidence against the null hypothesis after correcting for multiple comparisons when using an alpha of .05. Therefore, our conclusion is: Based on the desired error rate specified by the authors, we can't reject a correlation of zero between amygdalar activity and participants' score on the perceived stress scale. With a 95% CI that ranges from $r = 0$ to $r = 0.85$, it seems clear that effects that would be considered interesting cannot be rejected in an equivalence test. Thus, the results are inconclusive.

Analysis and interpretation by Morey and Homer

The first thing that should be noted about this data set is that it contains a meager 13 data points. The linear correlation by Tawakol et al. (2017) depends on assumptions that are for all intents and purposes unverifiable with this few participants. Add to this the difficulty of interpreting the independent variable — a sum of ordinal responses — and we are justified being skeptical of any hard conclusions from these data.

Suppose, however, that the relationship between these two variables was best characterized by a linear correlation, and set aside any worries about assumptions. The large observed correlation coupled with the marginal p value should

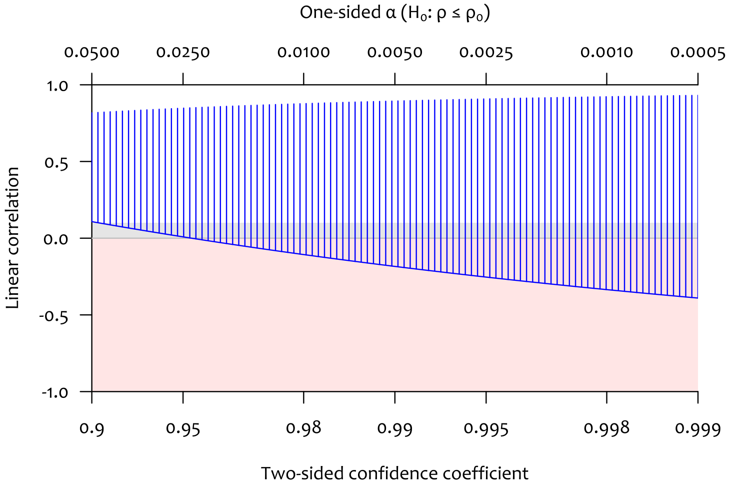


Figure 3.6: Confidence intervals and one-sided tests for the Pearson correlation as a function of the confidence coefficient. The vertical lines represent the confidence intervals (confidence coefficient on lower axis), and the curve represents the value that is just rejected by the one-sided test (α on the upper axis).

signal to us that a wide range of correlations are not ruled out by the data. Consider that the 95% CI on the Pearson correlation is [0.009,0.848]; the 99.5% CI is [−0.254,0.908]. Negligible correlations are not ruled out; due to the small sample size, any correlation from "essentially zero" to "the correlation between height and leg length" (that is, very high) is consistent with these data. The solid curve in Figure 3.6 shows the lower bound of the confidence interval on the linear correlation for a wide range of confidence levels; they are all negligible or even negative.

Finally, the authors did not show that this correlation is selective to the amygdala; it seems to us that interpreting the correlation as evidence for their model requires selectivity. It is important to interpret the correlation in the context of the relationship between amygdala resting state activity, stress, and cardiovascular disease. If one could not show that amygdala resting-state activation showed a substantially higher correlation with stress than other brain regions not implicated in the model, this would suggest that the correlation cannot be used to bolster their case. Given the uncertainty in the estimate of the correlation, there is little chance of being able to show this.

All in all, we're not sure that the information in these thirteen participants is enough to say anything beyond "the correlation doesn't appear to be (very) negative."

Analysis and interpretation by van Doorn, Gronau, and Wagenmakers

We applied Harold Jeffreys's test for a Pearson correlation coefficient ρ (Jeffreys, 1961; Ly, Marsman, & Wagenmakers, 2018) as implemented in JASP (`jasp-stats.org`; JASP Team, 2020).[6] Our first analysis focuses on estimation and assigns ρ a uniform prior from -1 to 1. The result, shown in the left panel of Figure 3.7, indicates that the posterior median equals 0.47, with a 95% credible interval ranging from -0.01 to 0.81. As can be expected with only 13 observations, there is great uncertainty about the size of ρ.

Our second analysis focuses on testing and quantifies the extent to which the data support the skeptic's $\mathcal{H}_0 : \rho = 0$ versus the proponent's $\mathcal{H}_1$. To specify $\mathcal{H}_1$ we initially use Jeffreys's default uniform distribution, truncated at zero to respect the directionality of the hypothesized effect: $\mathcal{H}_+ : \rho \sim U[0,1]$.

As can be seen from the right panel of Figure 3.7, the observed data are predicted about 3.9 times better by $\mathcal{H}_+$ than by $\mathcal{H}_0$. This degree of support is relatively modest: when $\mathcal{H}_+$ and $\mathcal{H}_0$ are equally likely a priori, the Bayes factor of 3.9 raises the posterior plausibility of $\mathcal{H}_+$ from 0.50 to 0.80, leaving a non-negligible 0.20 for $\mathcal{H}_0$.

To investigate the robustness of this result we explored a continuous range of alternative prior distributions for ρ under $\mathcal{H}_+$; specifically, we assigned ρ a stretched Beta(a,a) distribution truncated at zero, and studied how the Bayes factor changes with $1/a$, the parameter that quantifies the prior width and governs the extent to which $\mathcal{H}_+$ predicts large values of r. The results of this sensitivity analysis are shown in Figure 3.8 and confirm that the data provide modest support for $\mathcal{H}_+$ across a wide range of priors. When the precision is selected post-hoc to maximize the support for $\mathcal{H}_+$ this yields $\text{BF}_{+0} = 4.35$, which –under a position of equipose– raises the plausibility of $\mathcal{H}_+$ from 0.50 to about 0.81, leaving a posterior probability of 0.19 for $\mathcal{H}_0$.

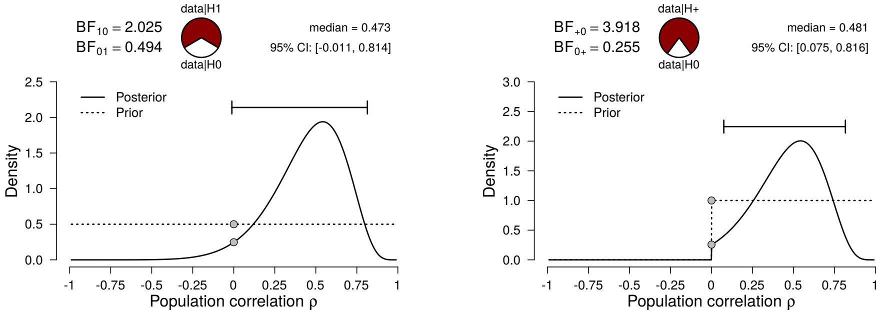
A similar sensitivity analysis could be conducted for $\mathcal{H}_0$ by assuming a 'perinull' (Tukey, 1995, p. 8) — a distribution tightly centered around $\rho = 0$ rather than a point mass on $\rho = 0$. The results will be qualitatively similar.

In sum, the claim that 'PTSD patients with high resting state amygdalar activity experience more stress' receives modest but not compelling support from the data. The 13 observations do not warrant categorical statements, neither about the presence nor about the strength of the hypothesized effect.

Analysis and interpretation by Gelman

I summarized the data with a simple scatterplot and a linear regression of logarithm of perceived stress on logarithm of amygdalar activity, using log scales because the data were all-positive and it seemed reasonable to model a multiplicative relation. The scatterplot revealed a positive correlation and no other

[6] JASP is an open-source statistical software program with a graphical user interface that supports both frequentist and Bayesian analyses.



(a) Estimation results

(b) Testing results

Figure 3.7: van Doorn, Gronau, and Wagenmakers' Bayesian analysis of the amygdala data set. The left panel shows the result of estimating the Pearson correlation coefficient ρ under $\mathcal{H}_1$ with a two-sided uniform prior. The right panel shows the result of testing $\mathcal{H}_0 : \rho = 0$ versus the one-sided alternative hypothesis $\mathcal{H}_+ : \rho \sim U[0,1]$. Figures from JASP.

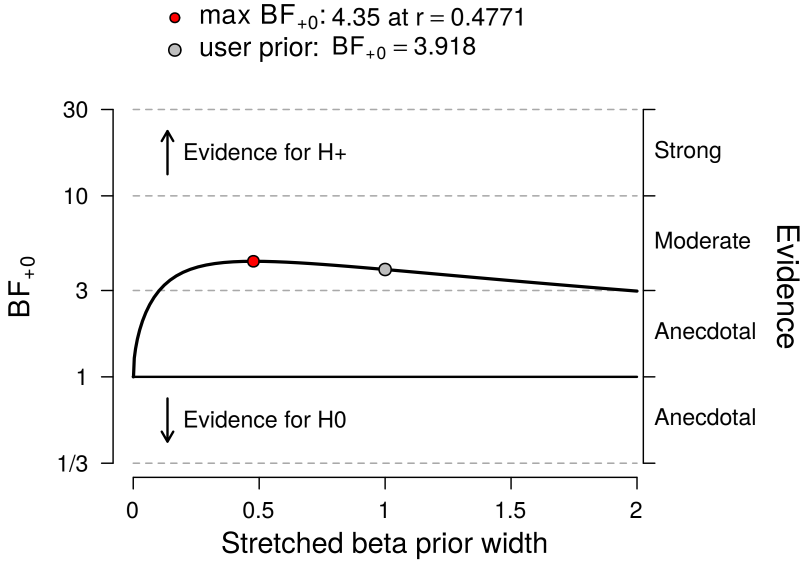


Figure 3.8: Sensitivity analysis for the Bayesian one-sided correlation test. The Bayes factor BF$_{+0}$ is a function of the prior width parameter $1/a$ from the stretched Beta(a,a) distribution. Figure from JASP.

striking patterns, and the regression coefficient was estimated at 0.6 with a standard error of 0.4. I performed the analyses in R using rstanarm (code available at https://osf.io/nh4gc/). and the standard error is based on the median absolute deviation of posterior simulations (see help("mad") in R for more on this).

I then looked up the article, "Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study", by Tawakol et al. (2017). The goal of the research is "to determine whether [the amygdala's] resting metabolic activity predicts risk of subsequent cardiovascular events." Here are some items relevant to our current task:

1. Perceived stress is an intermediate outcome, not the ultimate goal of the study.

2. Any correlation or predictive relation will depend on the reference population. The people in this particular study are "individuals with a history of post-traumatic stress disorder" living in the Boston area.

3. The published article reports that "Perceived stress was associated with amygdalar activity ($r = 0.56$; $p = 0.0485$)." Performing the correlation (or, equivalently, the regression) on the log scale, the result is not statistically significant at the 5% level. This is no big deal given that I don't think that it makes sense to make decisions based on a statistical significance threshold, but it is relevant when considering scientific communication.

4. Comparing my log-scale scatterplot to the raw-scale scatterplot (Figure 5A in the published article), I'd say that the unlogged scatterplot looks cleaner, with the points more symmetrically distributed. Indeed, based on these data alone, I'd move to an unlogged analysis–that is, the estimated correlation of 0.56 reported in the paper.

To address the question, "Do PTSD patients with high resting state amygdalar activity experience more stress?", we need two additional decisions or pieces of information. First, we must decide the population of interest; here there is a challenge in extrapolating from people with PTSD to the general population. Second, we need a prior distribution for the correlation being estimated. It is difficult for me to address either of these issues: as a statistician my contribution would be to map from assumptions to conclusions. In this case, the assumptions about the prior distribution and the assumptions about extrapolation go together, as in both cases we need some sense of how likely it is to see large correlations between the responses to a subjective stress survey and a biomeasurement such as amygdalar activity. It could well be that there is a prior expectation of positive correlation between these two variables in the general population, but that the current data do not provide much information beyond our prior for this general question.

3.4 Round-Table Discussion

As described above, the two data sets have each been analyzed by four teams. The different approaches and conclusions are summarized in Table 5. The discussion was carried out via email and a transcript can be found online at https://osf.io/f4z7x/. Below we highlight and summarize the central elements of a discussion that quickly proceeded from the data analysis techniques in the concrete case to more fundamental philosophical issues. Given the relative agreement among the conclusions reached by different methodological angles, our discussion started out with the following deliberately provocative statement:

In statistics, it doesn't matter what approach is used. As long as you do conduct your analysis with care, you will invariably arrive at the same qualitative conclusion.[7]

In agreement with this claim, Hennig stated that "we all seem to have a healthy skepticism about the models that we are using. This probably contributes strongly to the fact that all our final interpretations by and large agree. Ultimately our conclusions all state that 'the data are inconclusive'. I think the important point here is that we all treat our models as tools for thinking that can only do so much, and of which the limitations need to be explored, rather than 'believing in' our relying on any specific model" (Email 29). On the other hand, Hennig wonders "whether differences between us would've been more pronounced with data that wouldn't have allowed us to sit on the fence that easily" (Email 29) and Lakens wonders "about what would have happened if the data were clearer" (Email 32). In addition, Morey points out that "none of us had anything invested in these questions, whereas almost always analyses are published by people who are most invested" (Email 33). Wagenmakers responds that "we [referring to the group that organized this study] wanted to use simple problems that would not pose immediate problems of either one of the paradigms...[and] we tried to avoid data sets that met Berkson's 'inter-ocular traumatic test' (when the data are so clear that the conclusion hits you right between the eyes) where we would immediately agree" (Email 31). In addition, the focus was on the analyses and discussion as free as possible from other consideration (e.g., personal investment in the questions).

However, differences between the analyses were emphasized as well. First, Morey argued that the differences (e.g., research planning, execution, analysis, etc.) between the Bayesian and frequentist approach are critically important and not easily inter-translatable (Email 2). This gave rise to an extended discussion about the frequentist procedures' dependence of the sampling protocol, which Bayesian procedures lack. While Bayesians such as Wagenmakers see this as a critical objection against the coherent use of frequentist procedures (e.g., in cases where the sampling protocol is uncertain), Hennig contends that one can still interpret the p-value as indicating the compatibility of the data with the null model, assuming a hypothetical sampling plan (see Email 5-11, 16-18, 23, and 24). Second, Lakens speculated that, in the cases where the approaches differ, there "might be more variation within specific approaches, than between" (Email 1). Third, Hennig pointed out that differences in prior specifications could lead

[7] The statement is based on Jeffreys' claims "[a]s 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, 1939, p. 365) and "it would only be once in a blue moon that we [Fisher and Jeffreys] would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (Bennett, 1990, p. 162).

to discrepancies between Bayesians (Email 4) and Homer pointed out that differences in alpha decision boundaries could lead to discrepancies between frequentists' conclusions (Email 11). Finally, Gelman disagreed with most if what had been said in the discussion thus far. Specifically, he said: "I don't think 'alpha' makes any sense, I don't think 95% intervals are generally a good idea, I don't think it's necessarily true that points in 95% interval are compatible with the data, etc etc." (Email 15).

A concrete issue concerned the equivalence test used by Lakens and Hennig for the first data set. Wagenmakers objects that it does not add relevant information to the presentation of a confidence interval (Email 12). Lakens responds that it allows to reject the hypothesis of a >10% difference in proportions at almost any alpha level, thereby avoiding reliance on default alpha levels, which are often used in a mindless way and without attention to the particular case (Email 13).

A more foundational point of contention with the two data sets and their analysis was about the question of how to formulate Bayesian priors. For these concrete cases, Hennig contends that the subject-matter information cannot be smoothly translated into prior specifications (Email 27), which is the reason why Morey and Homer choose a frequenstist approach, while Gelman considers it "hard for me to imagine how to answer the questions without thinking about subject-matter information" (Email 26).

Lakens raised the question of how much the approaches in this paper are representative of what researchers do in general (Email 43 and 44). Wagenmakers' discussion of p-values echoes this point. While Lakens describes p-values as a guide to "deciding to act in a certain way with an acceptable risk of error" and contends many scientists conform to this rationale (Email 32), Wagenmakers has a more pessimistic view. In his experience, the role of p-values is less epistemic than social: they are used to convince referees and editors and to suggest that the hypothesis in question is true (Email 37). Also Hennig disagrees with Lakens, but from a frequentist point of view: they should not guide binary accept/reject-decisions, they just indicate the degree to which the observed data is compatible with the model specified by the null hypothesis (Email 34).

The question of how data analysis relates to learning, inference and decision-making was also discussed regarding the merits (and problems) of Bayesian statistics. Hennig contends that there can be "some substantial gain from them [priors] only if the prior encodes some relevant information that can help the analysis. However, here we don't have such information" and the Bayesian "approach added unnecessary complexity" (Email 23). Wagenmakers reply is that "prior is not there to help or hurt an analysis: it is there because without it, the models do not make predictions; and without making predictions, the hypotheses or parameters cannot be evaluated" and that "the approach is more complex, but this is because it includes some essential ingredients that the classical analysis omits" (Email 24). In fact, he insinuates that frequentists learn from data through "Bayes by stealth": the observed p-values, confidence intervals and other quantities are used to update the plausibility of the models in an "inexact and intuitive" way. "Without invoking Bayes' rule (by stealth) you can't learn much from a classical analysis, at least not in any formal sense." (ibid.) According to Hennig there is more to learning than "updating epistemic probabilities of certain parameter values being true. For example I find it informative to learn that 'Model X is compatible with the data' " (Email 25). However, Wagenmakers considers Hennig's example of learning as a synonymous to observing. Though he agrees that "it is informative to know when a specific model is or is not compatible with data; to learn anything, however, by its very definition requires some form of knowledge updating" (Email 30). This discussion evolved, finally, into a general discussion about the philosophical principles and ideas underlying schools of statistical inference. Ironically, both Lakens (decision-theoretically oriented frequentism) and Gelman (falsificationist Bayesianism) claim the philosophers of science Karl Popper and Imre Lakatos, known for their ideas of accumulating scientific progress through successive falsification attempts, as one of their primary inspirations, although they spell out their ideas in a different way (Emails 42 and 45).

Hennig and Lakens also devoted some attention to improving statistical practice and either directly or indirectly questioned the relevance of foundational debates. Concerning the above issue with using p-values for binary decision-making, Hennig suspects that "if Bayesian methodology would be in more widespread use, we'd see the same issue there ... and then 'reject' or 'accept' based on whether a posterior probability is larger than some mechanical cutoff" (Email 34) and "that much of the controversy about these approaches concerns naive 'mechanical' applications in which formal assumptions are taken for granted to be fulfilled in reality" (Email 29). In addition, Lakens points out that "whether you use one approach to statistics or another doesn't matter anything in practice. If my entire department would use a different approach to statistical inferences (now everyone uses frequentist hypothesis testing) it would have basically zero impact on their work. However, if they would learn how to better use statistics, and apply it in a more thoughtful manner, a lot would be gained" (Email 32). Homer provides an apt conclusion to this topic by stating "I think a lot of problems with research happen long before statistics get involved. E.g. Issues with measurement; samples and/or methods that can't answer the research question; untrained or poor observers" (Email 35).

Finally, an interesting distinction was made between a prescriptive use of statistics and a more pragmatic use of statistics. As an illustration of the latter, Hennig has a more pragmatic perspective on statistics, because a strong prescriptive view (i.e., fulfillment of modeling assumptions as a strict requirement for statistical inference) would often mean that we can't do anything in practice (Email 2). To clarify this point, he adds: "Model assumptions are never literally fulfilled so the question cannot be whether they are..., but rather whether there is information or features of the data that will mislead the methods that we want to apply" (Email 23). The former is illustrated by Homer: "I think that assumptions are critically important in statistical analysis. Statistical assumptions are the axioms which allow a flow of mathematical logic to lead to a statistical inference. There is some wiggle room when it comes to things like 'how robust is this test, which assumes normality, to skew?' but you are on far safer ground when all the assumptions are/appear to be met. I personally think that not even checking the plausibility of assumptions is inexcusable sloppiness (not that I feel anyone here suggested otherwise)" (Email 11). From what has been said in the discussion, there is general consensus that not all assumptions need to be met and not all rules need to be strictly followed. However, there is great disagreement about which assumptions are important; which rules should be followed and how strictly; and what can be interpreted from the results when (it is uncertain if) these rules and assumptions are violated. The interesting subtleties of this topic and the discussants' views on use of statistics can be read in the online supplement (model assumptions: Emails 4, 11, 23, and 24; alpha-levels, *p*-values, Bayes factors, and decision procedures: Emails 2, 9, 11, 14, 15, 24, and 31-40; sampling plan, optional stopping, and conditioning on the data: Emails 2, 5-11, 16-18, 23, and 24).

In summary, dissimilar methods were used that resulted in similar conclusions and varying views were discussed on how statistical methods are used and should be used. At times it was a heated debate with interesting arguments from both (or more) sides. As one might expect, there was disagreement about particularities of procedures and consensus on the expectation that scientific practice would be improved by better general education on the use of statistics.

3.5 Concluding Comments

Four teams each analyzed two published data sets. Despite substantial variation in the statistical approaches employed, all teams agreed that it would be premature to draw strong conclusions from either of the data sets. Adding to this cautious attitude are concerns about the nature of the data. For instance, the first data set was observational, and the second data set may require a correction for multiplicity. In addition, for each scenario, the research teams indicated that more background information was desired; for instance, "when is the difference in birth defects considered too small to matter?"; "what are the effects for similar drugs?"; "is the correlation selective to the amygdala?"; and "what is the prior distribution for the correlation?". Unfortunately, in the routine use of statistical procedures such information is provided only rarely.

It also became evident that the analysis teams not only needed to make assumptions about the nature of the data and any relevant background knowledge, but they also needed to interpret the research question. For the first data set, for instance, the question was formulated as "Is cetirizine exposure during pregnancy associated with a higher incidence of birth defects?". What the original researchers wanted to know, however, is whether or not cetirizine is safe – this more general question opens up the possibility of applying alternative approaches, such as the equivalence test, or even a statistical decision analysis: should pregnant women be advised not to take cetirizine? We purposefully tried to steer clear from decision analyses because the context-dependent specification of utilities adds another layer of complexity and requires even more background knowledge than was already demanded for the present approaches. More generally, the formulation of our research questions was susceptible to multiple interpretation: as tests against a point null, as tests of direction, or as tests of effect size. The goals of a statistical analysis can be many, and it is important to define them unambiguously – again, the routine use of statistical procedures almost never conveys this information.

Despite the (unfortunately near-universal) ambiguity about the nature of the data, the background knowledge, and the research question, each analysis team added valuable insights and ideas. This reinforces the idea that a careful statistical analysis, even for the simplest of scenarios, requires more than a mechanical application of a set of rules; a careful analysis is a process that involves both skepticism and creativity. Perhaps popular opinion is correct, and statistics is difficult. On the other hand, despite employing widely different approaches, all teams nevertheless arrived at a similar conclusion. This tentatively supports the Fisher-Jeffreys conjecture that, regardless of the statistical framework in which they operate, careful analysts will often come to similar conclusions.

Table 3.5: *Overview of the approaches and results of the research teams.*

Research Team	Data Set I: Cetirizine & Birth Defects	Data Set II:Amygdalar Activity
Lakens & Hennig	– Frequentist test of equivalence – 10% equivalence region – Data deemed inconclusive	– Frequentist correlation, $p = .047$ – Concerns about multiple comparisons – Data deemed inconclusive
Morey & Homer	– Frequentist logistic model, $p = .287$ – Observational, so possible confounds – Data deemed inconclusive	– Frequentist correlation, $p = .047$ – Small n means assumptions unverifiable – Is the effect specific for amygdala? – Data deemed inconclusive
Gronau, Van Doorn, & Wagenmakers	– Default Bayes factor $\text{BF}_{01} = 1.6$ – Evidence “not worth more than a bare mention” – Data deemed inconclusive	– Default Bayes factor $\text{BF}_{10} = 2$ – Small n – Data deemed inconclusive
Gelman	– Bayesian analysis needs good prior – Data likely to be inconclusive – A key question is who takes the drug in the population	– Bayesian analysis needs good prior – Problems with generalizing to population – Data likely to be inconclusive

Chapter 4

An In-Class Demonstration of Bayesian Inference

Abstract

Sir Ronald Fisher's venerable experiment "The Lady Tasting Tea" is revisited from a Bayesian perspective. We demonstrate how a similar tasting experiment, conducted in a classroom setting, can familiarize students with several key concepts of Bayesian inference, such as the prior distribution, the posterior distribution, the Bayes factor, and sequential analysis.

4.1 Introduction

Over 80 years ago, Sir Ronald Fisher conducted the famous experiment "The Lady Tasting Tea" in order to test whether his colleague, Dr. Muriel Bristol, could taste if the tea infusion or the milk had been added to the cup first (Fisher, 1935, p. 11). Dr. Bristol was presented with eight cups of tea and the knowledge that four of these had the milk poured in first. Dr. Bristol was then asked to identify these four cups. Fisher analyzed the results using null hypothesis significance testing: (1) assume the null hypothesis to be true (i.e., Dr. Bristol lacks any ability to discriminate the cups); (2) calculate the probability of encountering results at least as extreme as those observed; (3) if that probability is sufficiently low, consider the null hypothesis discredited. This probability is now known as the p-value and it features in many statistical analyses across empirical sciences such as biology, economics, and psychology (for recent critique, see R. Wasserstein & Lazar, 2016; Benjamin et al., 2018).

Decades later, Dennis Lindley (1993) used an experimental procedure similar to that of Fisher to highlight some limitations of the p-value paradigm. Specifically, the calculation of the p-value depends on the sampling plan, that is, the *intention* with which the data were collected. Consider the Lindley setup: the lady is offered six pairs of cups, where each pair consists of a cup where the tea was poured first, and a cup where the milk was poured first. She is then asked to judge, for each pair, which cup has had the tea added first. A possible outcome is

This chapter is published as van Doorn, J.B., Matzke, D., & Wagenmakers, E.–J. (2020). An In-Class Demonstration of Bayesian Inference. *Psychology Learning and Teaching, 19*, 36–45.

the sequence RRRRRW, indicating that she was right for the first five pairs, and wrong for the last pair. However, as Lindley demonstrated, the original sampling plan is crucial in calculating the *p*-value. Was the goal to have the lady taste six pairs of cups –no more, no less– or did she need to continue until she made her first mistake? The observed data are compatible with either sampling plan; yet in the former case, the *p*-value equals 0.109, whereas in the latter case the *p*-value equals 0.031. The difference lies in the inclusion of more extreme cases. In the 'test six cups' plan, the only more extreme outcome is RRRRRR (i.e., the binomial sampling distribution), whereas for the 'test until error' plan the more extreme outcomes include sequences such as RRRRRRW and RRRRRRRW (i.e., the negative binomial sampling distribution). It seems undesirable that the *p*-value depends on hypothetical outcomes that are in turn determined by the sampling plan. Harold Jeffreys summarized: "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; see also Berger & Wolpert, 1988).

In this chapter we revisit Fisher's experimental paradigm to demonstrate several key concepts of Bayesian inference, specifically the prior distribution, the posterior distribution, the Bayes factor, and sequential analysis. Furthermore, we highlight the advantages of Bayesian inference, such as its straightforward interpretation, the ability to monitor the result in real-time, and the irrelevance of the sampling plan. For concreteness, we analyze the outcome of a tasting experiment that featured 57 staff members and students of the Psychology Department at the University of Amsterdam; these participants were asked to distinguish between the alcoholic and non-alcoholic version of the Weihenstephaner Hefeweissbier, a German wheat beer. We describe how classroom tasting experiments can acquaint students with Bayesian inference, noting that beer can be substituted with anything else suitable (e.g., red and green M&M's, Coca Cola and Pepsi, decaf and regular coffee). We analyze and present the results in the open-source statistical software JASP (JASP Team, 2020).

4.2 The Tasting Experiment

On a Friday afternoon, May 12th 2017, an informal beer tasting experiment took place at the Psychology Department of the University of Amsterdam. The experimental team consisted of three members: one to introduce the participants to the experiment and administer the test, one to pour the drinks, and one to process the data. Participants tasted two small cups filled with Weihenstephaner Hefeweissbier, one with alcohol and one without, and indicated which one contained alcohol. Participants were also asked to rate the confidence in their answer (measured on a scale from 1 to 100, with 1 being completely clueless and 100 being absolutely sure), and to rate the two beers in tastiness (measured on a scale from 1 to 100, with 1 being the worst beer ever and 100 being the best beer ever). The experiment was double-blind, such that the person administering the test and interacting with the participants did not know which of the two cups contained alcohol. For ease of reference, each cup was labeled with a random integer between 1 and 500, and each integer corresponded either to the alcoholic or non-alcoholic beer. A coin was flipped to decide which beer was tasted first. The setup was piloted with 9 participants; subsequently, we tested as many people as possible within an hour, and also recorded which of the two beers was tasted first. On average, testing took approximately 30 seconds per participant, yielding a total of 57 participants. Of the 57 participants, 42 (73.7%) correctly identified the beer that contained alcohol: in other words, there were $s = 42$ successes and $f = 15$ failures.[1]

4.3 Theoretical Analysis

In order to assess statistically whether and to what extent participants were able to discriminate between alcoholic and non-alcoholic beer we apply the binomial model, where the rate parameter θ governs the probability of a correct response for each of the participants. Chance performance corresponds to $\theta = 1/2$. Above-chance performance corresponds to values of θ higher than $1/2$, with $\theta = 1$ indicating perfect performance.

In the Bayesian framework, we start by specifying a prior distribution. The prior distribution quantifies our beliefs about the parameter of interest before seeing the data. For convenience, we may specify a beta distribution: a probability distribution on the domain $[0,1]$ governed by two shape parameters, a and b. Setting $a = b = 1$ yields a uniform distribution, and implies that all values of rate θ are equally likely a priori. Setting $a > b$ assigns more prior probability mass to values of θ higher than $1/2$, whereas setting $a < b$ assigns more mass to values of θ lower than $1/2$.[2]

The beta prior distribution is then updated to a posterior distribution using Bayes' rule, such that values of θ that predicted the data well receive a boost in credibility, whereas values of θ that predicted the data poorly suffer a decline (Rouder & Morey, 2019; Wagenmakers et al., 2016):

$$\underbrace{p(\theta \mid s, f)}_{\text{Posterior}} = \underbrace{p(\theta)}_{\text{Prior}} \times \underbrace{\frac{\overbrace{p(s, f \mid \theta)}^{\text{Prediction for specific } \theta}}{p(s, f)}}_{\substack{\text{Average prediction} \\ \text{across all } \theta\text{'s}}} . \tag{4.1}$$

The right-most term is the predictive updating factor that quantifies the change from prior to posterior beliefs brought about by the data. This predictive updating factor indicates how well each value of θ predicted the data, relative to the average prediction across all values of θ. When a specific value of θ pre-

[1] Three video recordings of the procedure are available at https://osf.io/428pb/

[2] A Shiny app to examine the shape of different beta distributions is available at http://shinyapps.org/, under "A first lesson in Bayesian inference".

dicted the data better than average, the posterior density at that point will be higher than the prior density.

We used the binomial likelihood to assess the quality of each value's prediction (i.e., the likelihood of observing s successes and f failures, given a specific value of θ). Because we used the binomial likelihood and a beta prior distribution, the updated posterior distribution will also be a beta distribution – a property known as conjugacy (Gelman, 2013).

The obtained posterior distribution can be used for both parameter estimation and hypothesis testing. For parameter estimation, either a point estimate or an interval estimate can be obtained. Commonly used point estimates include the posterior median and posterior mean. Interval estimation can be done with a so-called credible interval, which is an interval that contains x% of the posterior mass[3] and can be interpreted as follows: there is an x% probability that the true parameter lies in this interval. For example, if we obtain a 95% credible interval of [0.6, 0.9] for θ, we can be 95% sure that the true value of θ lies between 0.6 and 0.9.

The posterior distribution can also be used for hypothesis testing, where the traditional goal is to examine specific values of θ. For instance, we can test the hypothesis $\mathcal{H}_0 : \theta = 1/2$ (i.e., chance performance) by comparing its predictive adequacy to that of an alternative hypothesis $\mathcal{H}_1 : \theta \neq 1/2$. In other words, $\mathcal{H}_0$ represents the idealized position of a skeptic who believes that the data can be accounted for purely by chance. This 'chance only' model is pitted against an alternative that allows θ to take on values different from 1/2.

As before, hypotheses that predict the data well receive a boost in credibility, whereas hypotheses that predict the data poorly suffer a decline. In the Bayesian framework, hypothesis testing is traditionally achieved through the Bayes factor (Kass & Raftery, 1995; Etz & Wagenmakers, 2017).[4] The Bayes factor can be seen as a weighing of one hypothesis' predictive quality relative to that of another. The following equation illustrates this principle, and is very similar to equation (4.1):

$$\underbrace{\frac{p(\mathcal{H}_1 \mid s,f)}{p(\mathcal{H}_0 \mid s,f)}}_{\substack{\text{Posterior beliefs}\\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs}\\ \text{about hypotheses}}} \times \underbrace{\frac{p(s,f \mid \mathcal{H}_1)}{p(s,f \mid \mathcal{H}_0)}}_{\text{Bayes factor}} \tag{4.2}$$

It is important to note here that the Bayes factor is a *relative* metric of the hypotheses' predictive quality. For instance, if the Bayes factor equals 5, this means that the data are 5 times as likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. The relative nature of the Bayes factor stands in stark contrast with the frequentist paradigm, where only the null hypothesis is under consideration.

[3] Two popular ways of creating a credible interval are the highest density credible interval, which is the narrowest interval containing the specified mass, and the central credible interval, which is created by cutting off $\frac{100-x}{2}$% from each of the tails of the posterior distribution. In the remainder of this chapter, we use the central credible interval.

[4] For an alternative procedure to test parameter values, see for instance Kruschke (2011, 2018).

The computation of the Bayes factor is usually not straightforward; however, when the two hypotheses are nested, a convenient computational shortcut can be used, known as the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The shortcut entails that the Bayes factor equals the ratio of the prior density and the posterior density at the test value θ_0. For instance, in the current study, $\theta_0 = 1/2$ so we have the following ratio:

$$\mathrm{BF}_{10} = \frac{p(\theta = 1/2)}{p(\theta = 1/2 \mid \mathrm{data})}, \tag{4.3}$$

where the numerator indicates the prior ordinate and the denominator indicates the posterior ordinate evaluated at the test value, $\theta = 1/2$. BF denotes the Bayes factor, and the subscript indicates which hypotheses are compared. BF_{10} indicates the Bayes factor in favor of $\mathcal{H}_1$ $\left(\text{i.e., } \frac{p(\mathrm{data}|\mathcal{H}_1)}{p(\mathrm{data}|\mathcal{H}_0)}\right)$, whereas BF_{01} indicates the Bayes factor in favor of $\mathcal{H}_0$ $\left(\text{i.e., } \frac{p(\mathrm{data}|\mathcal{H}_0)}{p(\mathrm{data}|\mathcal{H}_1)}\right)$. For instance, if $\mathrm{BF}_{10} = 1/5$, then $\mathrm{BF}_{01} = 5$.

We stress that the mathematical details are not critical for students' understanding of the Bayesian procedures. The following section shows how the example and the associated graphs suffice to clarify the key Bayesian concepts at an intuitive level.

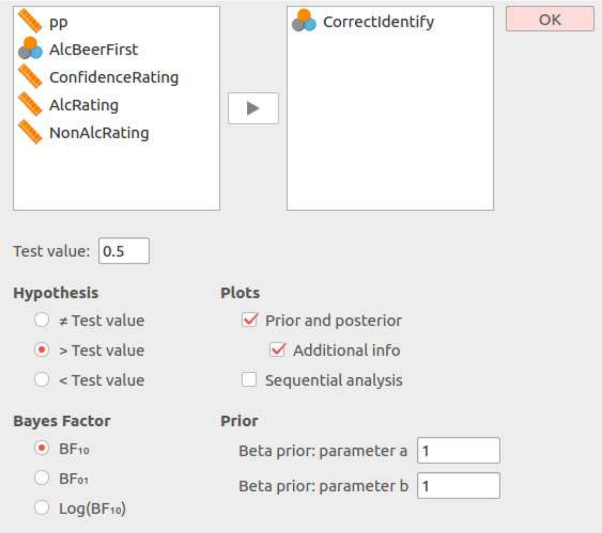
4.4 Bayesian Inference with JASP

When the statistical explanation does not resonate with students, a practical demonstration of the analysis might. This can be done with the statistical software JASP, which offers a graphical user interface for conducting Bayesian (and frequentist) analyses. In order to analyze the collected data, the Bayesian binomial test can be used, which can be found under the menu labeled "Frequencies". Several settings are available for the binomial test, allowing students to explore different analysis choices. Figure 4.1 presents a screenshot of the options panel in JASP. For this analysis, we specify a test value of 1/2 (i.e., chance performance), and $a = b = 1$ for the prior distribution of θ under $\mathcal{H}_1$. Note that in a sensitivity or robustness analysis, other values for a and b may be explored to assess their impact on the posterior distribution.

The null hypothesis postulates that participants performed at chance level, whereas the alternative hypothesis postulates that this is not the case. For instance, in the case of two-sided hypothesis testing, the hypotheses are specified as follows

$$\begin{aligned}\mathcal{H}_0 &: \theta = 1/2\\ \mathcal{H}_1 &: \theta \sim \mathrm{beta}(1, 1).\end{aligned} \tag{4.4}$$

However, since we wish to test whether or not participants' discriminating ability exceeds chance, we can specify the alternative hypothesis to allow only values of



pp
AlcBeerFirst
ConfidenceRating
AlcRating
NonAlcRating

CorrectIdentify

Test value: 0.5

Hypothesis
○ ≠ Test value
● > Test value
○ < Test value

Plots
☑ Prior and posterior
☑ Additional info
□ Sequential analysis

Bayes Factor
● BF_{10}
○ BF_{01}
○ $Log(BF_{10})$

Prior
Beta prior: parameter a 1
Beta prior: parameter b 1

Figure 4.1: The input panel for the Bayesian binomial test in JASP. The upper left box displays all available variables. The upper right box displays the tested variables. Below are other options, such as setting the test value, the alternative hypothesis, and the shape parameters of the beta prior.

θ greater than $1/2$ (note the '+' in the subscript):

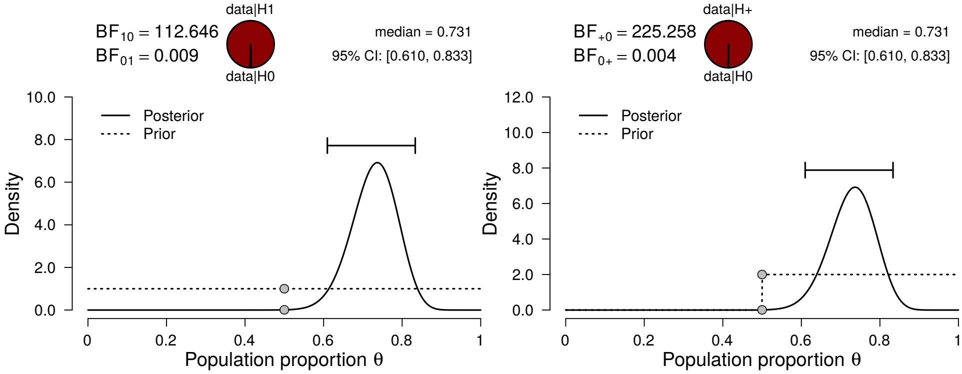
$$\mathcal{H}_+ : \theta \sim \text{beta}(1,1)\text{I}(1/2, 1), \tag{4.5}$$

where I indicates truncation of the beta distribution to the interval $[1/2, 1]$.

Figure 4.2 illustrates the results of the binomial test. The left panel shows the prior and the posterior distribution of θ for the two-sided alternative hypothesis, along with the median and credible interval of the posterior distribution. The posterior median equals 0.731 and the 95% credible interval ranges from 0.610 to 0.833, indicating a substantial deviation of θ from $1/2$. For each value of θ, the change from prior distribution to posterior distribution is quantified by predictive adequacy: for those values of θ that predict the data better than average, the posterior density exceeds the prior density (see equation (4.1))). The left panel shows inference for the two-sided alternative hypothesis (i.e., $\mathcal{H}_1 : \theta \neq 1/2$) compared to the null hypothesis (i.e., $\mathcal{H}_0 : \theta = 1/2$). The resulting Bayes factor is 122.65 in favor of the alternative hypothesis, that is, the observed data are about 112.65 times more likely to occur under $\mathcal{H}_1$ than under $\mathcal{H}_0$.

The right panel shows inference for the one-sided positive hypothesis (i.e., $\mathcal{H}_+ : \theta \geq 1/2$) compared to the null hypothesis: the resulting Bayes factor is 225.26 in favor of the alternative hypothesis. Note that the posterior distribution itself has hardly changed: the posterior median still equals 0.731 and the 95% credible interval ranges from 0.610 to 0.833. Because virtually all posterior mass was already to the right of $1/2$ in the two-sided case, the posterior distribution was virtually unaffected by changing from $\mathcal{H}_1$ to $\mathcal{H}_+$. However, in the right panel, $\mathcal{H}_+$ only predicts values greater than $1/2$, which is reflected in the prior distribution:

all prior mass is now located in the interval (1/2, 1), and as a result, the prior mass in the interval (1/2, 1) has doubled. Since the posterior density at the point of testing is the same in both panels, but the prior density is doubled in the right panel, the Bayes factor for the directed hypothesis doubles as well.



(a) $\mathcal{H}_1 : \theta \sim \text{beta}(1,1)$ **(b)** $\mathcal{H}_1 : \theta \sim \text{beta}(1,1)\text{I}(1/2,1)$

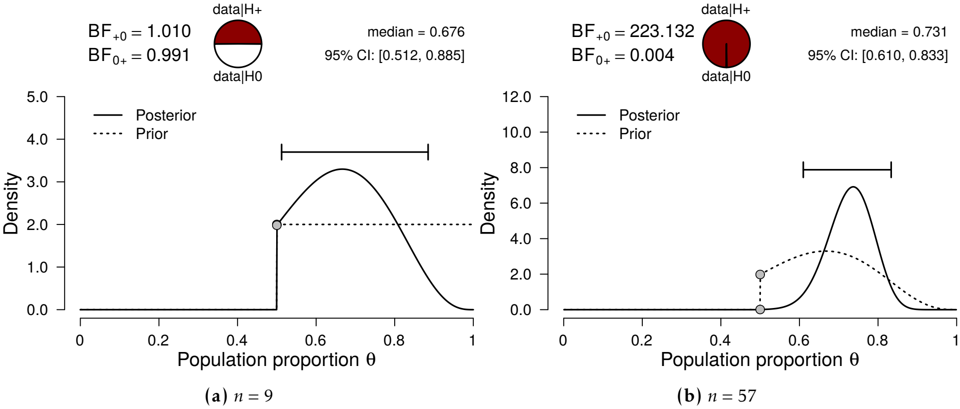
Figure 4.2: Bayesian binomial test for the rate parameter θ. The probability wheel at the top illustrates the ratio of the evidence in favor of the two hypotheses. The two gray dots indicate the prior and posterior density at the test value - the ratio of these is the Savage-Dickey density ratio. The median and the 95% credible interval of the posterior distribution are shown in the top right corner. The left panel shows the two-sided test and the right panel shows the one-sided test. Both figures from JASP.

The experimental procedure also highlights one of the main strengths of Bayesian inference: real-time monitoring of the incoming data. As the data accumulate, the analysis can be continuously updated to include the latest results. In other words, the results may be updated after every participant, or analyzed all at one, without affecting the resulting inference. To illustrate this, we can use Equation 4.1 to compute the posterior distribution for the first 9 participants of the experiment for which $s = 6$ and $f = 3$. Specifying the same beta prior distribution as before, namely a truncated beta distribution with shape parameters $a = b = 1$, and combining this with the data, yields a truncated beta posterior distribution with shape parameters $a = 6 + 1 = 7$ and $b = 3 + 1 = 4$.[5] The resulting posterior distribution is presented in the left panel of Figure 4.3. Now, we can take the remaining 48 participants and conduct the Bayesian binomial test. Because we already have knowledge about the population's rate parameter θ, namely the results of the first 9 participants, we can incorporate this in the analysis through the prior distribution, following Lindley's maxim "today's posterior is tomorrow's prior" (Lindley, 1972).

In this case, we can specify a truncated beta prior distribution with $a = 7$ and $b = 4$, and update this with the data of the remaining 48 participants using

[5] Due to the property of conjugacy, where the posterior distribution has the same form as the prior distribution, the shape parameters of the beta posterior distribution can be obtained by summing the a and b parameters of the prior distribution with the observed number of successes and failures, respectively.

Equation 4.1. Out of the 48 participants, 36 were correct, and 12 were incorrect. Updating the prior distribution with this data yields a posterior distribution with shape parameters $a = 7 + 36 = 43$ and $b = 4 + 12 = 16$, which is exactly the same posterior distribution obtained when analyzing the full data set at once. This two-step procedure is illustrated in Figure 4.3. The left panel shows the prior distribution (i.e., the truncated beta distribution with $a = 1, b = 1$) and the posterior distribution for the first 9 participants. The right panel shows the inference for the remaining 48 participants, while incorporating the knowledge gained from the first 9 participants in the prior distribution by specifying a truncated beta distribution with $a = 7, b = 4$.



(a) $n = 9$

(b) $n = 57$

Figure 4.3: Sequential updating of the Bayesian binomial test. The left panel shows results from a one-sided Bayesian binomial test for the first $n = 9$ participants ($s = 6$, $f = 3$). The shape parameters of the truncated beta prior were set to $a = 1$ and $b = 1$. The right panel shows results from a one-sided binomial test for the remaining 48 participants. Here, the specified prior is the posterior distribution from the left panel: a truncated beta distribution with $a + s = 7$ and $b + f = 4$. The resulting posterior distribution is identical to the posterior distribution in Figure 4.2b. In order to obtain the total Bayes factor in Figure 4.2b, the component Bayes factors in Figures 4.3a and 4.3b can be multiplied (Jeffreys, 1937). Both figures from JASP.

The ability to monitor the data in real-time and update the inference accordingly prevents wasteful data collection: if there is sufficient evidence to discredit either hypothesis with 50 observations, why collect another 10? Wasteful testing is a serious issue, and monitoring the evidence is important in fields such as medicine, biology, and industry. The Bayesian framework for planning experiments is discussed in more detail by Rouder (2014), Schönbrodt et al. (2017) and Schönbrodt & Wagenmakers (2018). Figure 4.4 shows the evolution of the Bayes factor as more data are collected. Initially the evidence is inconclusive, but after 30 participants the evidence increasingly supports $\mathcal{H}_1$.

4.5 Concluding Comments

This chapter has outlined a teaching tool for familiarizing students with the basics of Bayesian inference. The educational advantage of the Bayesian binomial

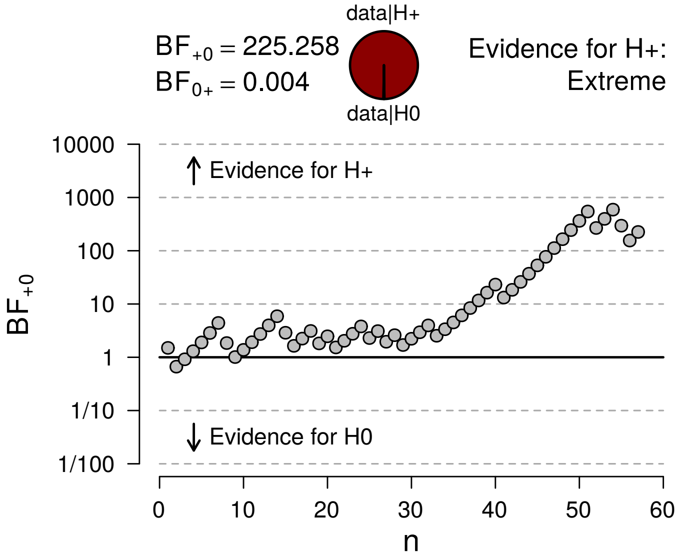


Figure 4.4: Sequential analysis, showing the evolution of the Bayes factor as n, the number of observed participants, increases. After an initial period of inconclusiveness, the Bayes factor strongly favors $\mathcal{H}_1$. Figure from JASP.

test is that both the likelihood function and the parameterization of the prior and posterior distributions are intuitive and straightforward. The tasting experiment allows students to analyze their own data, collected on the fly, making the inferential process more concrete and relevant. Table 4.1 summarizes the concepts that are introduced during the tasting experiment, as well as how these concepts can be practically demonstrated. The experiment is aimed at introducing college level students to these concepts. We have positive experiences using it as a teaching tool in both introductory workshops and undergraduate courses in Bayesian inference. We have created an Open Science Framework repository that contains the original data set, as well as a fully annotated JASP-file that presents additional analyses, such as a t-test on the difference in ratings for the alcoholic and non-alcoholic beer. The repository can be found at https://osf.io/428pb/

	Bayesian Concept	Demonstration
1.	Irrelevance of sampling plan for Bayesian updating	Analyzing the data as they come in
2.	Evidence for $\mathcal{H}_0$ is possible, as it is for $\mathcal{H}_1$	Computing the Bayes factor
3.	Conjugate prior distribution	Using the binomial likelihood to update a beta prior distribution
4.	Savage-Dickey density ratio for computation of Bayes factors	Interpreting posterior plots (e.g., Figure 4.2)
5.	Analysis of sensitivity of results to choice of prior distribution	Changing the parameters of the beta prior distribution and observing the corresponding changes in the posterior distribution and the Bayes factor
6.	Bayesian one-sided testing	Specifying different alternative hypotheses
7.	Principle of parsimony in Bayesian inference	Comparing two-sided results with one-sided results; comparing $\mathcal{H}_0$ with $\mathcal{H}_1$

Table 4.1: Bayesian concepts that students will learn during the tasting experiment and how these concepts can be demonstrated.

Chapter 5

Strong Public Claims May Not Reflect Researchers' Private Convictions

How confident are researchers in their own claims? Augustus De Morgan (De Morgan, 1847/2003) suggested that researchers may initially present their conclusions modestly, but afterwards use them as if they were a “moral certainty”. To prevent this from happening, De Morgan proposed that whenever researchers make a claim, they accompany it with a number that reflects their degree of confidence (Goodman, 2018). Current reporting procedures in academia, however, usually present claims without the authors’ assessment of confidence.

Here, we report the partial results from an anonymous questionnaire on the concept of evidence that we sent to 162 corresponding authors of research articles and letters published in Nature Human Behaviour (NHB). We opted for NHB because of its broad scope and because the majority of its articles include the main claim in the title (e.g., from the first issue, “Pathogen prevalence is associated with cultural changes in gender equality” (Zhou et al., 2016), or “Attention modulates perception of visual space” (Varnum & Grossmann, 2016)), which made it convenient to directly reference the claim in the questionnaire. We selected 129 articles with a claim in the title published between January 2017 and April 2020. The list of selected articles as well as a description of the selection procedure can be found in Appendix A of the online supplement (https://osf.io/zjnpm/). We received 31 complete responses (response rate: 19%). A complete overview of the questionnaire can be found in online Appendices B, C, and D.

As part of the questionnaire, we asked respondents two questions about the claim in the title of their NHB article: “In your opinion, how plausible was the claim before you saw the data?” and “In your opinion, how plausible was the claim after you saw the data?”. Respondents answered by manipulating a sliding bar that ranged from 0 (i.e., “you know the claim is false”) to 100 (i.e., “you know the claim is true”), with an initial value of 50 (i.e., “you believe the claim is equally likely to be true or false”).

This chapter has been submitted for publication as van Doorn, J.B., van den Bergh, D., Dablander, F., Derks, K., van Dongen, N.N.N., Evans, N. J., Gronau, Q. F., Haaf, J.M., Kunisato, Y., Ly, A., Marsman, M., Sarafoglou, A., Stefan, A., & Wagenmakers, E.–J. (in press). Strong Public Claims May Not Reflect Researchers’ Private Convictions. *Significance*. Also available as *PsyArXiv preprint*: https://psyarxiv.com/pc4ad

Figure 1 shows the responses to both questions. The blue dots quantify the assessment of prior plausibility. The highest prior plausibility is 75, and the lowest is 20, indicating that (albeit with the benefit of hindsight) the respondents did not set out to study claims that they believed to be either outlandish or trivial. Compared to the heterogeneity in the topics covered, this range of prior plausibility is relatively narrow.

The lines in Figure 1 connect prior to posterior plausibility for each respondent and their positive slopes indicate that all 31 respondents believed that the data increased the plausibility of the claim from the title of their article. However, with a median of only 80, the posterior plausibility for their claims is surprisingly low. From the difference between prior and posterior odds we can derive the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995), that is, the extent to which the data changed researchers' conviction. The median of this derived Bayes factor is 3, corresponding to the interpretation that the data are 3 times more likely to have occurred under the hypothesis that the claim is true than under the hypothesis that the claim is false. A Bayes factor of 3 equals Jeffreys's threshold value for labeling the evidence "not worth more than a bare mention" (Jeffreys, 1961), further underscoring the authors' modesty and/or seemingly weak convictions of their article's main claim.

The authors' modesty appears excessive. It is not reflected in the declarative title of their NHB articles, and it could not reasonably have been gleaned from the content of the articles themselves. Perhaps authors grossly overestimated the prior plausibility of their claims (due to hindsight bias); or perhaps they were afraid to come across as overconfident; or perhaps they felt that the title claim was overly general. It is also possible that authors were not sufficiently attuned to the response scale, although none of the respondents indicated that the scales were unclear.

Empirical disciplines do not ask authors to express the confidence in their claims. When an author publishes a strong claim in a top-tier journal such as NHB, one may expect this author to be relatively confident. While the current academic landscape does not allow authors to express their uncertainty publicly, our results suggest that they may well be aware of it. Encouraging authors to express this uncertainty openly may lead to more honest and nuanced scientific communication (Kousta, 2020).

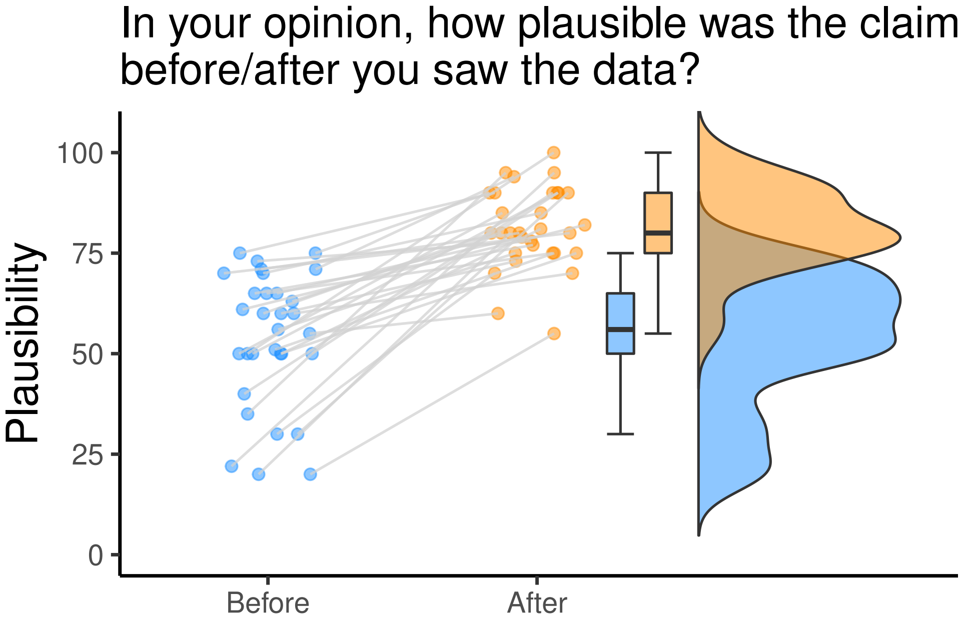


Figure 5.1: All 31 respondents indicated that the data made the claim in the title of their NHB article more likely than it was before. However, the size of the increase is modest. Before seeing the data, the plausibility centers around 50 (median = 56); after seeing the data, the plausibility centers around 75 (median = 80). The gray lines connect the responses for each respondent.

Chapter 6

Bayes Factors for Mixed Models

Abstract

Although Bayesian mixed models are increasingly popular for data analysis in psychology and other fields, there remains considerable ambiguity on the most appropriate Bayes factor hypothesis test to quantify the degree to which the data support the presence or absence of an experimental effect. Specifically, different choices for both the null model and the alternative model are possible, and each choice constitutes a different definition of an effect resulting in a different test outcome. We outline the common approaches and focus on the impact of aggregation, the effect of measurement error, the choice of prior distribution, and the detection of interactions. For concreteness, three example scenarios showcase how seemingly innocuous choices can lead to dramatic differences in statistical evidence. We hope this work will facilitate a more explicit discussion about best practices in Bayes factor hypothesis testing in mixed models.

6.1 Introduction

In a typical response time experiment, multiple participants complete multiple trials in multiple conditions. For example, in a lexical decision task (Meyer & Schvaneveldt, 1971), 30 participants may be instructed to decide as quickly and accurately as possible whether or not 100 individually presented letter strings are words (e.g., FISH) or nonwords (e.g., DRAPA). A possible experimental manipulation may concern the type of motor effector – on half of the trials participants have to press the response buttons with their thumbs, and on the other half they have to use their index fingers. In such two-condition within-participant designs, researchers are generally interested in the effect of the experimental manipulation. As a first step, researchers often address the question of whether or not the manipulation may be said to have had an effect, for instance, whether or not response times (RTs) differ when people respond with their thumbs rather than with their index fingers. Several statistical methods are available to test for such a difference between conditions and the choice among them cannot be based on

This chapter has been submitted for publication as van Doorn, J.B., Aust, F., Haaf, J.M., Stefan, A., & Wagenmakers, E.–J. (2020). Bayes Factors for Mixed Models.

statistical considerations alone—each of these approaches instantiates a different interpretation of the main question of interest.

The oldest and most common analysis approach is to conduct a repeated-measures (RM) analysis of variance (ANOVA), which in the case of two conditions is equivalent to a paired-samples *t*-test. In the scenario above, participants' RTs for individual trials are first averaged within each condition, resulting in two average RTs per participant, one for each condition. We term this averaging process *aggregation*. Following aggregation, participants' average RTs are then subjected to a one-way RM ANOVA.[1]

This method accounts for the correlation between the averaged observations that is caused by some participants generally being faster or slower than others (i.e., the presence of *baseline differences* or random intercepts). This is in contrast to a between participants ANOVA, which is not designed to account for correlated observations. Nonetheless, both types of ANOVA have in common that they are applied to observations averaged across multiple trials. Aggregating individual response times loses information and limits the questions that can be addressed. For example, aggregated RM ANOVA cannot be used to assess whether the experimental manipulation affects all participants alike, or whether the effect of the manipulation differs per participant.

In contrast, *mixed effects models* (also referred to as hierarchical or multilevel models) make use of the full (i.e., unaggregated) data set. These models typically account for the nested data structure by modelling baseline differences in general response speed across participants (as in RM ANOVA) as well as differences in the magnitude of the condition effect across participants (i.e., random slopes). By modelling individual RTs, mixed effects models enable researchers to ask more specific questions. As in RM ANOVA, mixed effects models estimate the average effect of condition (i.e., the fixed effect), but additionally they can be used to examine the extent to which the effect of condition differs between participants (i.e., the random effect).

The example given here can be generalized in two ways. First, while condition is a categorical variable, the same framework can be applied to continuous predictor variables. Second, the random effects *grouping factor* (in the example above, "participant" is the grouping factor) can be any categorical variable in the design for which there are multiple observations. For instance, instead of modeling differences between individual participants, we could model a difference in the effect of the manipulation for left-handed people, compared to right-handed people. Furthermore, in the case of multiple grouping factors, the random effects can either be nested or crossed. In the case of nested random effects, not all levels of one grouping factor are measured for the other grouping factor. For example, if both "participant" and "handedness" are used as grouping factors, the structure is nested, because each participant will be either left-handed or right-handed. In the case of crossed random effects, all levels of one grouping factor are measured for the other grouping factor. For example, if both "participant" and "item" are

[1] We assume that interest centers on RT for correct responses to word stimuli. For simplicity, we also assume that the untransformed trial-level RTs are normally distributed, which is usually not the case.

used as grouping factors, the structure is crossed, because for each participant, there are observations for each item (for more examples, see Baayen et al., 2008; Quené & Van den Bergh, 2008; Singmann & Kellen, 2019).

Although alluded to by Fisher (1935) and Yates (1935), the first explicit definition of random intercepts was given by Jackson (1939), who proposed to account for individual differences in intelligence in order to more accurately assess the reliability of mental tests. Since their introduction, mixed effects models have seen an increase in statistical development (e.g., Scheffe 1956; Kempthorne 1975; Efron & Morris 1977; Nelder 1977; Lindstrom & Bates 1990), and arguably rank among the most important statistical ideas of the last 50 years (Gelman & Vehtari, 2020). The application of mixed effects models has been particularly stimulated by software implementations (e.g., `lme4`, Bates, Mächler, et al. 2015; `nlme`, Pinheiro & Bates 2000; Pinheiro et al. 2020; and `afex`, Singmann et al. 2020) and tutorial papers (e.g., Baayen et al., 2008; Judd et al., 2012, 2017; Singmann & Kellen, 2019).

Here we focus on Bayesian inference for mixed effects models, and specifically on Bayes factor hypothesis tests (e.g., Rouder et al., 2012; Clyde et al., 2011).[2] Despite the availability of Bayesian tutorials (Shiffrin et al., 2008; Rouder et al., 2013; Sorensen et al., 2016) and software alternatives (e.g., Morey & Rouder, 2018; Carpenter et al., 2017; Goodrich et al., 2020; Bürkner, 2017; Thalmann & Niklaus, 2018; JASP Team, 2020), there remains a lack of clarity and consensus about how to best conduct Bayesian model comparison when considering mixed effects.

Examining the effect of a manipulation requires the specification of both a null model, which assumes no effect of the manipulation, and an alternative model. In the frequentist framework, a well-cited recommendation for the specification of the alternative mixed effects model of the full data is to specify a “maximal” model (i.e., the model that includes all fixed and random effects justified by the study design). In particular, failure to include random slopes can inflate Type 1 and Type 2 error probabilities (Barr et al., 2013; Berkhof & Kampen, 2004; Schielzeth & Forstmeier, 2008; Heisig & Schaeffer, 2019, but see Bates, Kliegl, et al. 2015; Matuschek et al. 2017). Despite the fact that there are multiple suitable null models that the maximal model can be compared to, the appropriate specification of the null model is much less discussed. This is problematic because the choice of the null model (just like the alternative model) defines the question we ask about the condition effect.

Several decisions need to be made when testing for the effect of a manipulation in an experimental within-participant designs: Which model comparisons are both suitable and sensible, whether or not to aggregate, how to quantify effects, and how to set prior distributions. The aim of the current paper is to list the available options and demonstrate their impact on inference. We hope our exposition provides a common starting ground for a discussion among experts in the field of Bayes factor model comparison. We further hope that this discussion will foster the development of a much needed set of guiding principles for the

[2] We use the terms “hypothesis test” and “model comparison” interchangeably.

applied researcher who ventures into the realm of Bayesian mixed models.

The outline of this paper is as follows. We start by defining the possible models that can be compared when random effects are considered. Then, we present a simple synthetic data set to illustrate the differences in model comparisons, as well as the effect of aggregating the data. The second example demonstrates how the different mixed model comparisons behave when analyzing data sets with either few accurate measurements or many noisy measurements. As a third example, we present a more real-life data set that underscores these modeling questions, and highlights the added complexity of having multiple independent variables of interest.

6.2 The Candidate Models

In this section we define the candidate models for a one-factorial design. Suppose I participants each observe M trials in each of J conditions. For this research scenario, there are the following six candidate models, each with different theoretical underpinnings:

Model 1. Intercept (μ) only: no fixed effect of condition and no random effects for participants. With subscript i for the i^{th} participant, j for the j^{th} condition, and m for the m^{th} trial, the model for the observed values Y_{ijm} can be written as a function of the grand mean μ and the error variance σ_ϵ^2. We give this definition below, and then expand it for each subsequent model:

$$Y_{ijm} \sim \mathrm{N}(\mu, \sigma_\epsilon^2). \tag{6.1}$$

Model 2. Fixed effect ν of condition, but no random effects for participants. The term x_j is a design element that encodes condition (i.e., $x_1 = -1/2$, $x_2 = 1/2$ if $J = 2$), which ensures the sums-to-zero constraint for the fixed effects (Rouder et al., 2012).[3] The resulting model can be written as follows:

$$Y_{ijm} \sim \mathrm{N}(\mu + x_j\nu, \sigma_\epsilon^2). \tag{6.2}$$

Model 3. No fixed effect of condition, but random intercepts α_i specific to the ith participant. In contrast to Models 1 and 2, this model includes baseline differences. The random intercepts are distributed normally around the grand mean μ, with standard deviation σ_α. Random intercepts can also be understood as a main effect of participant. When σ_α is 0, Model 6.3 reduces to Model 6.1. In one-way RM ANOVA, this model is typically used as the null model. The model can be written as follows:

$$\begin{aligned} Y_{ijm} &\sim \mathrm{N}(\alpha_i, \sigma_\epsilon^2), \\ \alpha_i &\sim \mathrm{N}(\mu, \sigma_\alpha^2). \end{aligned} \tag{6.3}$$

[3] This setup is known as effect coding, and implies that the μ parameter is the grand mean.

Model 4. Fixed effect ν of condition and random intercepts α_i for participants. In one-way RM ANOVA, this model is used as the alternative model. The model can be written as follows:

$$\begin{aligned} Y_{ijm} &\sim \mathrm{N}(\alpha_i + x_j\nu, \sigma_\epsilon^2), \\ \alpha_i &\sim \mathrm{N}(\mu, \sigma_\alpha^2). \end{aligned} \tag{6.4}$$

Model 5. No fixed effect, but random intercepts α_i and slopes θ_i specific to the ith participant. The random slopes are distributed normally around 0, with standard deviation σ_θ. In general, random slopes can also be understood as an interaction effect between condition and participant (Nelder, 1977). When σ_θ is 0, Model 6.5 reduces to to Model 6.3. In essence, this model postulates that there is an effect of condition in each participant, but that it varies across participants in a perfectly balanced way, such that the average effect is 0 across participants. The model can be written as follows:

$$\begin{aligned} Y_{ijm} &\sim \mathrm{N}(\alpha_i + x_j\theta_i, \sigma_\epsilon^2) \\ \alpha_i &\sim \mathrm{N}(\mu, \sigma_\alpha^2) \\ \theta_i &\sim \mathrm{N}(0, \sigma_\theta^2) \end{aligned} \tag{6.5}$$

Model 6. The full model, with fixed effect ν of condition, random intercepts α_i, and random slopes θ_i for participants. All previous models are restrictions of this model. For mixed models in the frequentist framework, this is the often-recommended alternative model (Barr, 2013). The model can be written as follows:

$$\begin{aligned} Y_{ijm} &\sim \mathrm{N}(\alpha_i + x_j\theta_i, \sigma_\epsilon^2) \\ \alpha_i &\sim \mathrm{N}(\mu, \sigma_\alpha^2) \\ \theta_i &\sim \mathrm{N}(\nu, \sigma_\theta^2) \end{aligned} \tag{6.6}$$

We do not entertain all possible combinations of parameters (e.g., models with random slopes but no random intercepts, or models without a grand mean), because we consider them both theoretically and statistically inappropriate in the current mixed modeling setting.

For all models, σ_ϵ^2 denotes the error variance, which is the variance in the data left unexplained by the model. The explained variance of a mixed model is the sum of the variance induced by the fixed effect, the variance of the random intercepts σ_α^2, and the variance of the random slopes σ_θ^2 (Rights & Sterba, 2019). Together the explained variance and σ_ϵ^2 make up the total variance of the data y. Random effects are a source of systematic variation that, if unaccounted for in the model, may be incorrectly attributed to the explained variance of a fixed effect, or the error variance, leading to conclusions about the fixed effect that are either overly permissive, or overly conservative, respectively (Barr et al., 2013).

6.2.1 The Model Comparisons

With the series of six models defined, we can use model comparisons to assess whether or not there is an effect of condition. Between the six models under consideration we can make $\frac{n(n-1)}{2} = 15$ model comparisons that can be applied to either the full data or the aggregated data. The models differ from each other with respect to the three parameters of interest: ν, σ_α, and σ_θ. The appropriate model comparison depends on the research question at hand, since each comparison answers a different question. Specifically, the combined choice of the null and alternative model constitute different definitions of what it means for a manipulation to have an effect: Model 6.4 posits that the fixed effect manifests in every participant, whereas Model 6.6 posits that the fixed effect is the average of participant-specific effects that vary in magnitude. Below, we consider three model comparisons that we consider to be primarily relevant to the current scenario.

We start by outlining the popular RM ANOVA procedure, which compares Model 6.3 to Model 6.4. This procedure uses one observation per participant and per level of each factor (i.e., $M = 1$). In cases where $M > 1$, the observations are typically aggregated first, even though aggregation is not strictly required.[4] The aggregation discards information about the distribution of each participants' observations within each condition. As a consequence, it becomes impossible to distinguish between systematic random slope variance and random error variance (i.e., aggregation confounds the random slopes variance with the residual variance). However, a benefit of aggregation is that it greatly reduces the impact of random slopes in the inference for a fixed effect and therefore eliminates the inflation of Type 1 and Type 2 error rates that ignoring random slopes typically entails (see Examples 1 & 2 for a demonstration). Comparing Model 6.3 to Model 6.4 on the full data, on the other hand, does suffer from this inflation, and we therefore do not consider it appropriate.

We now outline the comparisons of models that contain random slopes. In the frequentist framework, it is often recommended to use the maximal mixed model justified by the design (Barr, 2013). The presence of the fixed effect ν is typically tested by means of a t- or F-test. This procedure implicitly compares the full model (Model 6.6) to the full model without the fixed effect (Model 6.5). Since random slopes are in fact an interaction effect between the fixed effect of condition and the random effect of participant, specifying a model that includes random slopes without the corresponding fixed effect (i.e., Model 6.5) can be seen as conceptually problematic. Specifically, Rouder et al. (2016) argues that a model containing an interaction effect without the main effect is only plausible when the exact levels of each factor are picked such that the true main effects perfectly cancel, which in most practical applications seems implausible. If we accept this argument while still accounting for random slopes, the implied model comparison is between the full model (Model 6.6) and the model with only the

[4] Barr (2013) notes this as one of the three common misconceptions about conventional RM ANOVA. While Barr does not advise to conduct the typical RM ANOVA (i.e., without considering the random slopes) using the non-aggregated data, it is technically possible to do so.

random intercepts (Model 6.3). However, this model comparison comes with its own set of challenges: The increase in model validity coincides with a loss of diagnostic specificity: when Model 6.6 outperforms Model 6.3, we can only conclude that the data offer support for the presence of a fixed effect, random effect, or both a fixed and random effect.

Note that, for aggregated data, Models 6.5 and 6.6 are not identified.[5] We therefore only consider model comparisons involving Models 6.5 and 6.6 when applied to the full data.

Thus, based on different considerations, we identify three possible model comparisons, where the last two comparisons are named after the researchers (i.e., Klaus Oberauer and Jeff Rouder) who advocated these comparisons in an informal email discussion:

1. The RM ANOVA comparison: Model 6.3 vs Model 6.4 using the aggregated data
2. The Oberauer comparison: Model 6.5 vs Model 6.6 using the full data
3. The Rouder comparison: Model 6.3 vs Model 6.6 using the full data

6.3 Examples

Although all three comparisons outlined in the previous section can be viable options in an applied setting, they may lead to dramatically different conclusions. In order to illustrate the different behaviors of the three comparisons, we now discuss three data examples. We follow each example with several concrete questions that we hope will serve as useful starting points for discussion.

All Bayes factors presented below are computed using the `BayesFactor` package (Morey & Rouder, 2018), using the default settings for the multivariate Cauchy prior distributions (scale set to 0.5 and 1 for fixed effects and random effects, respectively). Each example also includes a reference to the analysis code in the online supplementary material.

6.3.1 Example 1: The Effect of Aggregation

We start with a relatively simple scenario, where $I = 20$ participants complete $M = 15$ trials in each of $J = 2$ conditions for a total of 600 observations.[6] The purpose of this example is to illustrate the effect of random slopes on the different model comparisons, and how each comparison reacts to the process of aggregation. Figure 6.1 shows both the full data and the aggregated data, where

[5] Technically, in the Bayesian framework random slopes can be included even for the aggregated data. In this case the estimates will be informed entirely by the prior distribution. Therefore, in most practical applications this approach is not useful.

[6] These data were generated using a Shiny app we developed to better understand these model comparisons under different population parameters. The app can be found at https://bayesianmixedmodels.shinyapps.io/mixedModelsMarkdown/ and the R-script for these specific data at https://osf.io/tjgc8/.

each color corresponds to a different participant. The data were simulated with a medium fixed effect ($\nu = 0.5$), random intercepts ($\sigma_\alpha^2 = 0.5$), and random slopes ($\sigma_\theta^2 = 1$). The difference between the top-left and top-right panel clearly underscores the process of aggregating, where a lot of information is discarded. The random slopes are evidenced by the different orientations of the lines in the plots in the bottom row: some participants exhibit an increase from condition 1 to condition 2, while for other participants this effect is reversed. To further demonstrate the effect of aggregation, we present the results for all three comparisons, for both the full and aggregated data.

The different model comparisons yield wildly different Bayes factors. For comparison purposes we report the natural logarithm of the Bayes factor throughout this manuscript. When $\log(\text{BF}_{A,B}) > 0$, Model A is preferred; when $\log(\text{BF}_{A,B}) < 0$, Model B is preferred. First, consider the results for the full data set:

1. The RM ANOVA comparison: $\log(\text{BF}_{4,3}) = 10.81$
2. The Oberauer comparison: $\log(\text{BF}_{6,5}) = 0.04$
3. The Rouder comparison: $\log(\text{BF}_{6,3}) = 65.5$

The RM ANOVA comparison on the full data highlights why it is important to include random slopes whenever possible. The true difference between the condition means is modest, and so is the sample size – yet this model comparison yields overwhelming evidence in favor of a fixed effect of condition, a result caused by the presence and pronounced influence of the random slopes. This behavior aligns with the inflation of Type 1 error probabilities in the frequentist framework as demonstrated by Barr et al. (2013), who therefore advised against performing the RM ANOVA comparison on the full data. Since the Oberauer comparison controls for random slopes by including the random slopes term in both models, it does not suffer from the overconfidence displayed in the RM ANOVA comparison. The Rouder comparison yields extreme evidence in favor of Model 6.6, but based on this comparison alone it is impossible to conclude whether this evidence is due to the random slopes, the fixed effect, or both.

Now, consider the results for the aggregated data:

1. The RM ANOVA comparison: $\log \text{BF}_{4,3} = 0.2$
2. The Oberauer comparison: $\log \text{BF}_{6,5} = 0.16$
3. The Rouder comparison: $\log \text{BF}_{6,3} = -0.21$

For the two comparisons where only one or none of the models include a random slope (i.e., the Rouder comparison and the RM ANOVA comparison, respectively), aggregation greatly impacts the Bayes factor. For both comparisons, the previously overwhelming Bayes factor plummets to around 0, leaving it undecided about which model best predicted the data. This demonstrates that aggregation eliminates the presence of random slopes: for the Rouder comparison, there is no longer any evidence for the alternative model, and for the RM ANOVA

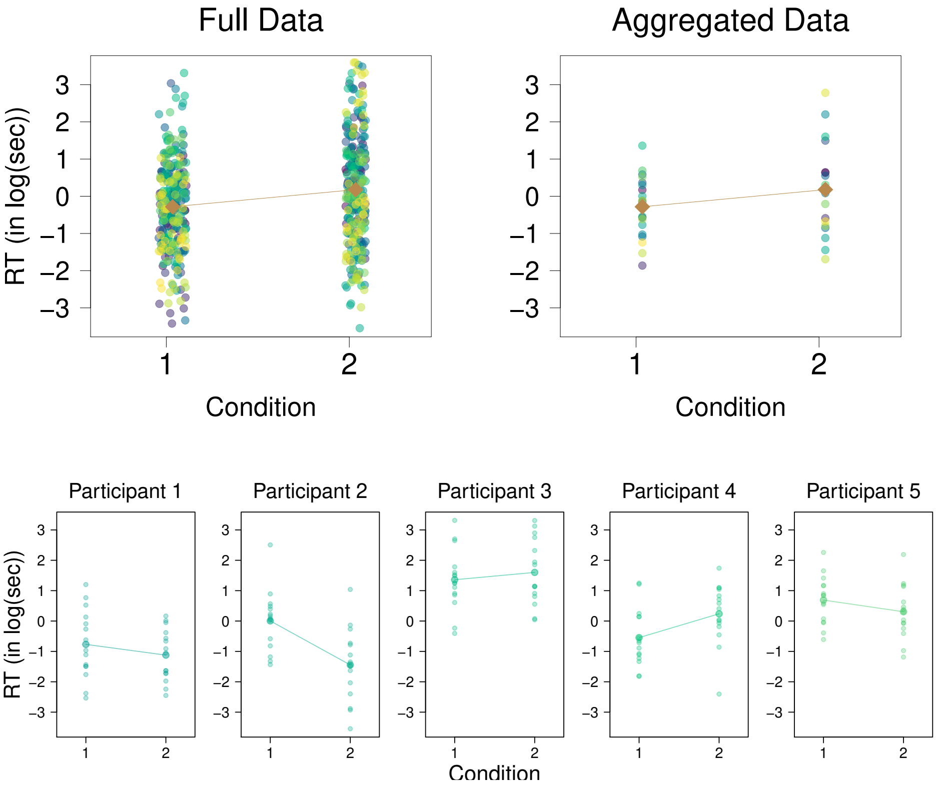


Figure 6.1: Synthetic data for Example 1: the effect of aggregation. The top left panel presents the full data set, and the top right panel the aggregated data, where the average value is taken per participant, per condition. The bottom row presents the full data for five example participants, including their condition means. Some participants display an increase as a result of the manipulation, whereas other participants display a decrease. Note that the overall and participant-specific condition means are exactly the same for both versions of the data. The different point colors in the top row correspond to different participants.

comparison there is no longer the inflation of the evidence in favor of a fixed effect.

In contrast, the Oberauer comparison appears relatively stable and is barely in favor of Model 6.6 in both cases, since the two rival models both include the random slopes. However, we should stress that conducting the Oberauer and Rouder comparisons on aggregated data is unorthodox (i.e., the random slopes estimates are entirely informed by the prior distribution), and that we present these two Bayes factors merely as an illustration.

Taken together, these results suggest that there are two valid methods to test for the presence of a fixed effect, and only the fixed effect, of condition in the presence of random slopes: either performing the Oberauer comparison on the full data, or performing the RM ANOVA comparison on the aggregated data. In order to avoid the demonstrated inflation of the fixed effect when performing the RM ANOVA comparison on the full data, we therefore only consider the this comparison for the aggregated data and not the full data in the remainder of this manuscript.

The considerations above motivate the following questions:

1. What are the relevant model comparisons for a one-factorial design?
   (a) When is aggregation an appropriate procedure?
2. Should more models be considered than the ones described here?
3. If strictly interested in the fixed effect only, when should the RM ANOVA comparison be used instead of the Oberauer comparison?

6.3.2 Example 2: The Effect of Measurement Error

For the second example, we again consider RTs of $I = 20$ participants in $J = 2$ conditions. However, the measurements are done with an instrument that can either measure quickly but inaccurately, or measure accurately but slowly. Thus, there is a trade-off between the measurement error and the number of trials that can be measured in the experiment. If this trade-off is perfectly balanced (i.e., the observed condition means, observed participant means within each condition, and the within-participant standard errors of the condition means are identical) does it matter which setting we choose? In other words, can a noisy measurement instrument be compensated for by collecting many data points per participant? The purpose of this example is to demonstrate how the different model comparisons behave as both the measurement error and number of trials decrease.

In order to implement the trade-off between number of trials and measurement error, we can start with the data set that has 100 trials per participant, per condition. Then, the average RT can be taken of every 10 trials a participant completes. This results in 10 scores per participant, per condition. For both of these data sets, the participant means for each condition and the within-participant standard errors for the fixed effect of condition in a hierarchical model are identical. The difference between these data sets lies in the trial level variance (i.e., the residual variance). Multiple data sets can be created this way by using different numbers of trials to average across. Doing so illustrates how the different model comparisons develop as the number of trials decreases, but the accuracy of those measurements increases. When the number of averages equals 100, the full data is used; when the number of averages equals 1, we obtain the fully aggregated data set. We also create intermediate data sets by taking 50, 10, 5, and 2 averages per participant, per condition. Since the RM ANOVA comparison is performed on the aggregated data and would remain identical, we consider only the Oberauer and Rouder comparisons for these different versions of the data.

Figure 6.2 shows how each Bayes factor changes as the number of trials decreases and the accuracy of each individual trial increases. The three panels correspond to three different models that generated the data. In the first panel, data were simulated under the model without a fixed effect or random slopes (i.e., Model 6.3). In the second panel, data were simulated for a fixed effect only (i.e., Model 6.4). In the third panel, data were simulated under a fixed effect and random slopes (i.e., Model 6.6). The code for the data generation and analysis is available at https://osf.io/jsgm3/.

The effect of decreasing the number of trials in the data is the most pronounced in the Rouder comparison, where the log Bayes factor gets less decisive (i.e, closer to 0) as the data set goes towards full aggregation in all three settings. As was illustrated in Example 1, the process of aggregation confounds the random slope variance with residual variance. This results in a less decisive Bayes factor for the Rouder comparison, as only one of the two models being compared includes the random slopes term. In the top panel, the data were generated under the null model of the Rouder comparison, and in the bottom panel the data were generated under the alternative model of that comparison. For these settings, it is not surprising that those models receive the most support in their favor, although the magnitudes of the Bayes factors seem too extreme in light of the relatively low sample sizes.

Surprisingly, the middle panel still depicts overwhelming evidence in favor of the alternative model (i.e., model 6), even though the data generating model did not include random effects. The Oberauer comparison in the middle panel also depicts evidence in favor of model 6, but to a much lesser extent than the Rouder comparison. Since both comparisons have the same alternative model, this difference in behavior is due to the null model. It seems that the null model in the Oberauer comparison (i.e., model 5) is able to model the data better than the null model in the Rouder comparison (i.e., model 3). This suggests that a random slopes term can, to some degree, account for a fixed effect in the data. Moreover, by definition, adding a random effect inherently increases a models robustness to the added variance induced by outliers.

The Oberauer comparison is relatively stable in the top and bottom panel, which confirms the balance between the number of trials and their accuracy. Since this comparison is between two models that both contain the random slopes factor, these Bayes factors do not reflect the effect of averaging on the random slope variance. However, the Oberauer comparison is not entirely stable across the different levels of averaging.

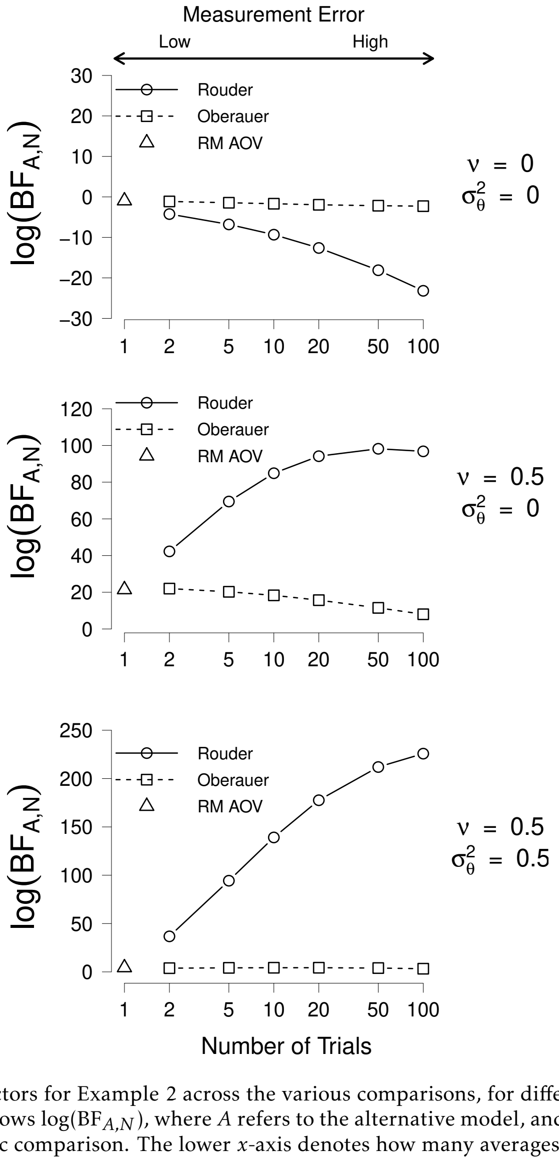


Figure 6.2: Bayes factors for Example 2 across the various comparisons, for different levels of aggregation. The y-axis shows $\log(\text{BF}_{A,N})$, where A refers to the alternative model, and N refers to the null model of that specific comparison. The lower x-axis denotes how many averages are taken; 100 indicates the use of the full data, and 1 indicates the use of the aggregated data. The upper x-axis denote the measurement error, which decreases as the number of trials decreases. The Rouder comparison is highly influenced by the presence or absence of random slopes, although this sensitivity dramatically decreases as the the number of trials decreases. The Oberauer comparison remains relatively stable around values of −1.5 and 5 in the top and bottom panel, respectively.

We suspect that the instability of the Oberauer comparison is largely due to a component of Bayesian mixed modeling that we have not addressed so far: the prior distribution. Until now, we have used the default settings for the scale settings of the multivariate Cauchy prior distributions, which are 0.5 and 1 for the fixed effects and random effects, respectively. The widths of these distributions reflect which standardized effect sizes are to be expected under each model. The standardization of the effect sizes is influenced by the measurement error, since observing the same mean difference between conditions, but with a smaller measurement error, results in a larger effect size. Thus, the prior distribution ought to reflect information about the expected measurement error: when this error is small, we can expect larger effect sizes and the prior distribution should be wider, and vice versa.

We suspect that using the same prior distribution for each level of aggregation is in part what leads to the extreme levels of evidence obtained for the full data sets in the Rouder comparisons, which seems overly sensitive to the presence of absence of random slopes in the data. In the case where there is only a fixed effect and no random slopes (middle panel of Figure 6.2), the Rouder comparison yields far more decisive Bayes factors than the Oberauer comparison, which does not seem desirable. We therefore wish to underscore the importance of having a sensible prior specification (i.e., accounting for the trial level variance) when random slopes are considered in only one of the two models under consideration.

Finally we focus on the difference between the Rouder and the Oberauer comparisons. In the former comparison, the null model (Model 6.3) postulates that none of the participants is affected by condition, whereas the alternative model (Model 6.6) postulates that participants are affected differently. For the Oberauer comparison, both the null model (Model 6.5) and the alternative model (Model 6.6) postulate that participants are affected differently by condition, but only the latter model postulates an overall effect. When there are more observations per participant, the error variance and the random slope variance can be disentangled more easily. Since the Rouder comparison focuses more on individual differences than the Oberauer comparison, collecting more data points will lead to more decisive Bayes factors in the Rouder comparison than in the Oberauer comparison.

The above considerations motivate the following questions:

1. How should the prior distributions for the fixed and random effect be constructed?
    (a) What is a meaningful standardized effect size in this scenario?
    (b) How can we construct an effect size that is meaningfully standardized? In other words, what variance should we standardize by?
2. Since there is overlap of the predictive space of Model 4 (fixed effect, but no random effect) and Model 5 (random effect, but no fixed effect), there is a certain degree of model mimicry (see also Figure 6.2); can we meaningfully disentangle a fixed effect and random effect, both statistically and theoretically?

6.3.3 Example 3: A Random Interaction Effect

Up to now, our discussion on mixed models has only dealt with the relatively simple case of a single independent variable of interest (e.g., condition). The purpose of the present example is to highlight how mixed model comparisons are affected by the presence of multiple independent variables of interest, and to explore which models to consider when testing for the presence of an interaction effect. Due to the addition of a second independent variable of interest, the possibility emerges to test for an interaction effect between the two variables that, just like a main effect, can have a fixed and a random component. Just as for the main effects, each cell of the interaction (i.e., each combination of levels from each factor) requires multiple measurements within each participant for the random interaction effect to be identifiable.

To demonstrate the decisions that arise when testing for an interaction effect, we consider a real world example by Lukács et al. (2020).[7] In this scenario, the hold-duration of a response button was measured in $I = 116$ participants, who completed an item recognition task where they used either thumb or index finger (factor A, with two levels) to respond to either a probed or irrelevant item (factor B, with two levels). For the RM ANOVA comparison, we can consider the aggregated case with one observation per participant, per cell of the design (i.e., per level of A, per level of B). For instance, the aggregated data contains one observation for the hold-duration of participant 4, where they responded with their thumb to an irrelevant item. Figure 6.3 presents these aggregated data.

For aggregated data, the analysis of choice is typically a RM ANOVA, where only the fixed effects of A, B, and A×B are considered. However, despite of the aggregation, it is possible to fit random slopes for A and B, because there are 2 observations for each level of A, and 2 observations for each level of B, for each participant. On the other hand, the aggregation prevents the calculation of random slopes for the interaction effect as there is only one observation for each combined cell of A×B.[8] Considering the full data instead of the aggregated data enables the fitting of random slopes for main and interaction effects. Figure 6.4 presents the full data that contains multiple observations per cell of the research design.

In this example, we are interested in whether there is an interaction effect A×B, as the original authors postulated that participants might keep the response button pressed for a longer period of time when responding to an irrelevant probe (factor B), and that this difference in hold-duration might differ per response mechanism (factor A). Because we previously defined the models in a scenario with only a single variable of interest, we will alter the models under considera-

[7] In fact, a forum post commenting on diverging results in the frequentist and the Bayesian RM ANOVA provided additional motivation for the current project. In the post, the p-value yielded evidence in favor of the interaction effect, while the Bayesian RM ANOVA yielded evidence against the interaction effect. Upon investigating the issue, it became clear that inference for an interaction effect, in the context of mixed effects, is not a straightforward endeavor.

[8] In general terms, aggregation of the data limits random slopes to only be fitted to $(K-1)$-order effects, where K is the number of categorical independent variables measured within each participant. In the case above, $K = 2$, so we can still fit random slopes for first order effects.

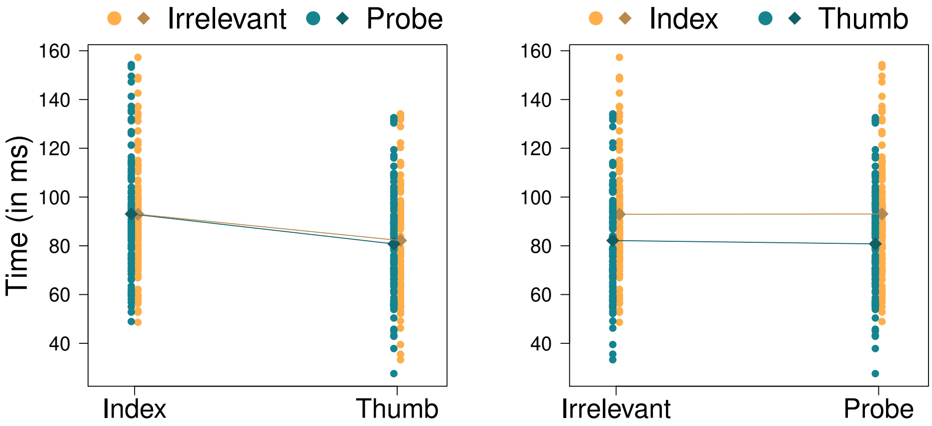


Figure 6.3: Aggregated data for Example 3, where we consider only the average observation per participant, per cell of the design. Each point in the plot represents one aggregated hold-duration. The left panel has factor A on the x-axis and factor B indicated by the colors of the points. The right panel has factor B on the x-axis and factor A indicated by the colors of the points. The lines connect the condition means in order to illustrate whether or not there is an interaction effect. If the two lines are not parallel, this is an indication of an interaction effect. There appears to be a main effect of factor A (i.e., responses made with the thumb are faster than those made with the index finger).

tion. We list the models under consideration in Table 6.1, and below we describe the process of constructing these models.

We start with the commonalities. Previously, this was only μ and σ_ϵ, but now this includes all parameters that are essential: the main effects of A and B (due to marginality; see also Wagenmakers, Love, et al., 2018b, and references therein), and the random intercepts for each participant (due to the repeated measures design). This defines a new version of Model 6.1. Next, we add the fixed interaction effect of the two factors, A×B, and create a new version of Model 6.2. However, these two models can also include random slopes for the main effect of A and B, since these are now identifiable. Thus, we can define Model 3 and 4, which are similar to the updated Models 1 and 2, but with added random slopes for A and B. Finally, we can add the random slopes for the interaction effect to these newly defined models 3 and 4, and create the updated versions of models 6.5 and 6.6, respectively.

With the updated models, we can consider the different model comparisons again. The RM ANOVA comparison, which is based on the aggregated data, can be either between Model 1 and Model 2, or between Model 3 and Model 4, based on whether or not the random effects for A and B are included. We will refer to the former comparison as the “minimal” RM ANOVA comparison, as it includes no random slopes at all. As before, both versions of the RM ANOVA take the approach of minimizing the random effects through aggregation, in order to focus on the fixed effect at hand. The Oberauer comparison (Model 5 vs Model 6), on the other hand, accounts for the random effect by including it in both models that are compared, such that the only difference between the models is the fixed effect of interest. The Rouder comparison (Model 3 vs Model 6) makes a differ-

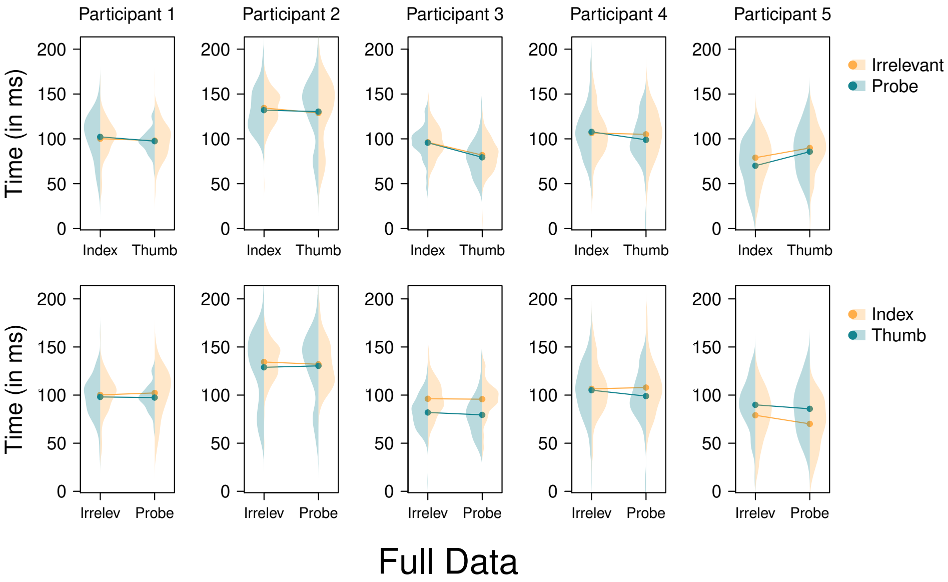


Figure 6.4: Full data for Example 3, where we consider all observations per participant, per cell of the design. The distributions of hold-duration for five example participants for each combination of conditions A and B are shown in two rows. The top row shows factor A on the x-axis, and indicates factor B with the different colors. The bottom row shows factor B on the x-axis, and indicates factor A with the different colors. The points indicate the participant means for level of A and B. The lines are drawn between the points to indicate the change in hold-duration. If these two lines are not parallel, this is an indication of an interaction effect. A random interaction effect then means that different participants exhibit varying degrees of the two lines not being parallel.

ent statement. Analogous to the earlier examples, the difference between the two models under consideration is the combination of both the fixed and random effect. It therefore quantifies evidence for the presence or absence of a *general* effect of condition.

The differences between these four comparisons are again reflected in the diverging Bayes factors:[9]

1. The minimal RM ANOVA comparison: $\log \text{BF}_{2,1} = -1.75$

2. The RM ANOVA: $\log \text{BF}_{4,3} = 2.26$

3. The Oberauer comparison: $\log \text{BF}_{6,5} = -1.59$

4. The Rouder comparison: $\log \text{BF}_{6,3} = -34.35$

The RM ANOVA comparison is the only case where there is evidence in favor of an interaction effect. Interestingly, there is a discrepancy between the two RM ANOVA comparisons, which means that including the random effects for A and B has consequences for the interaction effect. A possible explanation for this is

[9] The R-script with the analysis code can be found at `https://osf.io/cw5jd/`.

Model	Specification			
(1)	A + B + id			
(2)	A + B + id	+ A×B		
(3)	A + B + id		+ B×id + A×id	
(4)	A + B + id	+ A×B	+ B×id + A×id	
(5)	A + B + id		+ B×id + A×id	+ A×B×id
(6)	A + B + id	+ A×B	+ B×id + A×id	+ A×B×id

Table 6.1: Model definitions for a 2x2 design when analyzing an interaction effect. All models contain the fixed effect of A and B, and the random intercept for each participant. In the model specification, "id" refers to a random effect: "+id" refers to the random intercept, while "×id" refers to the random slope (e.g., B×id denotes random slopes for the main effect of B). Models 2, 4, and 6 contain the fixed interaction effect of A and B. Models 3-6 contain random slopes for the main effects of A and B. Models 5 and 6 contain random slopes for the interaction effect of A and B.

the presence of a strong random and fixed effect for A. This result stands in contrast to the frequentist results in Barr (2013), who demonstrated that excluding the non-critical random slopes yields similar results to the approach that does include the non-critical slopes.

The Oberauer comparison agrees with the minimal RM ANOVA comparison and provides moderate evidence against the presence of a fixed interaction effect. The Rouder comparison also agrees but yields a much stronger Bayes factor, which implies that there is also no evidence for a random interaction effect. Table 6.2 shows all possible Bayes factor comparisons between the 6 models outlined here, for both the full and aggregated data. From these comparisons, it is clear that there is evidence in favor of random effects of A and/or B, because of the overwhelming Bayes factors comparing Models 3, 4, 5, and 6 (i.e., the models with the random effects of A and B) to Models 1 and 2 (i.e., the models without the random effects of A and B). For instance, while Model 1 is marginally better than Model 2 for the full data ($\log(\text{BF}_{1,2}) = 1.03$), Model 1 is heavily outperformed by Model 3 ($\log(\text{BF}_{1,3}) = -4286.09$). Based on the table, Model 4 (i.e., the model with the random and fixed main effects) performed the best for both versions of the data: all the Bayes factors in row 4 are positive, indicating that there is at least moderate support for Model 4 compared to the model in the column, for both aggregated and full data.

The considerations above motivate the following questions:

1. What are the relevant model comparisons for a main effect in a two-factorial design?
2. What are the relevant model comparisons for an interaction effect in a two-factorial design?
    (a) Should the random main effects be included in all comparisons?
3. How can we explain the difference between the two versions of the RM ANOVA comparison?

4. For this example, is it theoretically meaningful to analyze random main effects when the data is aggregated?

Model	(1)	(2)	(3)	(4)	(5)	(6)
(1)		1.03	-4286.09	-4286.47	-4253.33	-4251.74
(2)	-1.75		-4287.12	-4287.50	-4254.36	-4252.77
(3)	248.67	250.42		-0.38	32.76	34.35
(4)	250.93	252.68	2.26		33.14	34.73
(5)	247.84	249.59	-0.83	-3.09		1.59
(6)	250.67	252.42	2.00	-0.26	2.83	

Table 6.2: Bayes factors for all pairs of models defined in Table 6.1. The cell entries are $\log(\text{BF}_{R,C})$, where R refers to the model in the row, and C refers to the model in the column. Bayes factors above the diagonal are for the full data, and under the diagonal are for the aggregated data.

6.4 Concluding Comments

This manuscript illustrated the three main choices faced by researchers who apply mixed models: when and why to aggregate, which model comparisons to use when testing hypotheses about the presence or absence of an effect, and whether or not to collect more (albeit noisier) observations per participant. Testing for a fixed effect is not straightforward in the presence of random effects, and we presented three approaches to do so. First, the data can be aggregated, which minimizes the impact of the random effects in the inference for a fixed effect. Second, two models can be compared that both include the random effects, which controls for the random effects. Third, the fixed and random effect can be considered together, instead of trying to dissect the general effect into its constituent elements. Each of these three approaches have their own implications for the three main choices, and –especially in the case where more than one variable is considered– the consequences of these different choices can be profound.

Our aim is for this manuscript to initiate a discussion on best practices in Bayes factor model comparison in mixed models. Table 6.3 outlines the specific questions and their relevant examples. Mixed model comparisons are surprisingly intricate, and a systematic discussion of the most pressing topics is long overdue. We hope that this discussion will result in broad consensus on best practices, even if this consensus is that those who apply mixed models should be aware what models are being compared and, consequently, what questions are being answered.

Question	Related Example
What are the appropriate model comparisons for a one-factorial design?	1, 2
What are the appropriate model comparisons for a two-factorial design?	3
What is the effect of aggregation?	1, 2
How should prior distributions be specified in the context of random effects?	2
Is it desirable to have different inference for many noisy observations, compared to few accurate observations?	2
How to cope with a growing model space, as the design becomes more complex?	3

Table 6.3: Summary of the different modeling questions faced when conducting a Bayes factor mixed model comparison. The right column indicates which of the presented examples are relevant to each question.

Part II

For Ranks

Chapter 7

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 modeling 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.

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

This chapter is published as van Doorn, J.B., Ly, A., Marsman, M., & Wagenmakers, E-J. (2018). Bayesian Inference for Kendall's Rank Correlation Coefficient. *The American Statistician*, *72*, 303–308.

7.1.1 Kendall's τ

Let $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$ be two data vectors each containing measurements of the same n units. For example, consider the association between French and math grades in a class of $n = 3$ children: Tina, Bob, and Jim; let $X = (8, 7, 5)$ be their grades for a French exam and $Y = (9, 6, 7)$ be their grades for a math 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 math 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:

$$\tau_{obs} = \frac{\sum_{1 \leq i < j \leq n}^{n} Q((x_i, y_i), (x_j, y_j))}{n(n-1)/2}, \tag{7.1}$$

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

$$Q((x_i, y_i)(x_j, y_j)) = \begin{cases} -1 & \text{if } (x_i - x_j)(y_i - y_j) < 0 \\ +1 & \text{if } (x_i - x_j)(y_i - y_j) > 0 \end{cases}. \tag{7.2}$$

Table 7.1 illustrates the calculation for our small data example. Applying Equation (7.1) gives $\tau_{obs} = 1/3$, an indication of a positive correlation between French and math 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 7.1: The pairs (i, j) for $1 \leq i < j \leq n$ and the concordance indicator function Q for the data example where $X = (8, 7, 5)$ and $Y = (9, 6, 7)$.

When $\tau_{obs} = 1$, all pairs of observations are concordant, and when $\tau_{obs} = -1$, 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 Figure 7.1 for an illustration of this method using our example data of French and math grades. These tools make for a straightforward and intuitive calculation and interpretation of Kendall's τ.

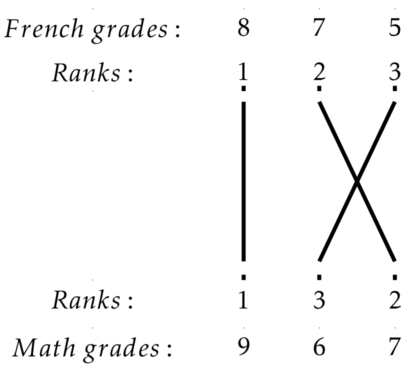


Figure 7.1: A visual interpretation of Kendall's τ_{obs} through the formula: $1 - \frac{4z}{n(n-1)}$, where z is the number of intersections of the lines. In this case, $n = 3$, $z = 1$, and $\tau_{obs} = 1/3$.

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 & Ghahramani, 2009; Labadi, Masuadi, & Zarepour, 2014) and Pòlya tree priors (Y. Chen & Hanson, 2014; Holmes, Caron, Griffin, & Stephens, 2015), this chapter will outline a different approach, one that permits an intuitive and direct interpretation.

7.1.2 Modeling Test Statistics

In order to compute Bayes factors for Kendall's τ we start with the approach pioneered by Johnson (2005) and Yuan & Johnson (2008). These authors established bounds for Bayes factors based on the sampling distribution of the standardized value of τ, denoted by T^*, which will be formally defined in section 7.2.1. Using the Pitman translation alternative, where a non-centrality parameter is used to distinguish between the null and alternative hypotheses (Randles & Wolfe, 1979), Johnson and colleagues specified the following hypotheses:

$$\mathcal{H}_0 : \theta = \theta_0, \tag{7.3}$$

$$\mathcal{H}_1 : \theta = \theta_0 + \frac{\Delta}{\sqrt{n}}, \tag{7.4}$$

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 (Hotelling & Pabst, 1936; Noether, 1955; Chernoff & Savage, 1958), with likelihoods

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

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

The prior on Δ is specified by Yuan and Johnson as

$$\Delta \sim N(0, \kappa^2),$$

where κ 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:

$$\text{BF}_{01} = \sqrt{1 + \frac{9}{4}\kappa^2} \, \exp\left(-\frac{\kappa^2 T^{*2}}{2\kappa^2 + \frac{8}{9}}\right). \tag{7.5}$$

Next, Yuan and Johnson calculated an upper bound on BF_{10}, (i.e., a lower bound on BF_{01}) by maximizing over the parameter κ.

7.1.3 Challenges

Although innovative and compelling, the approach advocated by Yuan & Johnson (2008) does have a number of non-Bayesian elements, most notably the data-dependent maximization over the parameter κ that results in a data-dependent prior distribution. Moreover, the definition of $\mathcal{H}_1$ depends on n: as $n \to \infty$, $\mathcal{H}_1$ 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 κ not by maximization 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 $\theta_0 = 0$.

7.2 Methods

7.2.1 Defining T^*

As mentioned above, Yuan & Johnson (2008) use the standardized version of τ_{obs}, denoted T^* (Kendall, 1938) which is defined as follows:

$$T^* = \frac{\sum_{1\leq i<j\leq n}^{n} Q((x_i, y_i), (x_j, y_j))}{\sqrt{n(n-1)(2n+5)/18}}. \tag{7.6}$$

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 & Gibbons, 1990).

7.2.2 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, Verhagen, & Wagenmakers (2016) use a symmetric beta prior distribution ($\alpha = \beta$) on the domain [-1,1], that is:

$$p(\rho) = \frac{2^{1-2\alpha}}{\mathcal{B}(\alpha, \alpha)} \times (1-\rho^2)^{(\alpha-1)}, \rho \in (-1,1), \tag{7.7}$$

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):

$$\tau = \frac{2}{\pi}\arcsin(\rho). \tag{7.8}$$

We can use this relationship to transform the beta prior in (7.7) on ρ to a prior on τ given by:

$$p(\tau) = \pi\frac{2^{-2\alpha}}{\mathcal{B}(\alpha, \alpha)} \times \cos\left(\frac{\pi\tau}{2}\right)^{(2\alpha-1)}, \tau \in (-1,1). \tag{7.9}$$

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 induces a non-uniform distribution on τ, that is,

$$p(\tau) = \frac{\pi}{4}\cos\left(\frac{\pi\tau}{2}\right). \tag{7.10}$$

Values of $\alpha \neq 1$ can be specified to induce different prior distributions on τ. 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 $\alpha = \beta = 1/2$. As is easily confirmed by entering these values in (7.9), this choice induces a uniform prior distribution for 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 & Sun, 2008; Gelman, 2013), by transforming ρ with the inverse of the expression given in (7.8):

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

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

7.2.3 Posterior Distribution and Bayes Factor

Removing $\sqrt{n}$ from the specification of $\mathcal{H}_1$ by substituting $\Delta\sqrt{n}$ for Δ, the likelihood function under $\mathcal{H}_1$ equals a normal density with mean $\mu = \frac{3}{2}\Delta\sqrt{n}$ and standard deviation $\sigma = 1$:

$$p(T^*|\theta_0 + \Delta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(T^* - (3/2)\Delta\sqrt{n})^2}{2}\right). \tag{7.11}$$

Combining this normal likelihood function with the prior from (7.9) yields the posterior distribution for Kendall's τ. Next, Bayes factors can be computed as the ratio of the prior and posterior ordinate at the point under test (i.e., the Savage-Dickey density ratio, Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). In the case of testing independence, the point under test is $\tau = 0$, leading to the following ratio: $\mathrm{BF}_{01} = \frac{p(\tau=0|y)}{p(\tau=0)}$, which is analogous to:

$$\mathrm{BF}_{01} = \frac{p(T^*|\theta_0)}{\int p(T^*|\theta_0 + \Delta)p(\Delta)\mathrm{d}\Delta}, \tag{7.12}$$

and in the case of Kendall's τ translates to

$$\mathrm{BF}_{01} = \frac{\exp(-\frac{T^{*2}}{2})}{\int\limits_{-1}^{1} \exp\left(-\frac{(T^* - (3/2)\tau\sqrt{n})^2}{2}\right)\left(\pi\frac{2^{-2\alpha}}{B(\alpha,\alpha)} \times \cos\left(\frac{\pi\tau}{2}\right)^{(2\alpha-1)}\right)\mathrm{d}\tau}. \tag{7.13}$$

7.2.4 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, Genest, & Hallin (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] Figure 7.2 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

[2] For more information on copulas see Nelsen (2006), Genest & Favre (2007), and Colonius (in press).

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

results were obtained using Frank, Clayton, and Gumbel copulas. As is the case under $\mathcal{H}_0$ (e.g., Ferguson et al., 2000; Kendall & 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., $\tau = 0$), the variance of these normal distributions equals 1. As the population correlation grows (i.e., $|\tau| \to 1$), 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 & Gibbons, 1990):

$$\sigma^2_{T^*} \leq \frac{2.5n(1-\tau^2)}{2n+5}. \tag{7.14}$$

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 favor 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., $\mathrm{BF}_{10} \geq 10$).

7.3 Results

7.3.1 Bayes Factor Behavior

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., $\alpha = \beta = 1$), Figure 7.3 illustrates BF_{10} as a function of n, for three values of τ_{obs}. The lines for $\tau_{obs} = 0.2$ and $\tau_{obs} = 0.3$ show that BF_{10} for a true $\mathcal{H}_1$ increases exponentially with n, as is generally the case. For $\tau_{obs} = 0$, the Bayes factor decreases as n increases.

7.3.2 Comparison to Pearson's ρ

In order to put the result in perspective, the Bayes factors for Kendall's tau (i.e., BF^{τ}_{10}) can be compared to those for Pearson's ρ (i.e., BF^{ρ}_{10}). The Bayes factors for Pearson's ρ are based on Jeffreys (1961, see also Ly et al., 2016), who used the uniform prior on ρ. Figure 7.4 shows that the relationship between BF^{τ}_{10} and BF^{ρ}_{10} 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, $\mathrm{BF}^{\tau}_{10} < \mathrm{BF}^{\rho}_{10}$ in the case of evidence in favor

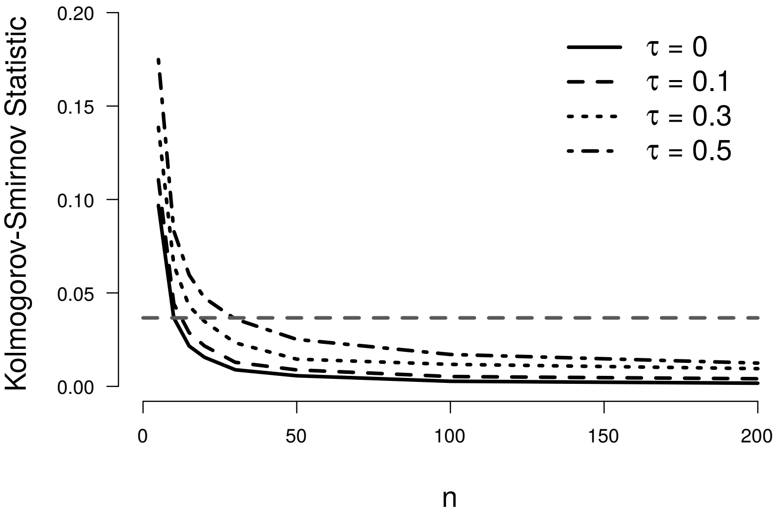


Figure 7.2: Quality of the normal approximation to the sampling distribution of T^*, as assessed by the Kolmogorov-Smirnov statistic. As n grows, the quality of the normal approximation increases exponentially. Larger values of τ necessitate larger values of n to achieve the same quality of approximation. 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".

of $\mathcal{H}_1$ (left panel of Figure 7.4). When evidence is in favor of $\mathcal{H}_0$, i.e. $\tau = 0$, BF_{10}^{τ} and BF_{10}^{ρ} perform similarly (right panel of Figure 7.4).

7.3.3 Real Data Example

Willerman, Schultz, Rutledge, & Bigler (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 gray matter pixels. The data are presented in the top left panel of Figure 7.5. Bayes factor hypothesis testing of Pearson's ρ yields $\text{BF}_{10}^{\rho} = 5.16$, which is illustrated in the middle left panel. This means 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 Figure 7.5 presents the effect of this monotonic transformation on the data. The middle right panel illustrates how the transformation decreases BF_{10}^{ρ} to 1.28. The bottom left panel presents our Bayesian analysis on Kendall's τ, which yields a BF_{10}^{τ} 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.

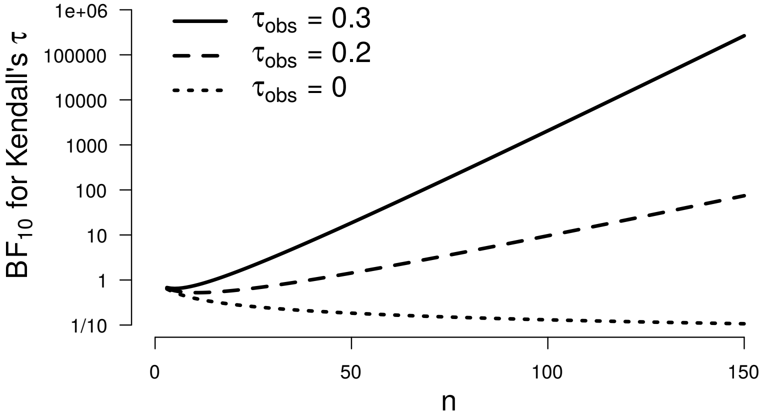


Figure 7.3: Relation between BF_{10} and sample size ($3 \le n \le 150$) for three values of Kendall's τ.

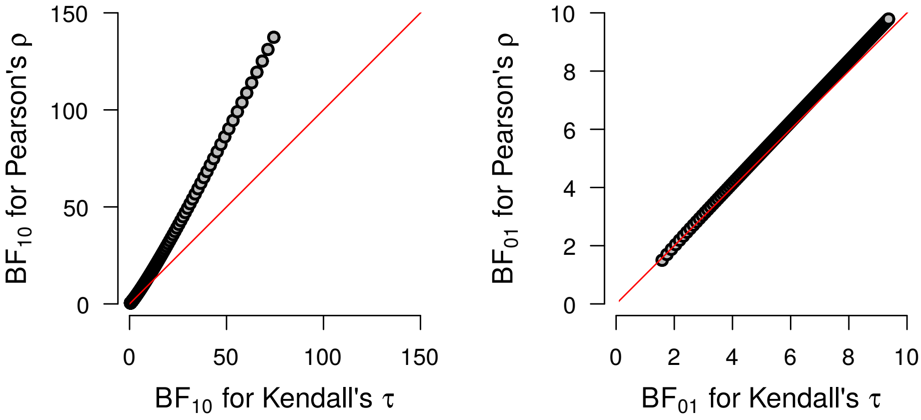


Figure 7.4: Relation between the Bayes factors for Pearsons ρ and Kendall's $\tau = 0.2$ (left) and Kendall's $\tau = 0$ (right) as a function of sample size (i.e., $3 \le n \le 150$). The data are normally distributed. Note that the left panel shows BF_{10} and the right panel shows BF_{01}. The diagonal line indicates equivalence.

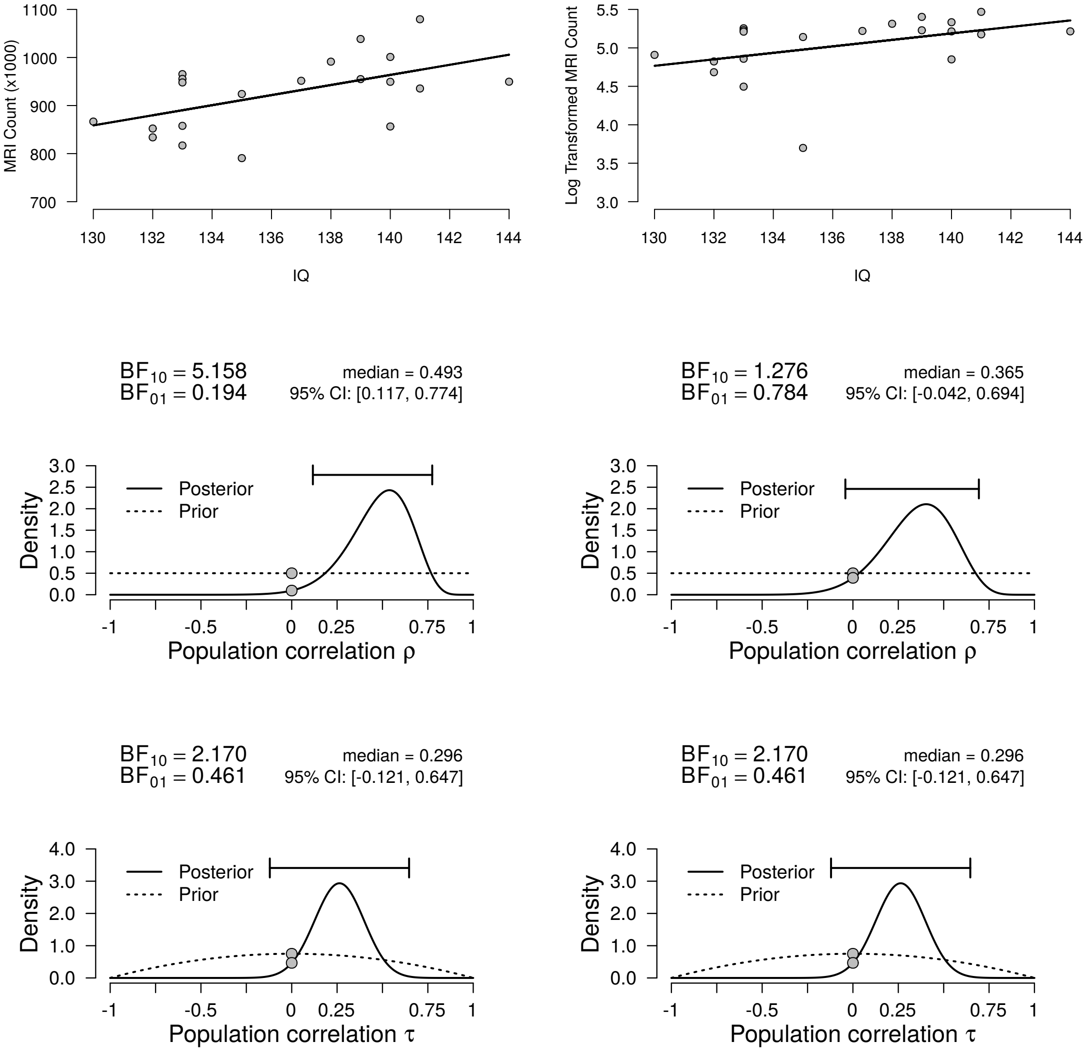


Figure 7.5: Bayesian inference for Kendall's τ illustrated with data on IQ and brain size (Willerman et al. 1991). The left column presents the relation between brain size and IQ, analyzed using Pearson's ρ (middle panel) and Kendall's τ (bottom panel). The right column presents the results after a log transformation of brain size. Note that the transformation affects inference for Pearson's ρ, but does not affect inference for Kendall's τ.

7.4 Concluding Comments

This chapter outlined a nonparametric Bayesian framework for inference about Kendall's tau. The framework is based on modeling test statistics and assigning a prior by means of a parametric yoking procedure. The framework produces a posterior distribution for Kendall's tau, and –via the Savage-Dickey density ratio test– also yields a Bayes factor that quantifies the evidence for the absence of a correlation.

Our general procedure (i.e., modeling 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 modeling of test statistics that have non-Gaussian limiting distributions.

Chapter 8

Bayesian Estimation of Kendall's τ Using a Latent Normal Approach

Abstract

The rank-based association between two variables can be modeled by introducing a latent normal level to ordinal data. We demonstrate how this approach yields Bayesian inference for Kendall's τ, improving on a recent Bayesian solution based on its asymptotic properties.

8.1 Introduction

Kendall's τ is a popular rank-based correlation coefficient. Compared to Pearson's ρ, Kendall's τ is robust to outliers, invariant under monotonic transformations, and has an intuitive interpretation (Kendall & Gibbons, 1990). Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two data vectors each containing ranked measurements of the same n units. For instance, x could be the rank ordered scores on a math exam and y the rank ordered scores on a geography exam, for n test-takers. A *concordant* pair is defined as a pair of subjects (i, j) where subject i has a higher score on x and y compared to subject j, whereas a *discordant* pair is defined as one where i scores higher on y, but j scores higher on x, or the other way around. Kendall's τ is defined as the difference between the number of concordant and discordant pairs, expressed as proportion of the total number of pairs:

$$\tau = \frac{\sum_{1 \le i < j \le n}^{n} Q((x_i, y_i), (x_j, y_j))}{n(n-1)/2}, \tag{8.1}$$

where the denominator is the total number of pairs and Q is the concordance indicator function, which is defined by:

$$Q((x_i, y_i)(x_j, y_j)) = \begin{cases} -1 & \text{if } (x_i - x_j)(y_i - y_j) < 0, \\ +1 & \text{if } (x_i - x_j)(y_i - y_j) > 0, \end{cases} \tag{8.2}$$

This chapter is published as van Doorn, J.B., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2019). Bayesian Estimation of Kendall's tau Using a Latent Normal Approach. *Statistics and Probability Letters*, *145*, 268–272. Also available as *ArXiv preprint*: https://arxiv.org/abs/1703.01805 which returns -1 if a pair is discordant, and returns $+1$ if a pair is concordant. However, due to the nonparametric nature of Kendall's τ and the lack of a likelihood function for the data, Bayesian inference is not trivial.

An innovative method for overcoming this problem was proposed by Johnson (2005), and involves the modeling of the test statistic itself, rather than the data. This method has been applied to Kendall's τ by Yuan & Johnson (2008), and was recently developed by van Doorn, Ly, Marsman, & Wagenmakers (2018). The inferential framework that follows from this work uses the limiting normal distribution of the test statistic T^* (Hotelling & Pabst, 1936; Noether, 1955), where

$$T^* = \tau\sqrt{\frac{9n(n-1)}{4n+10}}. \tag{8.3}$$

Under $\mathcal{H}_0$, this limiting normal distribution is the standard normal, whereas under $\mathcal{H}_1$, this distribution is specified with a non-centrality parameter Δ for the mean, and a sampling variance of 1.

However, the method —henceforth the *original asymptotic method*— might fall short on two counts. Firstly, the asymptotic assumptions only hold for sufficiently large n (i.e., $n \geq 20$, see van Doorn et al., 2018). Secondly, the variance of the sampling distribution of the test statistic depends on the population value of Kendall's τ. For $\tau = 0$, the sampling variance equals 1, but as $\mid \tau \mid \to 1$, the variance decreases to 0 (Kendall & Gibbons, 1990; Hotelling & Pabst, 1936).

In the current chapter, we will explore two corrections that aim to improve Bayesian inference for Kendall's τ:

1. Within the asymptotic framework, the observed value of Kendall's τ can be used to set its sampling variance. We label this the *enhanced asymptotic method*.

2. Within a Bayesian latent normal framework, a latent level correlation is obtained and transformed to Kendall's τ. We label this the *latent normal method*.

8.2 Correction Using The Sample τ

A first correction to consider is to use the sample value of Kendall's τ, denoted τ_{obs}, to estimate the sampling variance of T^*, denoted $\sigma^2_{T^*}$. A convenient expression for the upper bound of $\sigma^2_{T^*}$ in terms of τ is given in Kendall & Gibbons (1990):

$$\sigma^2_{T^*} \leq \frac{2.5n(1-\tau^2)}{2n+5}. \tag{8.4}$$

Using τ_{obs} as an estimate of τ provides a somewhat crude approximation to the sampling distribution of T^*. However, compared to using $\sigma^2_{T^*} = 1$ as in the original asymptotic method, working with the upper bound will result in a more narrow posterior for cases where $\tau \neq 0$. However, the enhanced asymptotic method still suffers from the use of asymptotic assumptions about the sampling distribution and variance of the test statistic.

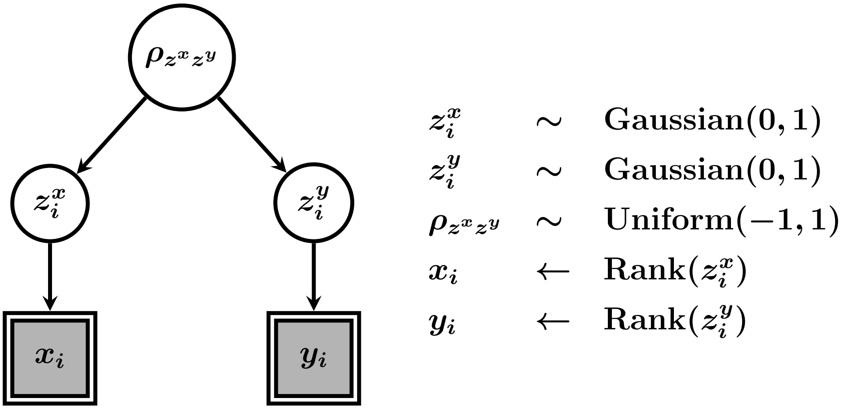

$\rho_{z^x z^y}$
z_i^x
z_i^y
x_i
y_i
$z_i^x \sim \text{Gaussian}(0,1)$
$z_i^y \sim \text{Gaussian}(0,1)$
$\rho_{z^x z^y} \sim \text{Uniform}(-1,1)$
$x_i \leftarrow \text{Rank}(z_i^x)$
$y_i \leftarrow \text{Rank}(z_i^y)$

Figure 8.1: A graphical model of the latent normal method. Here, x and y are observed rank data. The latent level is denoted with z^x and z^y, and $\rho_{z^x z^y}$ represents the latent correlation.

8.3 Correction Using The Latent Normal Approach

8.3.1 Latent Normal Models

Several latent variable models quantify the association between two ordinal variables. These methods often introduce a latent bivariate normal distribution to the ordinal variables, where the association between variables is modeled through a latent correlation (Pearson, 1900; Olssen, 1979; Pettitt, 1982; Albert, 1992b; Alvo & Yu, 2014). The observed rank data (x, y) can then be seen as the ordinal manifestations of the continuous latent variables (z^x, z^y), which have a bivariate normal distribution. Figure 8.1 offers a graphical representation of such a model. Using this methodology, the nonparametric problem of ordinal analysis is transformed to a parametric data augmentation problem.

8.3.2 Posterior Distribution for the Latent Correlation

The joint posterior can be decomposed as follows:

$$P(z^x, z^y, \rho_{z^x,z^y} \mid x, y) \propto P(x, y \mid z^x, z^y) \times P(z^x, z^y \mid \rho_{z^x,z^y}) \times P(\rho_{z^x,z^y}). \tag{8.5}$$

The second factor on the right-hand side is the bivariate normal distribution of the latent scores given the latent correlation:

$$\begin{pmatrix} z^x \\ z^y \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{z^x,z^y} \\ \rho_{z^x,z^y} & 1 \end{pmatrix}\right]. \tag{8.6}$$

The factor $P(x, y \mid z^x, z^y)$ consists of a set of indicator functions that map the observed ranks to latent scores, such that the ordinal information is preserved. For the value z_i^x, this means that its range is truncated by the lower and upper thresholds that are respectively defined as:

$$a_i^x = \max_{j:x_j<x_i} \left(z_j^x\right) \tag{8.7}$$

$$b_i^x = \min_{j:x_j>x_i} \left(z_j^x\right). \tag{8.8}$$

The third factor is the prior distribution on the latent correlation. In the remainder of this chapter, the prior is specified by a uniform distribution on $(-1, 1)$ (but see Berger & Sun 2008; Ly, Verhagen, & Wagenmakers 2016).

The general Bayesian framework for estimating the latent correlation involves data augmentation through a Gibbs sampling algorithm (Geman & Geman, 1984) for the latent values z^x and z^y, combined with a random walk Metropolis-Hastings sampling algorithm for ρ_{z^x,z^y}. At sampling step s:

1. For each value of z_i^x, sample from a truncated normal distribution:

   $$(z_i^x \mid z_{i'}^x, z_i^y, \rho_{z^x,z^y}) \sim N\left(z_i^y \rho_{z^x,z^y}, 1 - \rho_{z^x,z^y}^2\right) \mathbf{1}_{(a_i^x,b_i^x)}(z_i^x),$$

   where $\mathbf{1}_{(a_i^x,b_i^x)}(z_i^x)$ indicates truncation between the lower threshold a_i^x given in (8.7) and the upper threshold b_i^x given in (8.8).

2. For each value of z_i^y, the sampling procedure is analogous to step 1.

3. Sample a new proposal for ρ_{z^x,z^y}, denoted ρ_{z^x,z^y}^*, from the asymptotic normal approximation to the sampling distribution of Fisher's z-transform of ρ (Fisher, 1915):

   $$\tanh^{-1}(\rho_{z^x,z^y}^*) \sim N\left(\tanh^{-1}\left(\rho_{z^x,z^y}^{s-1}\right), \frac{1}{n-3}\right).$$

   The acceptance rate α is determined by the likelihood ratio of $(z^x, z^y \mid \rho_{z^x,z^y}^*)$ and $(z^x, z^y \mid \rho_{z^x,z^y}^{s-1})$, where each likelihood is determined by the bivariate normal distribution in (8.6):

   $$\alpha = \min\left(1, \frac{P(z^x, z^y \mid \rho_{z^x,z^y}^*)(1 - (\rho_{z^x,z^y}^*)^2)}{P(z^x, z^y \mid \rho_{z^x,z^y}^{s-1})(1 - (\rho_{z^x,z^y}^{s-1})^2)}\right),$$

   where $(1 - \rho^2)$ is the Jacobian of Fisher's z-transform.

Repeating the algorithm a sufficient number of times yields samples from ρ_{z^x,z^y} given z^x, z^y, thus, $P(\rho_{z^x,z^y} \mid z^x, z^y)$, and the posterior of z^x, z^y, that is, $P(z^x, z^y \mid x, y)$.

8.3.3 Relation to Kendall's τ

With the posterior distribution for the latent ρ_{z^x,z^y} in hand, the transition to the posterior distribution for Kendall's τ can be made using Greiner's relation (Greiner, 1909; Kruskal, 1958). This relation, defined in 7.8, enables the transformation of Pearson's ρ to Kendall's τ when the data follow a bivariate normal distribution. The latent normal framework thus models the posterior distribution for Kendall's τ as

$$P(\tau \mid x,y) = P(G(\rho) \mid x,y) = \int\int P(G(\rho) \mid z^x, z^y) P(z^x, z^y \mid x,y) dz^x dz^y.$$

Introducing the latent normal level to the observed variables enables the link between Pearson's ρ and Kendall's τ, and turns posterior inference for Kendall's τ into a parametric data augmentation problem that can be solved with the above MCMC-methods. Thus, Greiner's relation can be applied to the posterior samples of ρ_{z^x,z^y} to yield posterior samples of τ. Furthermore, the application of Greiner's relation in this manner implicitly alters the prior from a uniform distribution on the latent correlation to the prior distribution on Kendall's τ given in 7.9.

8.4 Results: Simulation Study

The performance of the original asymptotic method, the enhanced asymptotic method, and the latent normal method was assessed with a simulation study. For four values of τ (0, 0.2, 0.4, 0.7) and three values of n (10, 20, 50), 10,000 data sets were generated under four copula models: Clayton, Gumbel, Frank, and Gaussian (Sklar, 1959; Nelsen, 2006; Genest & Favre, 2007; Colonius, 2016). Using Sklar's theorem, copula models decompose a joint distribution into univariate marginal distributions and a dependence structure (i.e., the copula). The aforementioned copulas are governed by Kendall's τ, so the performance of each method can be assessed through a parameter recovery simulation study. Furthermore, the univariate marginal distributions can be transformed to any other distribution using the cumulative distribution function and its inverse. Because these functions are monotonic, this does not affect the copula or ordinal information in the synthetic data and therefore vastly increases the scope of the simulation study.

For each data set, a posterior distribution was obtained using the three methods and the population value of τ was estimated using the posterior median. Per combination of n and τ, this resulted in $10,000$ posterior distributions. For an overall view of each method's performance, Figure 8.2 shows the quantile averaged posterior distributions, along with a vertical line indicating the population value of τ. The data in Figure 8.2 were generated using the Clayton copula; other copula models yielded highly similar results. The quantile averaged posteriors indicate no difference between the inferential methods under $\mathcal{H}_0$, which corroborates the assumption of $\sigma^2_{T^*} = 1$ when $\tau = 0$. However, the difference in methods becomes pronounced in the scenario where $n = 10$ and $\tau = 0.7$. Both asymptotic approaches show a degree of underestimation, and yield a relatively broad posterior distribution. In the panels where $\tau \neq 0$, the misspecification of the sampling variance also becomes clear, as it is overestimated and results in a wider posterior distribution compared to the latent normal method. Although the assumption of latent normality is the price to pay for the Bayesian latent normal methodology, the simulation results indicate robustness of the method to various violations of this assumption.[1]

8.5 Concluding Comments

This chapter has outlined two methods of improving the Bayesian inferential framework in cases where n is low and/or τ is high. Although an extension of the asymptotic framework performs somewhat better than the original asymptotic framework in van Doorn et al. (2018), both are outperformed by the latent normal approach. Under $\mathcal{H}_0$, the methods do not differ from each other, underscoring the validity of the general framework.

The outlined methods are useful for both estimation and hypothesis testing. In the former case, the posterior distribution enables point estimation through the posterior median, or interval estimation through the credible interval. For hypothesis testing, the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010) can be used to obtain Bayes factors (Kass & Raftery, 1995). A concrete example is presented in the online appendix. Because the method uses only the ordinal information in the data, it retains the robust properties of Kendall's τ, such as invariance to monotone transformations, robustness to outliers or violations of normality, and ability to detect nonlinear monotone relations.

[1] R-code, plots, and further details of the simulation study are available at https://osf.io/u7jj9/.

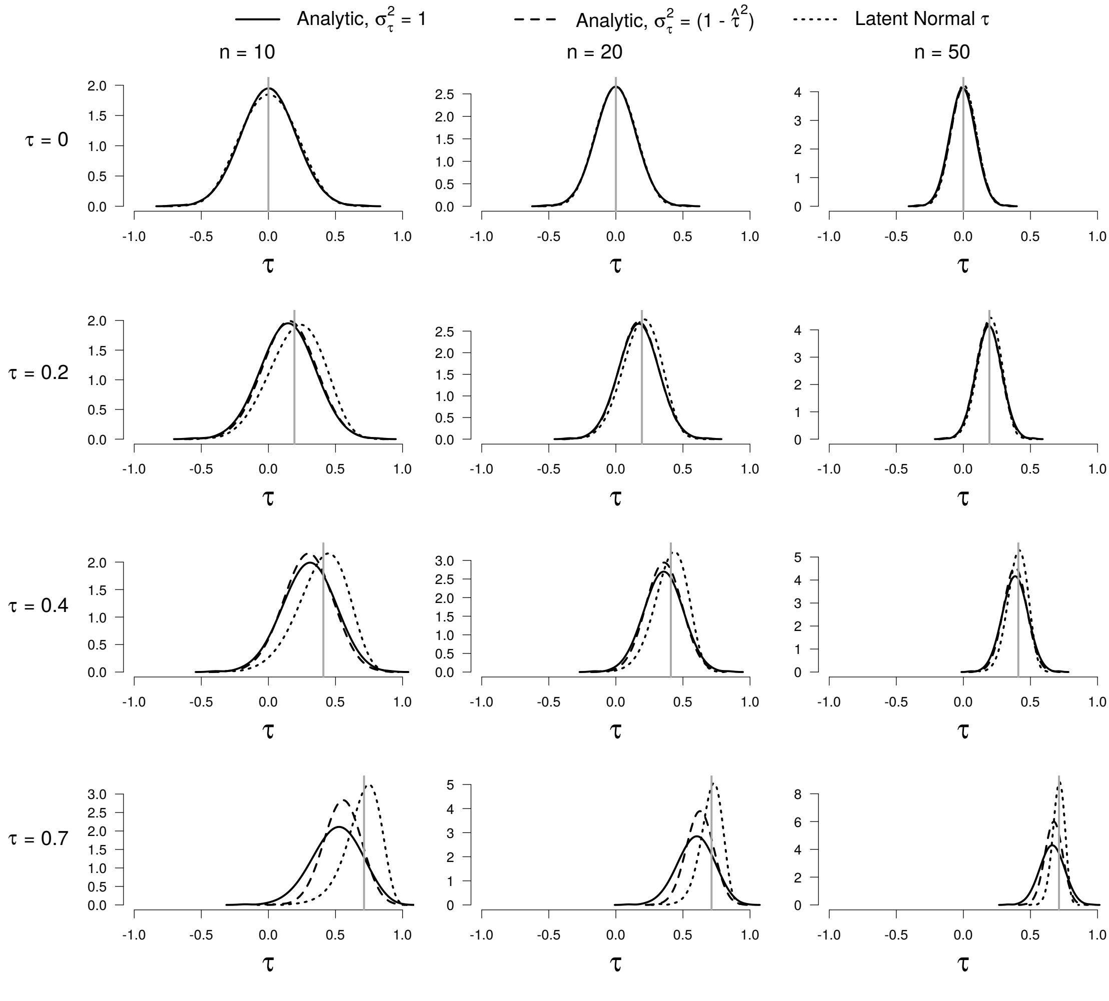


Figure 8.2: To illustrate the performance of the three methods, quantile averaged posterior distributions for several values of τ and n are shown. Each column corresponds to a value of n, and each row corresponds to a value of τ. The quantile averaged posterior distributions were obtained with 10,000 synthetic datasets per combination of n and τ. The vertical gray line indicates the population value of τ.

Chapter 9

Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ

Abstract

Bayesian inference for rank-order problems is frustrated by the absence of an explicit likelihood function. This hurdle can be overcome by assuming a latent normal representation that is consistent with the ordinal information in the data: the observed ranks are conceptualized as an impoverished reflection of an underlying continuous scale, and inference concerns the parameters that govern the latent representation. We apply this generic data-augmentation method to obtain Bayes factors for three popular rank-based tests: the rank sum test, the signed rank test, and Spearman's ρ_s.

9.1 Introduction

The debate on alternatives to null hypothesis significance tests based on p-values (R. Wasserstein & Lazar, 2016) has led to a renewed interest in the Bayesian alternative known as the Bayes factor. Advantages of such Bayesian tests include the ability to provide evidence in favor of *both* the null and the alternative hypotheses (Dienes, 2014), the ability to straightforwardly synthesize evidence to assess replicability (Ly, Etz, et al., 2018), and the ability to monitor the evidence as the data accumulate (Rouder, 2014); see Wagenmakers, Marsman, et al. (2018) and Dienes & McLatchie (2018) for further details on the advantages of Bayesian inference. These advantages are met by the recently proposed Bayes factors for the classical two- and one-sample t-tests (Rouder et al., 2009), as well as for the Bayes factor for Pearson's correlation (Ly, Marsman, & Wagenmakers, 2018). These tests

This chapter is published as van Doorn, J.B., Ly, A., Marsman, M., & Wagenmakers, E.–J. (2020). Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ. *Journal of Applied Statistics*, *47*, 2984–3006. Also available as *ArXiv preprint*: https://arxiv.org/abs/1712.06941 have become increasingly popular in the applied sciences. The goal of this paper is to extend these parametric Bayes factors to their rank-based counterparts.

Rank-based statistical procedures offer a range of advantages over their parametric counterparts. First, they are robust to outliers and to violations of distributional assumptions, which occur frequently in many practical applications, such as the analysis of questionnaire data. Second, they are invariant under monotonic transformations, which is desirable when interest concerns a hypothesized concept (e.g., rat intelligence) whose relation to the measurement scale is only weakly specified (e.g., brain volume or log brain volume could be used as a predictor; without a process model that specifies how brain physiology translates to rat intelligence, neither choice is privileged). Third, many data sets are inherently ordinal (e.g., Likert scales, where survey participants are asked to indicate their opinion on, say, a 7-point scale ranging from 'disagree completely' to 'agree completely'). Finally, rank-based procedures perform better than their fully parametric counterparts when assumptions are violated, with little loss of efficiency when the assumptions do hold (Hollander & Wolfe, 1973).

Prominent rank-based tests include the Mann-Whitney-Wilcoxon rank sum test (i.e., the rank-based equivalent of the two-sample t-test), the Wilcoxon signed rank test (i.e., the rank-based equivalent of the paired sample t-test), and Spearman's ρ_s (i.e., a rank-based equivalent of the Pearson correlation coefficient). These ordinal tests were developed within the frequentist statistical paradigm, and Bayesian analogues through Bayes factor hypothesis testing have, to the best of our knowledge, not yet been proposed. We speculate that the main challenge in the development of Bayesian hypothesis tests for ordinal data is the lack of a straightforward likelihood function. As stated by Harold Jeffreys (Jeffreys, 1939, pp. 178-179) for the case of Spearman's ρ_s:

> "The rank correlation, while certainly useful in practice, is difficult to interpret. It is an estimate, but what is it an estimate of? That is, it is calculated from the observations, but a function of the observations has no relevance beyond the observations unless it is an estimate of a parameter in some law. Now what can this law be? [...] the interpretation is not clear."

This difficulty can be overcome by postulating a latent, normally distributed level for the observed data (i.e., data augmentation). In other words, the rank data are conceptualized to be an impoverished reflection of richer latent data that are governed by a specific likelihood function. The latent normal distribution was chosen for computational convenience and ease of interpretation. This general procedure is widely known as data augmentation (Tanner & Wong, 1987; Albert & Chib, 1993), and Bayesian inference for the parameters of interest (e.g., a location difference parameter δ or an association parameter ρ) can then be achieved using Markov chain Monte Carlo (MCMC) sampling. In other words, we can use the latent normal approach to overcome the lack of a likelihood function, and thus enable a Bayesian approach to rank-based testing.

Below we first outline the general latent normal framework and then develop Bayesian counterparts for three popular frequentist rank-based procedures: the rank sum test, the signed rank test, and Spearman's rank correlation. Each of these developed Bayesian tests is accompanied by a simulation study that assesses the behavior of the test and a data example that highlights the desirable properties of rank-based inference, as well as the applicability of our proposed tests.

9.2 General Methodology

In the Bayesian framework, the posterior distribution of the parameter of interest θ is often used for hypothesis testing and parameter estimation. The posterior distribution is proportional to the likelihood, i.e., $f(\text{data} \mid \theta)$, times the prior, i.e., $\pi(\theta)$, that is,

$$\pi(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) \times \pi(\theta). \tag{9.1}$$

In the parametric case, this is often straightforward. For rank-based procedures, however, $f(\text{data} \mid \theta)$ is unavailable and to overcome this complication, we can use a latent normal framework.

9.2.1 Latent Normal Models

Latent normal models were first introduced by Pearson (1900) as a means of modeling data from a 2×2 cross-classification table. The method was later extended by Pearson & Pearson (1922) to accommodate $r \times s$ tables. Instead of modeling the count data directly for the 2×2 case, Pearson assumed a latent bivariate normal level with certain governing parameters. In the case of cross-classification tables, the governing parameter is the *polychoric correlation coefficient* (PCC) and refers to Pearson's correlation on the bivariate, latent normal level.

A maximum likelihood estimator for the PCC was developed by Olssen (1979) and Olssen et al. (1982), and a Bayesian framework for the PCC was later introduced by Albert (1992b). This idea was extended by Pettitt (1982) to rank likelihood models, where the latent boundaries are not estimated but determined directly by the latent scores (see also Hoff, 2007, 2009). For the two-sample location problem, a similar approach has been discussed by Savage (1956) and Brooks (1974, 1978), where a continuous distribution is assumed to be underlying the observed data. Further models for ordinal data are given in Mallows (1957), Fligner & Verducci (1986), Fligner & Verducci (1988), and Marden (1995). However, these methods omit Bayesian hypothesis testing through Bayes factors and/or lack a straightforward interpretation of the model parameters.

In general, the latent normal methodology allows one to transform ordinal problems to parametric problems. The resulting models that are discussed here have a data-generating process, are governed by easily interpretable parameters on the latent level, and enable Bayes factor hypothesis testing. A detailed sampling algorithm of the general methodology is presented in the next section.

9.2.2 Posterior Distribution

We elaborate the main idea of the latent normal approach with data consisting of two groups of samples. Let (r^x, r^y) be two vectors of ranked data, and z^x, z^y be the vectors of associated latent normal scores which depend on a model parameter θ. The latent normal posterior is then proportional to

$$\pi(z^x, z^y, \theta \mid x, y) \propto f(r^x, r^y \mid z^x, z^y) \times f(z^x, z^y \mid \theta) \times \pi(\theta) \tag{9.2}$$

Note how the parametric likelihood in (9.1) is now replaced by the product $f(r^x, r^y \mid z^x, z^y) \times f(z^x, z^y \mid \theta)$. As before, the third term on the right-hand side refers to the prior $\pi(\theta)$. The second term refers to the latent normal structure. For instance, in the two-sample case, we replace the generic θ by the population difference δ and take for $f(z^x, z^y \mid \theta)$ the product of two normal densities with unit variances, but a mean depending on δ, see below for further details. On the other hand, for inference on Spearman's ρ_s, we replace the generic θ by ρ, and take for $f(z^x, z^y \mid \theta)$ the centered bivariate normal density with unit variances, and correlation ρ.

The first term on the right-hand side of (9.2), i.e., $f(r^x, r^y \mid z^x, z^y)$ consists of a set of indicator functions, presented below, that connect the observed ranks to the unobserved latent normal scores, z^x, z^y such that the ordinal information (i.e., the ranking function) in the observations r^x, r^y is preserved. This is similar to the approach of Albert (1992a) and Albert & Chib (1993), who sampled latent scores for binary or polytomous response data from a normal distribution that was truncated with respect to the ordinal information of the data.

With (9.2) in hand, we have the specified the link between the data, the latent normal scores and parameters, and an MCMC sampler can be constructed in order to obtain the joint posterior distribution. This sampler takes as input the ordinal information of the observed data, and iteratively generates random parameter values θ as well as random latent scores z^x, z^y. The indicator function $f(r^x, r^y \mid z^x, z^y)$ ensures that the latent scores z^x, z^y retain the ordinal information in the data by truncating the latent normal likelihood $f(z^x, z^y \mid \theta)$. For the latent value z_i^x this means that its range is truncated by the lower and upper thresholds that are respectively defined as:

$$a_i^x = \max_{j:\, r_j^x < r_i^x} \left(z_j^x\right) \tag{9.3}$$

$$b_i^x = \min_{j:\, r_j^x > r_i^x} \left(z_j^x\right). \tag{9.4}$$

For example, suppose that on a particular MCMC iteration we wish to augment the observed ordinal value r_i^x to a latent z_i^x; on the latent scale, the lower threshold a_i^x is given by the maximum latent value associated with all r^x lower than r_i^x, whereas the upper threshold b_i^x is determined by the minimum latent value associated with all r^x higher than r_i^x. This dependence between the scores can make the sampler inefficient. In order to remedy the high degree of autocorrelation that data augmentation can induce (van Dyk & Meng, 2001), we included an additive decorrelating step documented by S. Liu & Sabatti (2000) and Morey et al. (2008).

9.2.3 Estimation and Testing

After obtaining the joint posterior distribution through the MCMC sampling algorithm outlined above, we can either focus on *estimation* and present the marginal posterior distribution for the parameter of interest θ, or we can conduct a Bayes factor *hypothesis test* and compare the predictive performance of a point-null hypothesis $\mathcal{H}_0$ (in which the parameter of interest is fixed at a predefined value θ_0) against that of an alternative hypothesis $\mathcal{H}_1$ (in which θ is free to vary; Kass & Raftery (1995); Jeffreys (1939); Ly et al. (2016)). The Bayes factor can be interpreted as a predictive updating factor, that is, degree to which the observed data drive a change from prior to posterior odds for the hypothesis of interest:

$$\underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{Bayes factor}_{10}} = \underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\text{Posterior odds}} \tag{9.5}$$

For instance, a Bayes factor $\text{BF}_{10} = 7$ implies that the data are seven times more likely under $\mathcal{H}_1$ then under $\mathcal{H}_0$, whereas $\text{BF}_{10} = 1/9$ indicates that the data are 9 times more likely under the null than under the alternative.

For nested models, the Bayes factor be easily obtained using the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), that is, the ratio of the posterior and prior ordinate for the parameter of interest θ, under $\mathcal{H}_1$, evaluated at the point of testing θ_0 specified under $\mathcal{H}_0$:

$$\text{BF}_{10} = \frac{p(\theta_0 \mid \mathcal{H}_0)}{p(\theta_0 \mid \text{data}, \mathcal{H}_1)}. \tag{9.6}$$

9.3 Case 1: Wilcoxon Rank Sum Test

9.3.1 Background

The ordinal counterpart to the two-sample t-test is known as the Wilcoxon rank sum test (or as the Mann-Whitney-Wilcoxon U test). It was introduced by Wilcoxon (1945) and further developed by Mann & Whitney (1947), who worked out the statistical properties of the test. Let $x = (x_1, ..., x_{n_1})$ and $y = (y_1, ..., y_{n_2})$ be two data vectors that contain measurements of n_1 and n_2 units, respectively. The aggregated ranks r^x, r^y (i.e., the ranking of x and y together) are defined as:

$$r_i^x = \text{rank of } x_i \text{ among } (x_1, \ldots, x_{n_1}, y_1 \ldots y_{n_2}),$$

$$r_i^y = \text{rank of } y_i \text{ among } (x_1, \ldots, x_{n_1}, y_1 \ldots y_{n_2}).$$

The test statistic U is then given by summing over either r^x or r^y, and subtracting $\frac{n_x(n_x+1)}{2}$ or $\frac{n_y(n_y+1)}{2}$, respectively. In order to test for a difference between the two groups, the observed value of U can be compared to the value of U that corresponds to no difference. This point of testing is defined as $\frac{n_1 n_2}{2}$.

To illustrate the procedure, consider the following hypothetical example. In the movie review section of a newspaper, three action movies and three comedy movies are each assigned a star rating between 0 and 5. Let $X = (4, 3, 1)$ be the star ratings for the action movies, and let $Y = (2, 3, 5)$ be the star ratings for the comedy movies. The corresponding aggregated ranks are $R^x = (5, 3.5, 1)$ and $R^y = (2, 3.5, 6)$. The test statistic U is then obtained by summing over either R^x or R^y, and subtracting $\frac{3(3+1)}{2} = 6$, yielding 3.5 or 5.5, respectively. Either of these values can then be compared to the null point which is equal to $\frac{n_1 n_2}{2} = 4.5$.

The range of U depends on the sample sizes and to avoid this dependence, we consider the rank-biserial correlation, which is a standardized effect size of U instead. The rank-biserial correlation, denoted ρ_{rb}, is the correlation coefficient used as a measure of association between a nominal dichotomous variable and an ordinal variable. The transformation is as follows:

$$\rho_{rb} = 1 - \frac{2U}{n_1 n_2}. \tag{9.7}$$

When $\rho_{rb} = 1$ we now know that U is at its maximum. The rank-biserial correlation can also be expressed as the difference between the proportion of data pairs where $x_i > y_j$ versus $x_i < y_j$ (Cureton, 1956; Kerby, 2014):

$$\rho_{rb} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Q(x_i - y_j)}{n_1 n_2}, \tag{9.8}$$

where $Q(d_i)$ is the sign indicator function defined as

$$Q(d_i) = \begin{cases} -1 & \text{if } d_i < 0 \\ +1 & \text{if } d_i > 0 \end{cases}. \tag{9.9}$$

This provides an intuitive interpretation of the test procedure: each data point in x is compared to each data point in y and scored -1 or 1 if it is lower or higher, respectively. In the movie ratings data example, there are three pairs for which $x_i > y_j$, five pairs for which $x_i < y_j$, and one pair for which $x_i = y_j$, yielding an observed rank-biserial correlation coefficient of $\frac{3-5}{9} = -0.22$, which is an indication that comedy movies receive slightly more positive reviews.

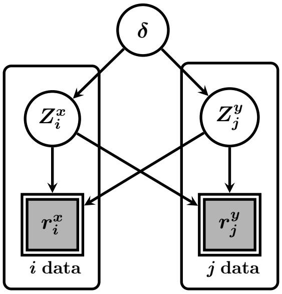
One argument to favor the Wilcoxon rank sum test over its parametric counterpart is provided by Pitman's asymptotic relative efficiency (ARE); that is, the ratio of the number of observations necessary to achieve the same level of power (Lehmann, 1999).[1] If ARE > 1 then we require fewer samples for U than for its parametric counterpart (van der Vaart, 2000).

[1] More precisely, let θ be a true parameter value and $\alpha, \beta \in (0, 1)$ fixed, then we denote by $N_{\text{T}}(\alpha, \beta, \theta)$ the number of samples necessary for a generic test statistic T at level α to reach the desired power of $1-\beta$ under θ computed using the asymptotical variance of the test statistic. The ARE of the parametric test over the Mann-Whitney-Wilcoxon U test is defined as $\text{ARE} = N_{\text{par}}(\alpha, \beta, \theta)/N_U(\alpha, \beta, \theta)$.

When the data are normally distributed as assumed under the parametric setting, then the rank sum test performs slightly poorer to the parametric two-sample t-test as ARE of $3/\pi \approx 0.955$ (Hodges & Lehmann, 1956; Lehmann, 1975). Thus, even when the distributional assumption of the t-test holds, the loss of the rank sum test in terms of sample sizes is about 4.5%. The ARE increases as the data distribution grows more heavy-tailed, with a maximum value of infinity. In addition, results for other distributions include the logistic distribution (ARE = $\pi^2/9 \approx 1.097$), the Laplace distribution (ARE = 1.5), and the exponential distribution (ARE = 3). Hence, relatively little is lost when using the Wilcoxon rank sum tests as compared to the parametric two-sample t-test when the parametric assumptions are met, but a lot is gained when the assumptions are violated.

9.3.2 Sampling Algorithm

For the Bayesian counterpart of the Wilcoxon rank sum test, we use the latent normal framework as elaborated on above. Specifically, the Bayesian data augmentation algorithm for the rank sum test follows the graphical model outlined in Figure 9.1. The ordinal information contained in the aggregated ranking constrains the corresponding values for the latent normal parameters Z^x and Z^y to lie within certain intervals (i.e., the ordinal information imposes truncation). The parameter of interest here is the effect size δ, the difference in location of the distributions for Z^x and Z^y. We follow Jeffreys (1961) and assign δ a Cauchy prior with scale parameter γ. For computational simplicity, this prior is implemented as a normal distribution with an inverse gamma prior on the variance, where the shape parameter is set to 0.5 and the scale parameter is set to $\gamma^2/2$ (Liang, German, Clyde, & Berger, 2008; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The difference with earlier work is that we set the latent normal variances σ to 1, as the rank data contain no information about the variance and the inclusion of σ in the sampling algorithm becomes redundant.



$$\begin{aligned}
\delta &\sim \text{Normal}(0, g)\\
g &\sim \text{Inverse Gamma}\left(\tfrac{1}{2}, \tfrac{\gamma^2}{2}\right)\\
Z_i^x &\sim \text{Normal}\left(-\tfrac{1}{2}\delta, 1\right)\\
Z_j^y &\sim \text{Normal}\left(\tfrac{1}{2}\delta, 1\right)\\
r_i^x &\leftarrow \text{Rank}(Z_i^x) \text{ among } (Z_1^x, ..., Z_n^x, Z_1^y, ..., Z_n^y)\\
r_j^y &\leftarrow \text{Rank}(Z_i^y) \text{ among } (Z_1^x, ..., Z_n^x, Z_1^y, ..., Z_n^y)
\end{aligned}$$

Figure 9.1: The graphical model underlying the Bayesian rank sum test. The latent, continuous scores are denoted by Z_i^x and Z_i^y, and their manifest rank values are denoted by x_i and y_j. The latent scores are assumed to follow a normal distribution governed by the parameter δ. This parameter is assigned a Cauchy prior distribution, which for computational convenience is reparameterized to a normal distribution with variance g (which is then assigned an inverse gamma distribution).

In order to sample from the posterior distributions of δ, Z^x and Z^y, we used

Gibbs sampling (Geman & Geman, 1984). Specifically, the sampling algorithm takes the aggregated ranks r^x, r^y as input and iteratively generates the latent δ, Z^x, and Z^y as follows, at sampling time point s:

1. For each i in $(1, \ldots, n_x)$, sample Z_i^x from a truncated normal distribution, where the lower threshold is a_i^x given in (9.3) and the upper threshold is b_i^x given in (9.4):

   $$(Z_i^x \mid z_{i'}^x, z_i^y, \delta) \sim \mathcal{N}_{(a_i^x,\, b_i^x)}\left(-\tfrac{1}{2}\delta, 1\right),$$

   where the subscripts of $\mathcal{N}$ indicate the interval that is sampled from.

2. For each i in $(1, \ldots, n_y)$, the sampling procedure for Z_i^y is analogous to step 1, with

   $$(Z_i^y \mid z_{i'}^y, z_i^x, \delta) \sim \mathcal{N}_{(a_i^y,\, b_i^y)}\left(\tfrac{1}{2}\delta, 1\right).$$

3. Sample δ from

   $$(\delta \mid z^x, z^y, g) \sim \mathcal{N}(\mu_\delta, \sigma_\delta),$$

   where

   $$\mu_\delta = \frac{2g(n_y\overline{z^y} - n_x\overline{z^x})}{g(n_x + n_y) + 4}$$

   $$\sigma_\delta^2 = \frac{4g}{g(n_x + n_y) + 4}.$$

4. Sample g from

   $$(G \mid \delta) \sim \text{Inverse Gamma}\left(1, \frac{\delta^2 + \gamma^2}{2}\right),$$

   where γ determines the scale (i.e., width) of the Cauchy prior on δ.

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of Z^x, Z^y, and δ. The posterior distribution of δ can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (9.6).

9.3.3 Simulation Study

In order to provide insight into the behavior of the inferential framework, a simulation study was performed. For three values of difference in location parameters, δ (0, 0.5, 1.5), and three values of n (10, 20, 50), 1,000 data sets were generated under various distributions: skew-normal, Cauchy, logistic, and uniform distributions. In one scenario, both groups have the same distributional shape (e.g., both follow a logistic distribution), and in a second scenario, one group follows the normal distribution and one group follows one of the aforementioned distributions.

First, the relationship between the observed rank statistic U and the latent normal Bayes factor was analyzed. Figure 9.2 illustrates this relationship, fitted with a cubic smoothing spline (J. Chambers & Hastie, 1992), for two logistic distributions ($\alpha = 20$). To show results for multiple values of n in one figure, the rank biserial correlation coefficient ρ_{rb} is plotted instead of U. The figure shows a clear relationship: when $\rho_{rb} = 0$, thus, U corresponds to the test value $n_1 n_2/2$, then the evidence in favor of $\mathcal{H}_0$ is at its maximum as one would expect. Similarly, when $|U|$ is maximal, that is, $|\rho_{rb}| = 1$, one has the most evidence against the null, which is apparent from the curves getting closer to 0. This relationship grows more decisive as n increases: both the peak at $\rho_{rb} = 0$ and the decay at $|\rho_{rb}| = 1$ are more prominent as n grows. The results are highly similar for the other distributions that were considered (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios). Since both statistics, ρ_{rb} and BF_{01}, depend solely on the ordinal information in the data, the observed relationship is not surprising. This result highlights and illustrates the robustness of the latent normal Bayes factor to violations of the assumptions of the parametric test: it illustrates the same robustness as the traditional W test statistic.

Second, the relationship between the latent normal Bayes factor and the parametric Bayes factor (Rouder et al., 2009) was analyzed. For both the parametric and rank-based Bayes factor, a default Cauchy prior with scale $1/\sqrt{2}$ is used. Figure 9.3 illustrates this relationship for all values of n and δ that were used, again in the scenario with two logistic distributions. Generally, the two Bayes factors are in agreement. In cases where δ deviates from 0, the parametric Bayes factor becomes more decisive (i.e., deviates from 1) compared to the latent normal Bayes factor. For distributions of data that violate the assumptions of the parametric test, such as the Cauchy distribution, the relationship between the two Bayes factors is notably less defined. In this case, the results of the rank-based Bayes factor are more reliable, which is expected based on the ARE results as the Cauchy is a heavy-tailed distribution. The parametric test greatly overestimates the variance and is no longer able to detect differences in location parameters (see the supplementary material), whereas the latent normal Bayes factor is unaffected by this. Note that the difference in performance is due to the use of the latent normal framework and not due to the prior, as both the parametric and rank-based Bayes factor use the same Cauchy prior.

9.3.4 Data Example

Cortez & Silva (2008) gathered data from 395 students concerning their math performance (scored between 1 and 20) and their level of alcohol intake (self-rated on a Likert scale between 1 and 5). Students passed the course if they scored ≥ 10, and we will test whether students who failed the course ($n_1 = 130$) had a higher self-reported alcohol intake than their peers who passed ($n_2 = 265$).

As alcohol intake was measured on a Likert scale, the data contain many ties and show extreme non-normality. These properties make this data set particularly suitable for the latent-normal rank sum test. The hypotheses are $\mathcal{H}_0 : \delta = 0$

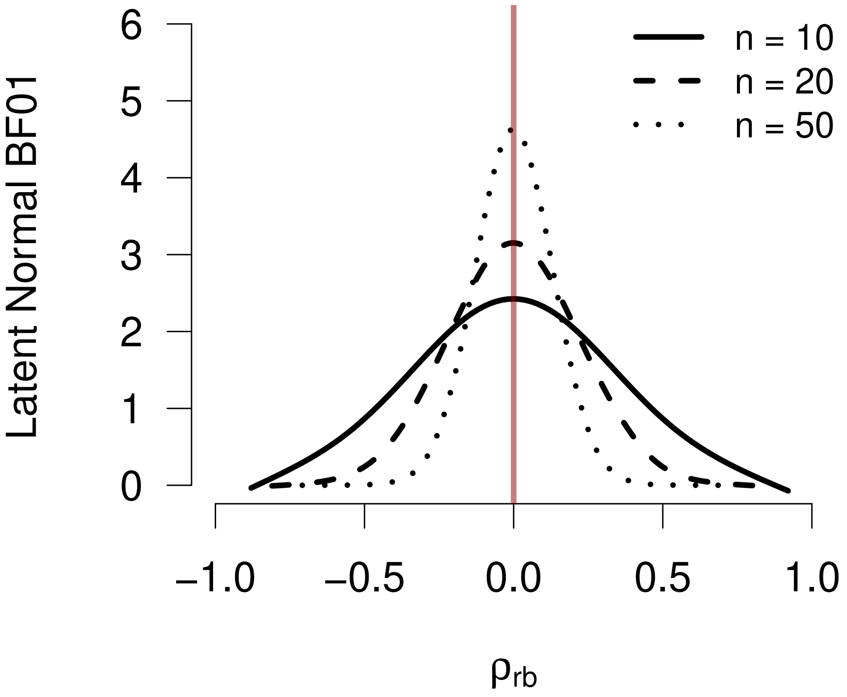


Figure 9.2: The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for logistic data. Because U is dependent on n, the rank biserial correlation coefficient is plotted on the x-axis instead of U. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when $\rho_{rb} = 0$. The further ρ_{rb} deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines depict smoothing splines fitted to the observed Bayes factors.

which is pitted against $\mathcal{H}_1 : \delta \neq 0$. For the rank-based Bayes factor we use the prior Cauchy prior with scale $1/\sqrt{2}$, that is, $\delta \sim \text{Cauchy}\left(0, \frac{1}{\sqrt{2}}\right)$. The null hypothesis posits that alcohol intake does not differ between the students who passed the course and those who failed. The alternative hypothesis posits the presence of an effect and assigns effect size a Cauchy distribution with scale parameter set to $1/\sqrt{2}$, as advocated by Morey & Rouder (2018). Figure 9.4 shows the resulting posterior distribution for δ under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for δ equals −0.121, with a 95% credible interval that ranges

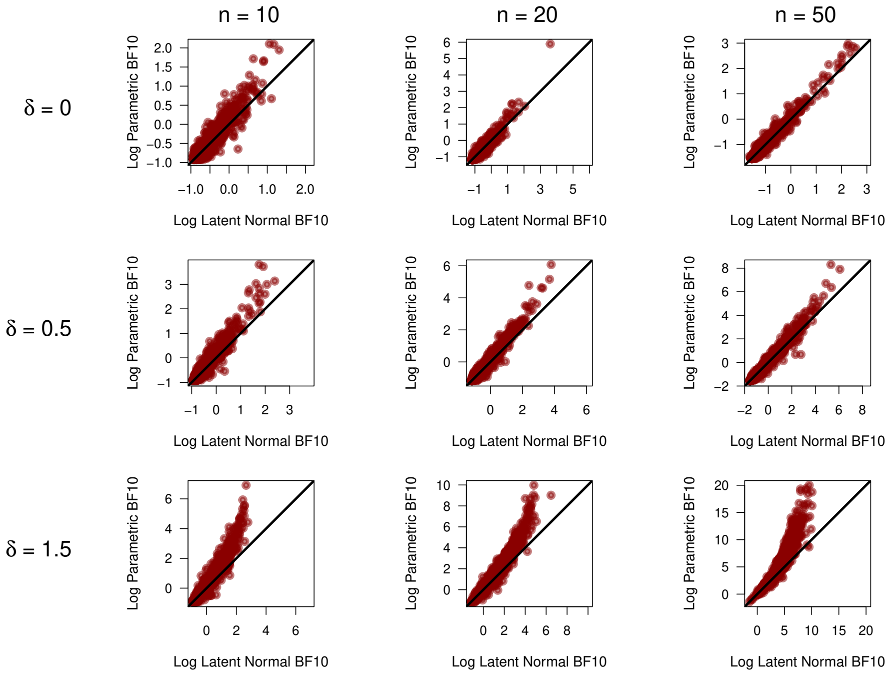


Figure 9.3: For all combinations of difference in location parameters δ, and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for logistic data. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement, as suggested by the ARE results in van der Vaart (2000).

from −0.373 to 0.120. The corresponding Bayes factor indicates that the data are about 4.694 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating moderate evidence against the hypothesis that self-reported alcohol intake differentiates between students who did and who did not pass the math exam. As a reference, the parametric t-test yields a Bayes factor of 7.138 in favor of $\mathcal{H}_1$, which is less conservative. However, due to the violated assumptions of the parametric t-test model, this result is meaningless.

9.4 Case 2: Wilcoxon Signed Rank Test

9.4.1 Background

The rank-based counterpart to the paired samples t-test was proposed by Wilcoxon (1945), who termed it the *signed rank test*. The test procedure involves taking the difference scores between the two samples under consideration and

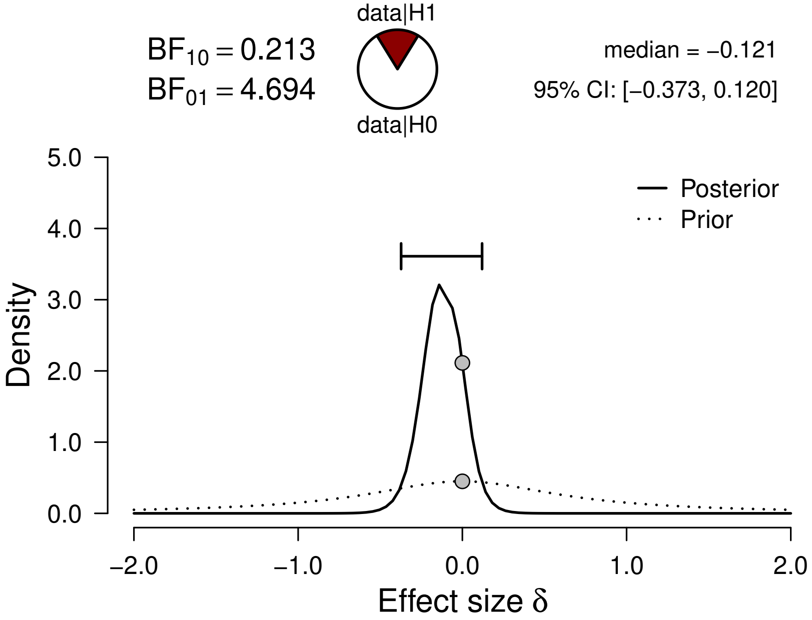


Figure 9.4: Do students who flunk a math course report drinking more alcohol? Results for the Bayesian rank sum test as applied to the data set from Cortez & Silva (2008). The dashed line indicates the Cauchy prior with scale $1/\sqrt{2}$. The solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\delta = 0$. The ratio of the ordinates gives the Bayes factor.

ranking the absolute values. The procedure may also be applied to one-sample scenarios by ranking the differences between the observed sample and the point of testing. These ranks are then multiplied by the sign of the respective difference scores and summed to produce the test statistic W. For the paired samples signed rank test, let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two data vectors each containing measurements of the same n units, and let $d = (d_1, ..., d_n)$ denote the difference scores. For the one-sample signed rank test, this process is analogous, except y is replaced by the test value. The test statistic is then defined as:

$$W = \sum_{1}^{n} [\text{rank}(|d_i|) \times Q(d_i)],$$

where Q is the sign indicator function given in (9.9).

To illustrate the procedure, consider the following hypothetical data example. Three students take a math exam, graded between 0 and 10, before and after receiving a tutoring session. Let $X = (5, 8, 4)$ be their scores on the exam before the session, and let $Y = (6, 7, 7)$ be their scores on the exam after the session. The difference scores, the ranks of the absolute difference scores, and the sign indicator function are presented in Table 9.1. In order to have a positive test statistic indicate an increase in scores, the difference scores are defined here as $(y_i - x_i)$. The test statistic W is then calculated by summing over the product of the fourth and fifth column: $1.5 - 1.5 + 3 = 3$. This value indicates a slight increase in math scores after the tutoring session.

i	$(y_i - x_i)$	d_i	rank($\|d_i\|$)	$Q(d_i)$
1	$6-5$	1	1.5	1
2	$7-8$	-1	1.5	-1
3	$7-4$	3	3	1

Table 9.1: The scores, difference scores, ranks of the absolute difference scores, and the sign indicator function Q for the hypothetical scenario where $X = (5, 8, 4)$ are the initial scores on a math exam and $Y = (6, 7, 7)$ are the scores on the exam after a tutoring session.

An often used standardized effect size for W is the matched-pairs rank-biserial correlation, denoted ρ_{mrb}, which is the correlation coefficient used as a within subjects measure of association between a nominal dichotomous variable and an ordinal variable (Cureton, 1956; Kerby, 2014). The transformation is as follows:

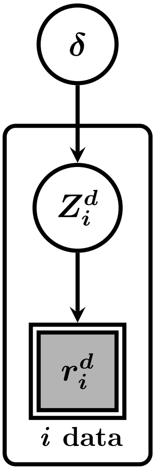
$$\rho_{mrb} = 1 - \frac{4W}{n(n+1)}. \tag{9.10}$$

The matched-pairs rank-biserial correlation can also be expressed as the difference between the proportion of data pairs where $x_i > y_i$ versus $x_i < y_i$. For the grades example, there is one pair for which $x_i > y_i$, and two pairs for which $x_i < y_i$, yielding a matched-pairs rank-biserial correlation coefficient of $\frac{2-1}{3} = \frac{2}{3}$, which is an indication that the tutoring session has increased students' math ability.

The signed rank test is similar to the sign test, where the procedure is to sum over the sign indicator function. The difference here is that the output of the sign indicator function is weighted by the ranked magnitude of the absolute differences. The signed rank test has a higher ARE than the sign test: a relative efficiency of $\frac{3}{2}$ for all distributions (Conover, 1999). For the one-sample scenario, the Pitman ARE of the signed rank test (compared to the fully parametric t-test) is similar to the ARE of the rank sum test for the unpaired two-sample scenario; for example, when the data follow a normal distribution the ARE equals $\frac{3}{\pi}$. For other distributions, especially when these are heavy-tailed, the signed rank test outperforms the t-test (Lehmann, 1999; van der Vaart, 2000).

9.4.2 Sampling Algorithm

The data augmentation algorithm is similar to that of the rank sum test and is outlined in Figure 9.5. Here d denotes the difference scores as ordinal manifestations of latent, normally distributed values Z^d. The parameter of interest is again the standardized location parameter δ, which is assigned a Cauchy prior distribution with scale parameter γ. Similar to the rank sum sampling procedure, the variance of Z^d is set to 1, as the ranked data contain no information about the



$$\begin{aligned}
\delta &\sim \text{Normal}(0, g)\\
g &\sim \text{Inverse Gamma}\left(\tfrac{1}{2}, \tfrac{\gamma^2}{2}\right)\\
Z_i^d &\sim \text{Normal}(\delta, 1)\\
r_i^d &\leftarrow \text{Rank}(|Z_i^d|) \times \text{sign}(Z_i^d)
\end{aligned}$$

Figure 9.5: The graphical model underlying the Bayesian signed rank test. The latent, continuous difference scores are denoted by Z_i^d, and their manifest signed rank values are denoted by d_i. The latent scores are assumed to follow a normal distribution governed by parameter δ. This parameter is assigned a Cauchy prior distribution, which for computational convenience is reparameterized to a normal distribution with variance g (which is then assigned an inverse gamma distribution).

variance. The computational complexity of sampling from the posterior distribution of δ is again reduced by introducing the parameter g. The Gibbs algorithm for the data augmentation and sampling δ is as follows, at sampling time point s:

1. For each value of i in $(1, \dots, n)$, sample Z_i^d from a truncated normal distribution, where the lower threshold is a_i^d given in (9.3) and the upper threshold is b_i^d given in (9.4):

$$(Z_i^d \mid z_{i'}^d, \delta) \sim \mathcal{N}_{\left(a_i^d, b_i^d\right)}(\delta, 1)$$

2. Sample δ from

$$(\delta \mid z^d, g) \sim \mathcal{N}\left(\mu_\delta, \sigma_\delta^2\right),$$

where

$$\mu_\delta = \frac{gn\overline{z^d}}{gn+1}$$

$$\sigma_\delta^2 = \frac{g}{gn+1}$$

3. Sample g from

$$(g \mid \delta) \sim \text{Inverse Gamma}\left(1, \frac{\delta^2 + \gamma^2}{2}\right),$$

where γ determines the scale (i.e., width) of the Cauchy prior on δ.

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of Z^d and δ. The posterior distribution of δ can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (9.6).

9.4.3 Simulation Study

Similar to the Wilcoxon rank sum test, a simulation study was performed to illustrate the behavior of the Bayesian signed rank test. For three values of difference in location parameters, δ (0, 0.5, 1.5), and three values of n (10, 20, 50), 1,000 data sets were generated under various distributions: skew-normal, Cauchy, logistic, and uniform distributions. In one scenario, both groups have the same distributional shape, and in a second scenario, one group follows the normal distribution and one group follows one of the aforementioned distributions. After the data were generated, the difference scores between the two groups were calculated, and used as input for the Bayesian latent normal test.

The same analyses were performed as for the Wilcoxon rank sum test. First, the relationship between the observed rank statistic W and the latent normal Bayes Factor was analyzed. Figure 9.6 illustrates this relationship, fitted with a cubic smoothing spline (J. Chambers & Hastie, 1992), when the difference scores were taken for two logistic distributions. To show results for multiple values of n in one figure, the matched-pairs rank-biserial correlation coefficient ρ_{mrb} is plotted instead of W. The Bayes factor shows a clear relationship with the rank-based test statistic, where the maximum evidence in favor of $\mathcal{H}_0$ is obtained when this statistic equals 0. Furthermore, the obtained Bayes factor grows more decisive as n increases. For other distributions of the data, highly similar results were obtained (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios).

Next to the relationship between W and the latent normal Bayes factor, the relationship between the latent normal Bayes factor and the parametric Bayes factor (Rouder et al., 2009) was analyzed. Figure 9.7 illustrates the results for all combinations of n and the difference in location parameters, δ. Note that differences in performance are due to the use of the latent normal framework and not due to the prior specification, as both the parametric and rank-based Bayes factor were based on the same Cauchy prior with scale $1/\sqrt{2}$. The two Bayes factors are generally in agreement, with the parametric Bayes factor accumulating evidence in favor of $\mathcal{H}_1$ faster when this is the true model. The latent normal Bayes factor demonstrates more instability, due to only using the ordinal information in the data. For distributions of the data that violate the assumptions of the parametric test, such as the Cauchy distribution, the parametric test greatly overestimates the variance and is no longer able to detect differences in location parameters (see the supplementary material). This misspecification does not affect the latent normal Bayes factor, underscoring its robustness.

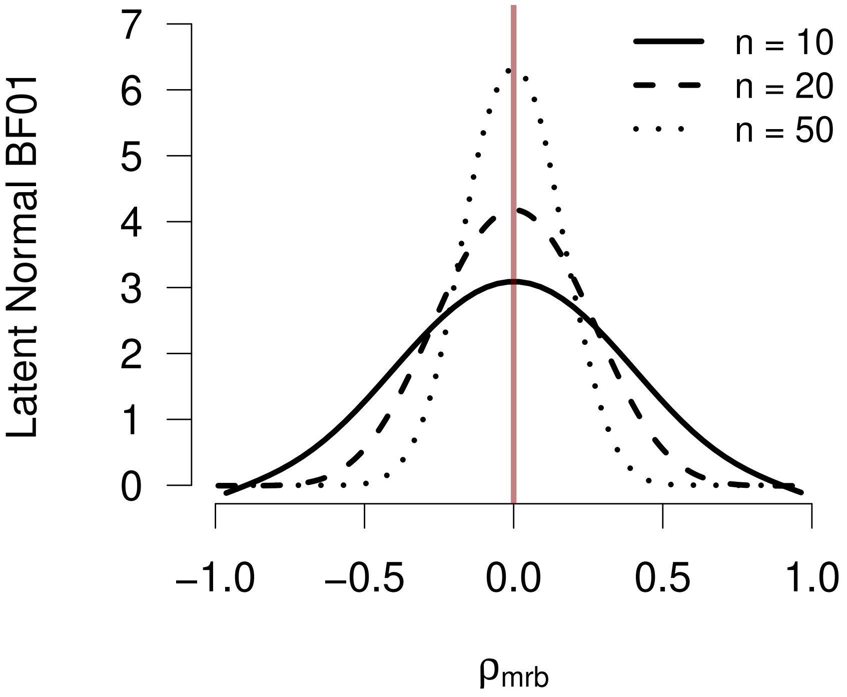


Figure 9.6: The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for logistic data. Because W is dependent on n, the matched-pairs rank-biserial correlation coefficient is plotted on the x-axis instead of W. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when $\rho_{mrb} = 0$. The further ρ_{mrb} deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines are smoothing splines fitted to the observed Bayes factors.

9.4.4 Data Example

Thall & Vail (1990) investigated a data set obtained by D. S. Salsburg concerning the effects of the drug progabide on the occurrence of epileptic seizures. During an initial eight week baseline period, the number of epileptic seizures was recorded in a sample of 31 epileptics. Next, the patients were given progabide, and the number of epileptic seizures was recorded for another eight weeks. In order to accommodate the discreteness and non-normality of the data, Thall & Vail (1990) applied a log-transformation on the counts.

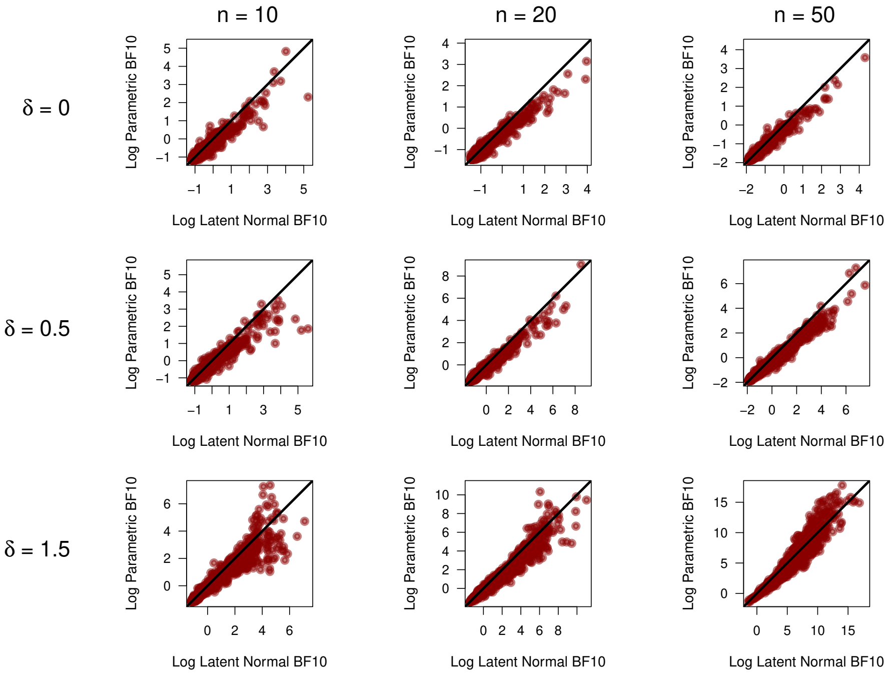


Figure 9.7: For all combinations of difference in location parameters δ, and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for logistic data. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement, with the latent normal Bayes factor accumulating evidence in favor of the true model faster.

This log-transformation has a clear impact on the outcome of a parametric Bayesian t-test (Morey & Rouder, 2018): $\text{BF}_{10} \approx 0.2$ for the raw data, whereas $\text{BF}_{10} \approx 2.95$ for the log-transformed data. Here we analyze the data with the signed rank test; because this test is invariant under monotonic transformations, the same inference will result regardless of whether or not the data are log-transformed.

The hypothesis specification here is similar to that of the setup of the rank sum example: $\mathcal{H}_0 : \delta = 0$ which is pitted against $\mathcal{H}_1 : \delta \neq 0$ and prior $1/\sqrt{2}$, that is, $\delta \sim \text{Cauchy}\left(0, \frac{1}{\sqrt{2}}\right)$. Figure 9.8 shows the resulting posterior distribution for δ under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for δ equals 0.207, with a 95% credible interval that ranges from −0.138 to 0.549. The corresponding Bayes factor indicates that the data are about 2.513 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating that, for the purpose of discriminating $\mathcal{H}_0$ from $\mathcal{H}_1$, the data are almost perfectly uninformative.

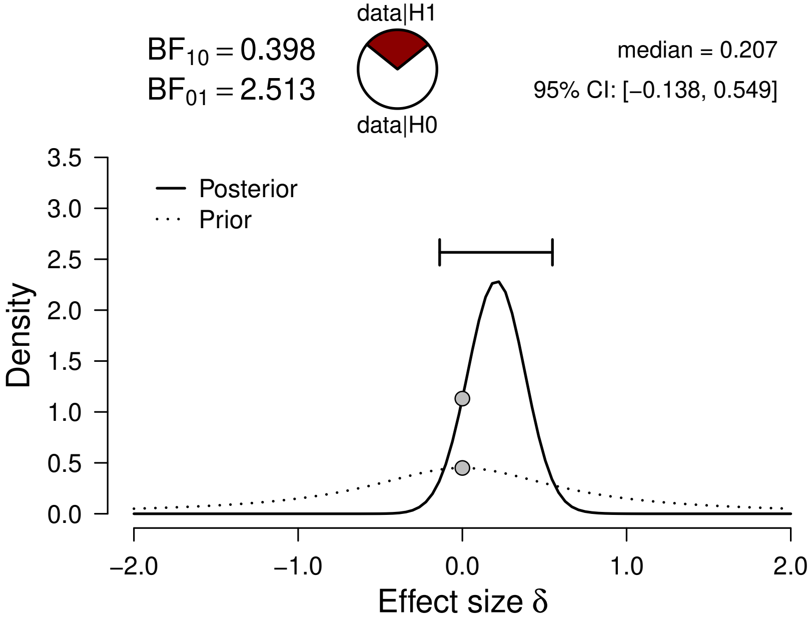


Figure 9.8: Does progabide reduce the frequency of epileptic seizures? Results for the Bayesian signed rank test as applied to the data set presented in Thall & Vail (1990). The dashed line indicates the Cauchy prior with scale $1/\sqrt{2}$. The solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\delta = 0$. The ratio of the ordinates gives the Bayes factor.

9.5 Case 3: Spearman's ρ_s

9.5.1 Background

Spearman (1904) introduced the rank correlation coefficient ρ in order to overcome the main shortcoming of Pearson's product moment correlation, namely its inability to capture monotonic but non-linear associations between variables. Spearman's method first applies the rank transformation on the data and then computes the product-moment correlation on the ranks. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two data vectors each containing measurements of the same n units, and let $r^x = (r_1^x, ..., r_n^x)$ and $r^y = (r_1^y, ..., r_n^y)$ denote their rank-transformed values, where each value is assigned a ranking within its variable. This then leads to the following formula for Spearman's ρ_s:

$$\rho_s = \frac{\text{Cov}_{r^x r^y}}{\sigma_{r^x}\sigma_{r^y}}.$$

The Pitman ARE of Spearman's ρ compared to parametric Pearson's ρ displays a similar pattern to the ARE's discussed before. When the data follow a bivariate normal distribution, the ARE equals $9/\pi^2$ (Hotelling & Pabst, 1936). Thus, under optimal conditions for the parametric test, it is marginally more efficienct compared to Spearman's ρ. As the data depart from normality, the rank-based test outperforms its parametric counterpart.

9.5.2 Sampling Algorithm

The graphical model in Figure 9.9 illustrates the data augmentation setup for inference on the latent correlation parameter ρ. The sampling method is a Metropolis-within-Gibbs algorithm, where data augmentation is conducted with a Gibbs sampling algorithm as before, but combined with a random walk Metropolis-Hastings sampling algorithm (Metropolis et al., 1953; Hastings, 1970) to sample from the posterior distribution of ρ (see also van Doorn et al., 2019).

The sampling algorithm for the latent correlation is as follows, at sampling time point s:

1. For each i in $(1, \ldots, n_x)$, sample Z_i^x from a truncated normal distribution, where the lower threshold is a_i^x given in (9.3) and the upper threshold is b_i^x given in (9.4):

$$(Z_i^x \mid z_{i'}^x, z_i^y, \rho_{z^x, z^y}) \sim \mathcal{N}_{(a_i^x,\, b_i^x)}\left(z_i^y \rho_{z^x,\, z^y},\, \sqrt{1 - \rho_{z^x, z^y}^2}\right)$$

2. For each i in $(1, \ldots, n_y)$, the sampling procedure for Z_i^y is analogous to step 1.

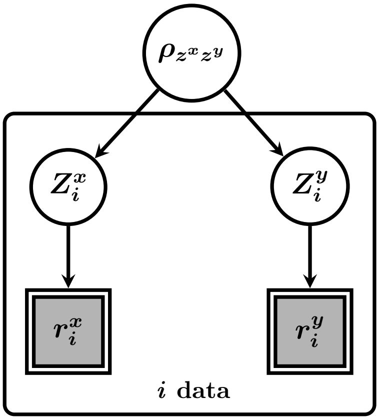
3. Sample a new proposal for $\rho_{z^x,\, z^y}$, denoted ρ^*, from the asymptotic normal approximation to the sampling distribution of Fisher's z-transform of ρ (Fisher, 1915):

$$\tanh^{-1}(\rho^*) \sim \mathcal{N}\left(\tanh^{-1}(\rho^{s-1}), \frac{1}{\sqrt{(n-3)}}\right).$$

   The acceptance rate α is determined by the likelihood ratio of $(z^x, z^y|\rho^*)$ and $(z^x, z^y \mid \rho^{s-1})$, where each likelihood is determined by the centered bivariate normal density with unit variances, and correlation ρ:

$$\alpha = \min\left(1, \frac{P(z^x, z^y \mid \rho^*)}{P(z^x, z^y \mid \rho^{s-1})}\right).$$

Repeating the algorithm a sufficient number of times yields samples from the posterior distributions of z^x, z^y, and $\rho_{z^x,\, z^y}$.



$$\begin{aligned}
Z_i^x &\sim \text{Normal}(0,1)\\
Z_i^y &\sim \text{Normal}(0,1)\\
\rho_{Z^x Z^y} &\sim \text{Uniform}(-1,1)\\
r_i^x &\leftarrow \text{Rank}(Z_i^x)\\
r_i^y &\leftarrow \text{Rank}(Z_i^y)
\end{aligned}$$

Figure 9.9: The graphical model underlying the Bayesian test for Spearman's ρ_s. The latent, continuous scores are denoted by Z_i^x and Z_i^y, and their manifest rank values are denoted by r_i^x and r_j^y. The latent scores are assumed to follow a normal distribution governed by parameter ρ (which is assigned a uniform prior distribution).

9.5.3 Transforming Parameters

The transition from Pearson's ρ to Spearman's ρ_s can be made using a statistical relation described in Kruskal (1958). This relation, defined as

$$\rho_s = \frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right).$$

enables the transformation of Pearson's ρ to Spearman's ρ_s when the data follow a bivariate normal distribution. Since the latent data are assumed to be normally distributed, this means that the posterior samples for Pearson's ρ can be easily transformed to posterior samples for Spearman's ρ_s. The posterior distribution of ρ_s can then be used to obtain a Bayes factor through the Savage-Dickey density ratio given in (9.6).

9.5.4 Simulation Study

Similar to the previous tests, the behavior of the latent normal correlation test was assessed with a simulation study. For four values of Spearman's ρ_s (0, 0.3, 0.8) and three values of n (10, 20, 50), 1,000 data sets were generated under four copula models: Clayton, Gumbel, Frank, and Gaussian (Sklar, 1959; Nelsen, 2006; Genest & Favre, 2007; Colonius, 2016). Using Sklar's theorem, copula models decompose a joint distribution into univariate marginal distributions and a dependence structure (i.e., the copula). This decomposition enables the generation of data for specific values of Spearman's ρ_s. Furthermore, the copula is independent of the marginal distributions of the data and can therefore encompass a wide range of distributions.

Similar to the previous tests, the relationship between the latent normal Bayes factor and the observed rank-based statistic was analyzed. Figure 9.10 illustrates this relationship, fitted with a cubic smoothing spline (J. Chambers & Hastie, 1992), for various values of n, for data generated with the Clayton copula. The relationship is similar to those shown for the previous tests: maximum evidence in favor of $\mathcal{H}_0$ is attained when the observed Spearman's ρ_s equals 0. The further the observed test statistic deviates from 0, the more evidence is accumulated in favor of $\mathcal{H}_1$. Furthermore, the obtained Bayes factor grows more decisive as n increases. Highly similar results were obtained for the other copulas that were considered (see the online supplementary material at https://osf.io/gny35/ for the results of these scenarios).

Secondly, the relationship between the latent normal Bayes factor and the parametric Bayes factor (Ly, Marsman, & Wagenmakers, 2018) for testing correlations was analyzed. For both Bayes factors, a uniform prior between -1 and 1 was used, such that differences in performance are due to the use of the latent normal framework and not due to the prior. Figure 9.11 shows the results for all combinations of n and ρ that were used, for the Clayton copula. The two Bayes factors are generally in agreement. An important remark here is that the marginal distributions of the data are not taken into account. The data generated with the copula method are located on the unit square, and if so desired, can then be transformed with the inverse cdf to follow any desired distribution. These transformations are monotonic, and therefore do not affect the rank-based Bayes factor, whereas the parametric Bayes factor can be heavily affected by this. This underscores an important property of the rank-based Bayes factor: it solely depends on the copula (i.e., the only component of the data that pertains to the dependence structure), and not on the marginal distribution of the data.

9.5.5 Data Example

We return to the data set from Cortez & Silva (2008) and examine the possibility that math grades (ranging from 0 to 20) are associated with the quality of family relations (self-reported on a Likert scale that ranges from $1-5$). The hypotheses are $\mathcal{H}_0 : \rho = 0$ which is pitted against $\mathcal{H}_1 : \rho \neq 0$. For the Bayes factor we use the uniform prior, that is, $\rho \sim \text{Uniform}[-1, 1]$. Thus, the null hypothesis specifies the lack of an association between the two variables and the alternative hypothesis assigns the degree of association a uniform prior distribution (e.g., Jeffreys (1961)). The parametric correlation test (Ly, Marsman, & Wagenmakers, 2018) yields a Bayes factor of 9.467, but since the data are ordinal measures and not normally distributed, the parametric correlation model is severely misspecified. Thus, conducting the rank-based analysis is more applicable and prudent here.

Figure 9.12 shows the resulting posterior distribution for ρ_s under $\mathcal{H}_1$ and the associated Bayes factor. The posterior median for ρ_s equals 0.059, with a 95% credible interval that ranges from -0.052 to 0.161. The corresponding Bayes factor indicates that the data are about 7.915 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, indicating moderate evidence against an association between math performance and the quality of family ties.

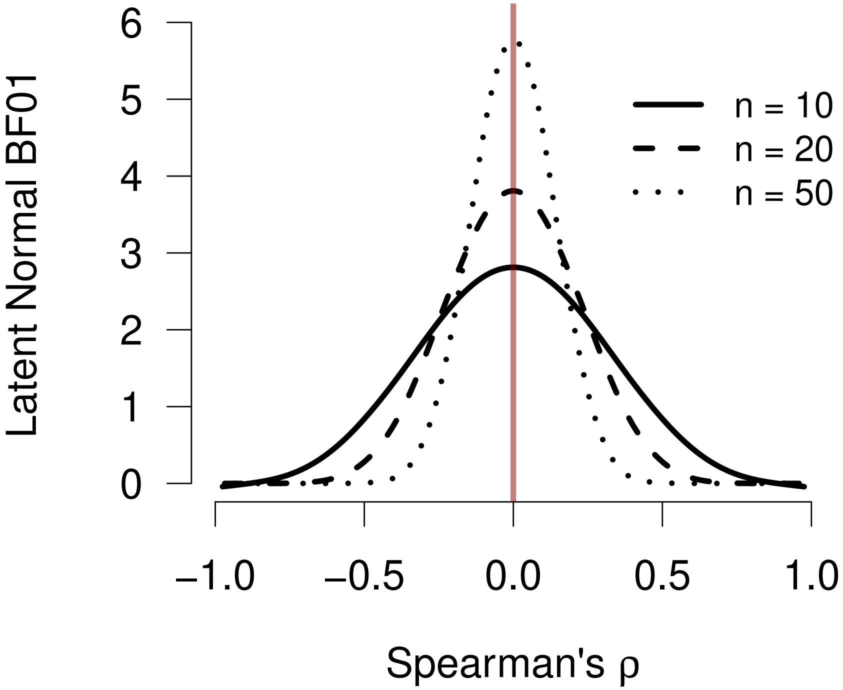


Figure 9.10: The relationship between the latent normal Bayes factor and the observed rank-based test statistic is illustrated for data generated with the Clayton copula. The relationship is clearly defined, and maximum evidence in favor of $\mathcal{H}_0$ is attained when Spearman's $\rho_s = 0$. The further Spearman's ρ_s deviates from 0, the stronger the evidence in favor of $\mathcal{H}_1$ becomes. The lines are smoothing splines fitted to the observed Bayes factors.

9.6 Concluding Comments

This chapter outlined a general methodology for applying conventional Bayesian inference procedures to ordinal data problems. Latent normal distributions are assumed to generate impoverished rank-based observations, and inference is done on the model parameters that govern the latent normal level. This idea, first proposed by Pearson (1900), yields all the advantages of ordinal inference including robustness to outliers and invariance to monotonic transformations. Moreover, the methodology also handles ties in a natural fashion, which is important for coarse data such as provided by popular Likert scales. Furthermore,

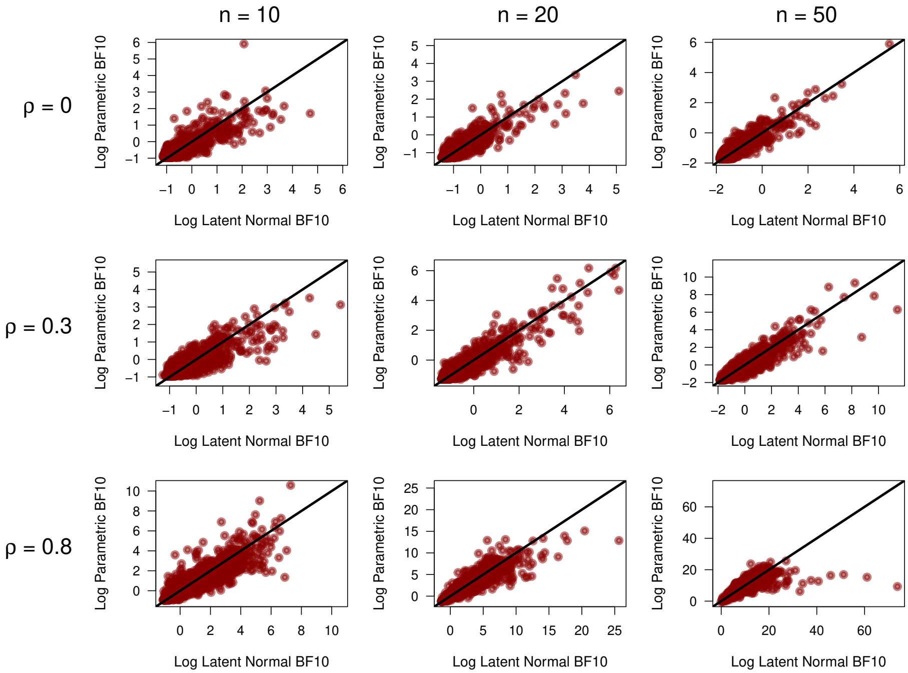


Figure 9.11: For all combinations of Spearman's ρ_s and n, the relationship between the latent normal Bayes factor and the parametric Bayes factor is shown for data generated with the Clayton copula. The black lines indicate the point of equivalence. The two Bayes factors are generally in agreement.

the robustness of the latent normal method is underscored by the simulation studies performed for each test. These results illustrate that the method provides accurate inference, even if the data are not normally distributed.

By postulating a latent normal level for the observed rank data, the advantages of ordinal inference can be combined with the advantages of Bayesian inference such as the ability to update uncertainty as the data accumulate, the ability to quantify evidence in favor of either hypothesis being tested, and the ability to incorporate prior information. It should be stressed that, even though our examples used default prior distributions, the proposed methodology is entirely general in the sense that it also applies to informed or subjective prior distributions (Gronau et al., 2018).

For computational convenience and ease of interpretation, our framework used latent normal distributions. This is not a principled limitation, however, and the methodology would work for other families of latent distributions as well (e.g., Albert (1992b)).

In sum, we have presented a general methodology to conduct Bayesian infer-

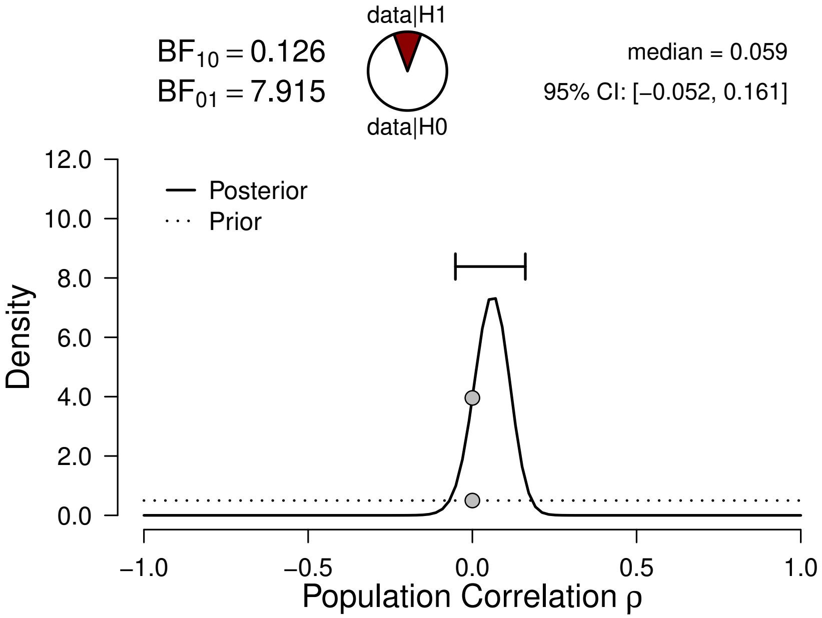


Figure 9.12: Is performance on a math exam associated with the quality of family relations? Results for the Bayesian version of Spearman's ρ_s as applied to the data set from Cortez & Silva (2008). The dashed line indicates the uniform prior distribution, and the solid line indicates the posterior distribution. The two grey dots indicate the prior and posterior ordinate at the point under test, in this case $\rho = 0$. The ratio of the ordinates gives the Bayes factor.

ence for ordinal problems, and illustrated its potential by developing Bayesian counterparts to three popular ordinal tests: the rank sum test, the signed rank test, and Spearman's ρ_s. Supplementary material, including simulation study results, R-code for each method and the example data used, is available at https:https://osf.io/gny35/. In the near future we intend to make these tests available in the open-source software package JASP (e.g., JASP Team (2020); jasp-stats.org), which we hope will further increase the possibility that the tests are used to analyze ordinal data sets for which the traditional parametric approach is questionable.

Chapter 10

Using the Weighted Kendall's Distance to Analyze Rank Data in Psychology

Abstract

Although Kendall's distance is a standard metric in computer science, it is less widely used in psychology. We demonstrate the usefulness of Kendall's distance for analyzing psychological data that take the form of ranks, lists, or orders of items. We focus on extensions of the metric that allow for heterogeneity of item importance, item position, and item similarity, as well showing how the metric can accommodate missingness in the form of top-k lists. To demonstrate how Kendall's distance can help address research questions in psychology, we present four applications to previous data. These applications involve the recall of events on September 11, people's preference rankings for the months of the year, people's free recall of animal names in a clinical setting, and expert predictions involving American football outcomes.

10.1 Introduction

Rank data are ubiquitous in psychological science. Any task that involves sequences of behavior, such as recalling items from memory or solving a problem through a series of decisions and actions, yields rank order data. Other common examples include Likert scale measures, consumer choice preferences, and psychophysical judgments.

An often-used statistical tool to analyze rank data is a rank correlation, such as Kendall's τ (Kendall, 1938) or Spearman's ρ (Spearman, 1904). The goal of these methods is to quantify the strength of a monotonic relation between two variables, without assuming this relation to be linear. The rank correlation coefficient is then frequently used to test hypotheses related to the presence or absence of such a relation. However, such a procedure often overlooks the wealth of information embedded in the value of the rank correlation coefficient. In computer science, for instance, rank correlations are a popular metric for aggregating

This chapter has been submitted for publication as van Doorn, J.B., Westfall, H., & Lee, M.D. (2020). Using the Weighted Kendall's Distance to Analyze Psychological Data and Models.

search engine results, fighting spam, and word association (Beg & Ahmad, 2003). Whereas psychological science predominantly uses the rank *correlation* to test for an association between two variables, the field of computer science focuses on the (non-standardized) rank *distance* to quantify degrees of similarity between two or more observed sequences of data points. In doing so, the distance metric becomes a function of the data that can, in turn, be used for further quantitative analysis.

In this chapter, we aim to bridge the gap between developments in computer science and psychological science by underscoring Kendall's distance metric as a useful tool for analyzing psychological data . First, we outline the basic distance metric, which has sometimes been used in psychology (e.g., Lee et al., 2014; Brandt et al., 2016; Selker et al., 2017). Second, we introduce three extensions that enable the weighting of item importance, item position, and item similarity, which are rarely used in psychology. Third, we illustrate how Kendall's distance can be modified to accommodate missingness in the data in the form of top-k lists. Each extension is first illustrated using a toy example, and then demonstrated more fully in practical applications to existing data sets in psychology previously collected to address specific research questions. In order to increase the ease of application of the discussed algorithms, we include a plug and play R-script, available at https://osf.io/6k9t8/

10.2 Kendall's Distance

Introduced by Kendall (1938), the Kendall's distance metric, often written as τ, is a popular rank-based coefficient for comparing two vectors of data points. It is based on the number of adjacent pairwise swaps required to transform one vector into the other.

In order to introduce the notation and computation of Kendall's distance and its extensions, we use a small toy example where two people are asked to rank $n = 4$ sodas—Coke, Pepsi, Sprite, and Fanta—in terms of tastiness. Let the ranking of person A be $A = (Coke, Pepsi, Fanta, Sprite)$, and the ranking of person B be $B = (Pepsi, Coke, Sprite, Fanta)$. We denote the ith item of A with A_i, such that $A_i = Pepsi$ when $i = 1$. Next, we denote the ranking of item c with $\sigma_A(c)$ for person A, and $\sigma_B(c)$ for person B. For instance, $\sigma_A(c) = 1$ and $\sigma_B(c) = 2$ for $c = Coke$. Combining these two notations allows us to denote the rank of the ith item in A, for person B. For instance, $\sigma_B(A_i) = 2$ when $i = 1$, because the first item in A (i.e., Coke) is ranked second by person B.

With these definitions in hand, we can compute Kendall's distance between person A and B. In order to sort B in such a way that it is identical to A, we need to swap *Coke* and *Pepsi*, and then *Fanta* and *Sprite*. In this example, Kendall's distance is therefore equal to 2. As a consequence, Kendall's distance is often referred to as the bubble sort distance (Shaw & Trimble, 1963).[1] Table 10.1 provides an illustration of this sorting procedure.

[1] See also https://www.youtube.com/watch?v=lyZQPjUT5B4 and https://www.geeksforgeeks.org/bubble-sort/ for accessible introductions.

In order to obtain the correlation coefficient, the distance is then standardized to be in the interval $[-1, 1]$, however, we focus on the distance in this chapter. The minimum value for the distance is 0, indicating perfect correspondence, and the maximum value for the distance is equal to $\frac{n(n-1)}{2}$, where n is the length of A and B.

A	B	B^1	B^2
Coke	Pepsi	Coke	Coke
Pepsi	Coke	Pepsi	Pepsi
Fanta	Sprite	Sprite	Fanta
Sprite	Fanta	Fanta	Sprite

Table 10.1: The two vectors A and B, and the adjacent pairwise swaps needed to transform B into A: B^1 denotes B after one swap, and B^2 denotes B after two swaps. Therefore, Kendall's distance between A and B equals 2.

Another way of calculating Kendall's distance is by comparing the ranks of items A_i and A_j in the vector B, for $i < j$. If item $\sigma_B(A_i)$ is greater than $\sigma_B(A_j)$, this means that person B ranked items A_i and A_j in the reverse order compared to person A. We refer to this as an *inversion*. A formal definition is given by the formula:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} \left[\sigma_B(A_i) > \sigma_B(A_j)\right], \tag{10.1}$$

which counts the number of pairwise inversions.

We now consider four extensions of Kendall's distance that we think can be especially useful for analyzing psychological data.

10.2.1 Item Weights

As presented by Kumar & Vassilvitskii (2010), item-specific weights may be incorporated in the distance metric. In the basic definition, the cost of swapping two items is set to 1, such that swapping two items adds 1 to the metric. However, it could be the case that some items contribute more, or less, to the dissimilarity between the soda preferences of person A and B. For instance, we could theorize that disagreement in taste is more prolific in the ranking of Fanta than for other sodas. For instance, the marketing team of Fanta may want two people who rank Fanta differently to be recognized as being more dissimilar than two people who rank Coke differently. In such cases, we can introduce the item specific weights w, where w_i denotes the cost of performing a swap that contains item A_i. This enables us to model different items as contributing more, or less, to differences between the rankings represented by the vectors A and B.

Formally, the extension to include item importance is given by the formula:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} w_i w_j \left[\sigma_B(A_i) > \sigma_B(A_j)\right]. \tag{10.2}$$

10.2.2 Position Weights

Another extension focuses on weighting different positions in a ranking, rather than different items. This can be achieved by making the cost of performing a swap dependent on the position i on which an inversion occurs (Kumar & Vassilvitskii, 2010). We can imagine a situation in which one's favorite soda is more important in determining taste preference than one's least favorite soda: if an inversion occurs early in B, this should lead to a greater value of Kendall's distance than if an inversion occurs at the end of B.

In order to assign these weights, we first define p:

$$p_i = p_{i-1} + \delta_i,$$

where $p_1 = 1$. The position weight δ_i denotes the cost of a pairwise swap of an item in the ith position. It therefore represents the importance of that position, relative to the first position. This weight can either be assigned arbitrarily, or through a specific algorithm. One popular method is called discounted cumulative gain (DCG; Järvelin & Kekäläinen, 2002), where the weights are calculated as a logarithmic function of the item positions:

$$\delta_i = \frac{1}{\log(i+1)} - \frac{1}{\log(i+2)}.$$

The intuition behind the DCG weighting is that the item at position i is about twice as important in determining the dissimilarity between A and B than the item at position $i-1$, so that when the item is sorted, its swaps have a lower cost. For an illustration of this, see Table 10.2.

i	A	B	δ	p
1	Coke	Pepsi	-	1
2	Pepsi	Coke	0.189	1.189
3	Fanta	Sprite	0.1	1.289
4	Sprite	Fanta	0.063	1.352

Table 10.2: Values of the position weights δ and p_i for the soda example, with δ calculated using the DCG algorithm.

With δ and p defined, we can now calculate the average cost of moving item i in B to the position of that item in A, remembering that this can involve multiple pairwise swaps. This average cost is: $\bar{p_i} = \frac{p_i - p_{\sigma_B(A_i)}}{i - \sigma_B(A_i)}$. For instance, the cost of moving item Coke from position 2 to position 1 in B is calculated as $\bar{p_1} = \frac{p_1 - p_{\sigma_B(Coke)}}{1 - \sigma_{B(Coke)}} = \frac{p_1 - p_2}{1-2} = \frac{1-1.189}{1-2} = 0.189$.

The general incorporation of position weights is provided by the formula:

$$\tau = \sum_{1 \le i < j \le n}^{n} \bar{p_i}\bar{p_j}\left[\sigma_B(A_i) > \sigma_B(A_j)\right]. \tag{10.3}$$

10.2.3 Similarity Weights

The third extension of Kendall's distance takes into account the similarities and differences between items. This means that, when two items are considered highly similar, the cost of swapping these two items is lower than the cost of swapping two items that are considered to be more different from one another. In our sodas example, this can be used to model the high similarity between Coke and Pepsi.[2] As such, the inversion of Coke and Pepsi has a lower cost than the inversion of Fanta and Sprite.

In order to incorporate item similarities, we define the distance matrix D, where element D_{ij} determines how similar items A_i and A_j are. When this is set to 0, items A_i and A_j are identical; as the values are set to the large values, the item pairs become more different.

The distance matrix is incorporated as follows in the formula for Kendall's distance:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} D_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right]. \tag{10.4}$$

10.2.4 Top-*k* Lists

Lastly, we discuss comparing top-k lists. When comparing two lists of k items, it may be the case that not all items appear on both lists. It could be that there is no predetermined set of items to rank. For example, instead of asking person A and B to rank four sodas, they could have been asked to list their top 4 favorite sodas. It could also be the case that one ranking contains missing information. For example, even if the same four sodas are being ranked, one person might only list their top three. This sets the current extension apart from the previous three extensions: whereas the other extensions are modeling choices, the top k extension is driven by the nature of the data.

Suppose that we observe the responses $A = (Coke, Pepsi, Fanta, Sprite)$ and $B = (7up, Sprite, GingerAle, Pepsi)$. In such a case, we cannot determine for all items if an inversion has occurred due to some items only appearing in one of the lists. A method introduced by Fagin, Kumar, & Sivakumar (2003) can be used to model the missingness of items.

The approach identifies four cases of how two items may appear in A and B, and outlines the cost of a swap z. We present these four cases for the toy example:

1. Both items appear in A and B (e.g., Coke and Pepsi). Since person A prefers Coke and person B prefers Pepsi, this is the traditional case of an inversion and therefore $z = 1$.

2. Both items appear in A, but only one item appears in B (e.g., Pepsi and Fanta). Since person B only includes Pepsi, we can conclude that they prefer Pepsi over Fanta. If person A shares this preference, $z = 0$. Otherwise, this is an inversion and therefore $z = 1$.

[2] The authors acknowledge that some readers might wildly disagree with this statement.

3. One item appears only in A, and the other item appears only in B (e.g., Coke and 7up). In a similar reasoning to the previous case, we know that person A prefers Coke over 7up and person B prefers 7up over Coke, because at least those sodas appear in the list. This is an inversion and we therefore set $z = 1$.

4. Both items appear in A, but neither appear in B (e.g., Coke and Fanta). Here there is no information on whether person B prefers Coke or Fanta, since neither appear in B. As a first option, Fagin et al. (2003) outline the optimistic approach, which is setting $z = 0$. In other words, this gives person B the "benefit of the doubt", and assumes that if they had included Coke and Fanta, they would have expressed the same preference as person A. Alternatively, the pessimistic approach sets $z = 1$, and assumes person B would have expressed the opposite ordering of the items. In general, the probability of person B preferring Coke over Fanta can be represented by parameter $0 \leq \theta \leq 1$, with $\theta = 0$ corresponding to the optimistic approach and $\theta = 1$ corresponding to the pessimistic approach. As such, specifying $\theta = \frac{1}{2}$ corresponds to a neutral approach, in which there is an equal probability for person B expressing the same or reverse order for the items. This still takes into account the missingness, while not making a statement about how the items would be ranked if they would be included in B.

Adding this extension to the Kendall's distance formula gives:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} \theta_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right]. \tag{10.5}$$

All of the extensions presented above can be combined to form the weighted partial Kendall's distance:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} w_i w_j \bar{p_i} \bar{p_j} D_{ij} \theta_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right]. \tag{10.6}$$

10.3 Applications

We have now defined the full metric that is capable of modeling item importance, item position, and item similarity, while also accommodating missingness in top-k lists. In this section, we present a series of four applications of Kendall's distance to previous psychological data, demonstrating how the various extensions can improve data analysis to address the motivating research questions.

10.3.1 Item Weights: Recall of Events on September 11

In order to study memory reconstruction, Altmann (2003) considered six events that occurred on September 11, 2001. The events, in their true temporal order, were (1) One plane hits the World Trade Center, (2) A second plane hits the World

Trade Center, (3) One Plane crashes into the Pentagon, (4) One tower at the World Trade Center collapses, (5) One Plane crashes in Pennsylvania, and (6) A second tower at the World Trade Center collapses.

The participant responses consist of individual's recalled temporal orderings of these events. Kendall's distance provides a natural single measure of response accuracy for each participant. However, as noted by Altmann (2003), the correct ranking of some of these events need not be driven by memory, but can be determined by logic. For example, it can be deduced that the planes hitting the World Trade Center occurs before the tower collapsing, and that the first plane hits before the second plane. In contrast, correctly recalling when the plane crash in Pennsylvania occurred needs to be memory driven. Thus, when a participant incorrectly orders the two planes hitting the towers, this can be due to poor memory or poor reasoning, while incorrectly ranking the Pennsylvania crash is more likely due to poor memory.

These considerations mean that if the research goal is to study memory ability in recall, rather than logic reasoning skill, events (3) and (5) should be weighted more heavily than items (1), (2), and (4). For example, if we consider the responses from two specific participants in the Altmann (2003) data, who recalled orders:

(A) Plane 2, Pentagon, Plane 1, Tower 1, Pennsylvania, Tower 2

(B) Plane 1, Pentagon, Plane 2, Pennsylvania, Tower 1, Tower 2

Both of these participants have the same number of inversions relative to the ground truth, and therefore yield an identical unweighted Kendall's distance of 2. However, participant A makes logical errors while participant B does not. As a consequence, assigning a weight of 2 to the memory driven items, and a weight of 1/2 to the logic driven items, changes the accuracy measures to 1.25 for participant A and 3 for participant B.

Figure 10.1 shows the change in Kendall's distance resulting from including item weights for all 158 participants from Altmann (2003). The standard unweighted measure is shown on the left, and the item-weighted measure is shown on the right, with lines connecting the same participant under each measure. It is clear that the recall accuracy of participants can increase, decrease, or stay the same once item weights are incorporated. It is also clear that the use of item weights also gives the Kendall's distance greater resolution as a measure of accuracy. Without weight, there theoretically are 15 possible outcomes for Kendall's distance, 9 of which are observed in the Altmann (2003) data. With item weighting there are 61 possible outcomes, including fractional counts, 21 of which are observed.

10.3.2 Position Weights: Month Preference

In the previous example, the participant responses were compared to a true ranking, in order to determine their accuracy. However, participants' responses can also be compared to each other, in order to determine similar response patterns.

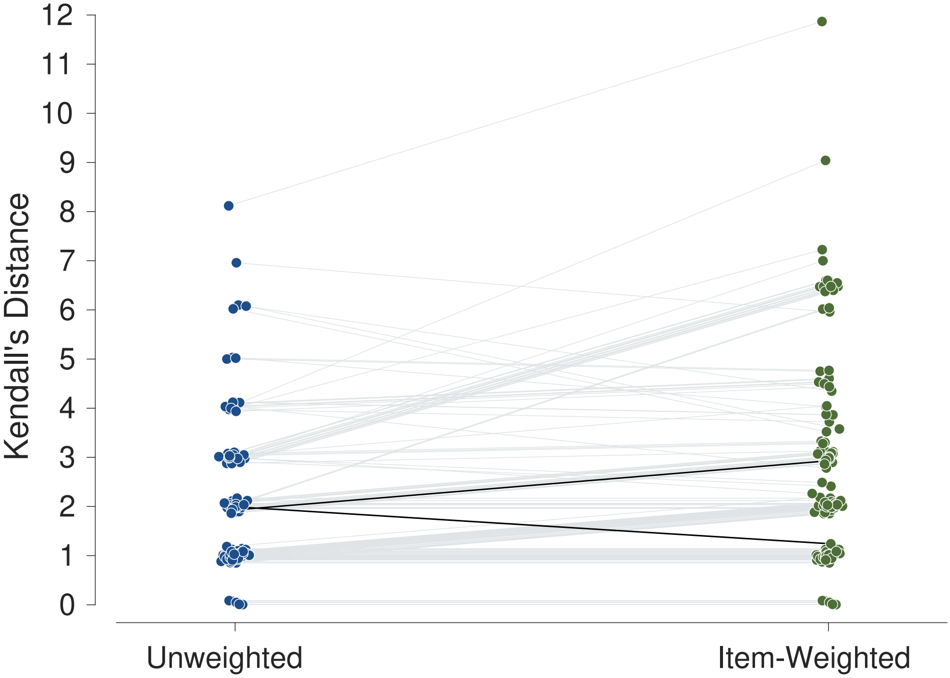


Figure 10.1: Unweighted and item-weighted Kendall's distance for 158 participants from the Altmann (2003) study of memory for the order of events on September 11. Each point in the unweighted column and item-weighted column corresponds to a participant, jittered around the Kendall's distance measure. The same participant for each measure is connected by a gray line. Participants A and B are highlighted by black lines. As a result of the weighting, Participant A sees an increase of Kendall's distance, whereas Participant B sees a decrease of Kendall's distance.

Accordingly, our second application involves people's preferences for the months of the year, as collected by the crowd-source opinion web site `ranker.com`. A total of 16 people ranked the 12 months from best worst.

A natural research question addressed by these data is whether there are individual differences in people's preference patterns. For instance, some people prefer the winter months to the summer months, while others may prefer summer to winter. One exploratory approach to identifying such patterns is through data visualization. We rely on multidimensional scaling (MDS: Borg & Groenen, 1997) algorithm to the pairwise Kendall's distances between people, using spaces of just two dimensions. This allows for a simple visualization that may reveal clusters of people based on the similarity of their preferences (i.e., groups of participants whose Kendall's distance scores are small with respect to each other).

There are two extensions of Kendall's distance that are potentially useful here. First, we can model the adjacent months as being fairly similar to each other. We can therefore reduce the cost of swapping, for instance, January and February from 1 to 0.5. Secondly, we can use position weights to capture assumptions about whether people's most or least favorite months are more indicative of their preference. For example, consider the rankings provided two `ranker.com` users:

(A) Dec, Jun, Oct, May, Jul, Nov, Aug, Apr, Sep, Mar, Feb, Jan

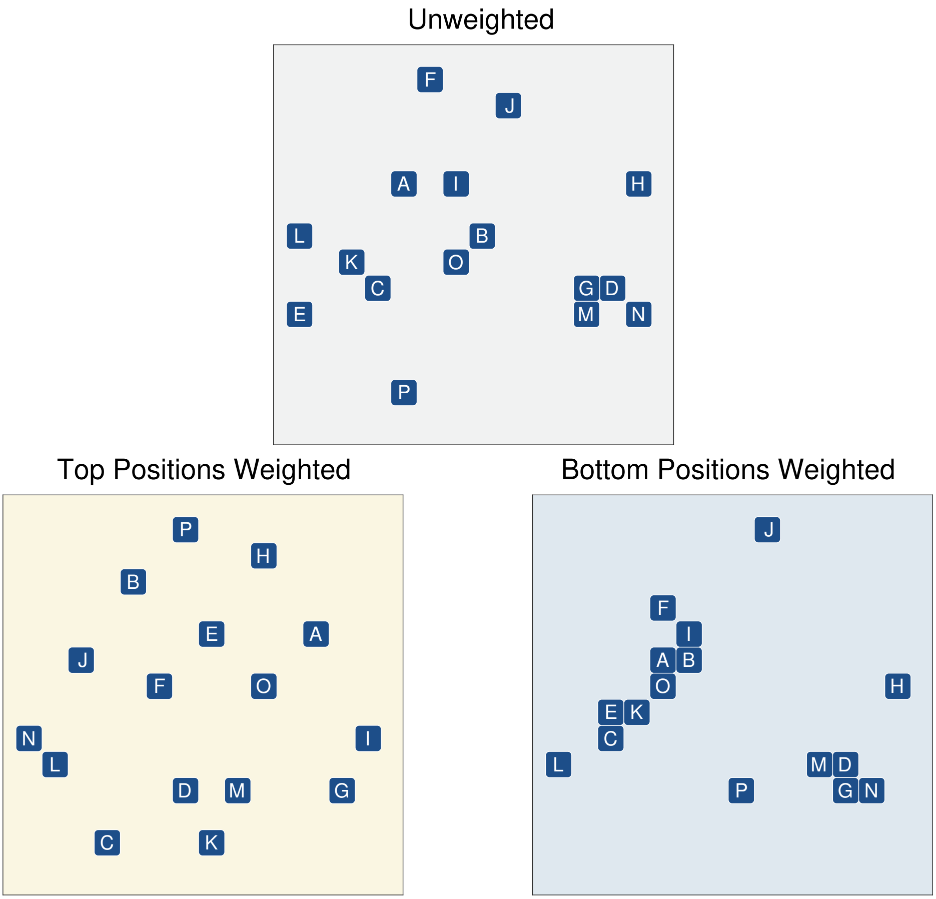
(B) May, Oct, Nov, Dec, Sep, Jun, Jul, Apr, Aug, Mar, Feb, Jan

Their favorite months are rather different, but their least favorite months are very similar. Whether these people are regarded as having similar preferences depends on the weighting given to their favorite vs least favorite months.

Figure 10.2 presents the MDS visualizations for all of the `ranker.com` people, considering three scenarios. The top panel shows the MDS visualization for the preference rankings that are weighted by similarity but are unweighted by position. The lower-left panel shows the MDS visualization for the preference rankings where the DCG algorithm was used to weight the best months more heavily. The bottom-right panel shows the MDS visualization for the preference rankings where the reverse DCG algorithm is applied, in order to weigh the worst months more heavily.

In this way, the difference between the bottom-left and bottom-right visualizations is based on whether the most favored or least favored months are treated as the most important in determining the similarity between people's preferences. Accordingly, in terms of the specific examples presented earlier, person A and person B are further apart in the top and bottom-left panels of Figure 10.2 than they are when the weighting is changed to emphasize the least favorite months, as in the bottom-right panel.

It is striking that the MDS visualization based on weighting the least favorite months, shown in the bottom-right panel, reveals a clear cluster structure. There is a divide between people who dislike the cold winter months, in the left half of the plot, and people who dislike the hot summer months, in the right half of the plot. The other visualizations lack this clear cluster structure, suggesting



Unweighted

Top Positions Weighted

Bottom Positions Weighted

Figure 10.2: MDS visualization based on unweighted, top-position weighted, and bottom-positions weighted Kendall distances between people's preferences for the months of the year.

that focusing on the months people like the least is a good way to understand the group structure of their preferences.

10.3.3 Similarity Weights: The Free Recall of Animals

Our third application involves measuring performance on a free recall memory task in a clinical setting, and focuses on the use of similarity weights. The data were collected using the Mild Cognitive Impairment Screen (MCIS: Shankle et al., 2009), one component of a routine assessment of Alzheimer's patients in a neurodegenerative disorders clinic. As part of this assessment, patients complete a triadic comparison task for nine animal names, where each of the animals is presented in a triad with each of the other animals and the patient must determine which of the three animal names is least like the other two. After a delay, patients complete a surprise free recall task of those nine animal names.

One important research goal is to identify and understand the different free recall response patterns. There is evidence that the semantic relationships between the animals influences the order in which their names are recalled (Bousfield & Sedgewick, 1944; Bousfield, 1953; Romney et al., 1993). In particular, it is common for the recalled list to be made up of sub-sequences of semantically-related animal names. For example,"zebra", "giraffe", "elephant", and "tiger" are likely to be recalled consecutively, as a cluster of African zoo animals. In clinical settings, the exact order in which a cluster like this is recalled is less important than the fact it is recalled largely as a cluster, since this suggests semantic memory is intact.

As a concrete example, consider the recall data for three people:

(A) Elephant, Giraffe, Sheep, Rat, Monkey, Chimpanzee, Rabbit, Zebra, Tiger

(B) Rat, Sheep, Giraffe, Zebra, Elephant, Monkey, Chimpanzee, Tiger, Rabbit

(C) Rat, Chimpanzee, Zebra, Giraffe, Elephant, Tiger, Rabbit, Sheep, Monkey

The unweighted Kendall's distance between A and B is 11, between A and C is 18, and between B and C is 13, which implies A and B behave most like one another. We implemented a similarity-weighted measure using the pairwise similarity between each pair of animals determined by an independent triadic comparison task (Lee et al., 2015; Westfall & Lee, 2020). Using this extension of the metric changes the distances between A and B to 30.1, between A and C to 41.7, and between B and C to 26.4, so that B and C become the most similar. Person A breaks the recall of the African zoo animals across extremes of the list, with "elephant" and "giraffe" first and "zebra" and "tiger" last. Persons B and C, in contrast, recall these animals near each other, although not in the same order as one another. The similarity weighting gives less penalty to the transposition of semantically-related animal names, which leads to B and C being measured as having given the most similar responses.

We again use MDS visualizations to explore the overall relationships between people's free recall patterns, based on the Kendall's distance measures. The left-hand panels of Figure 10.3 show the visualizations for the unweighted metric,

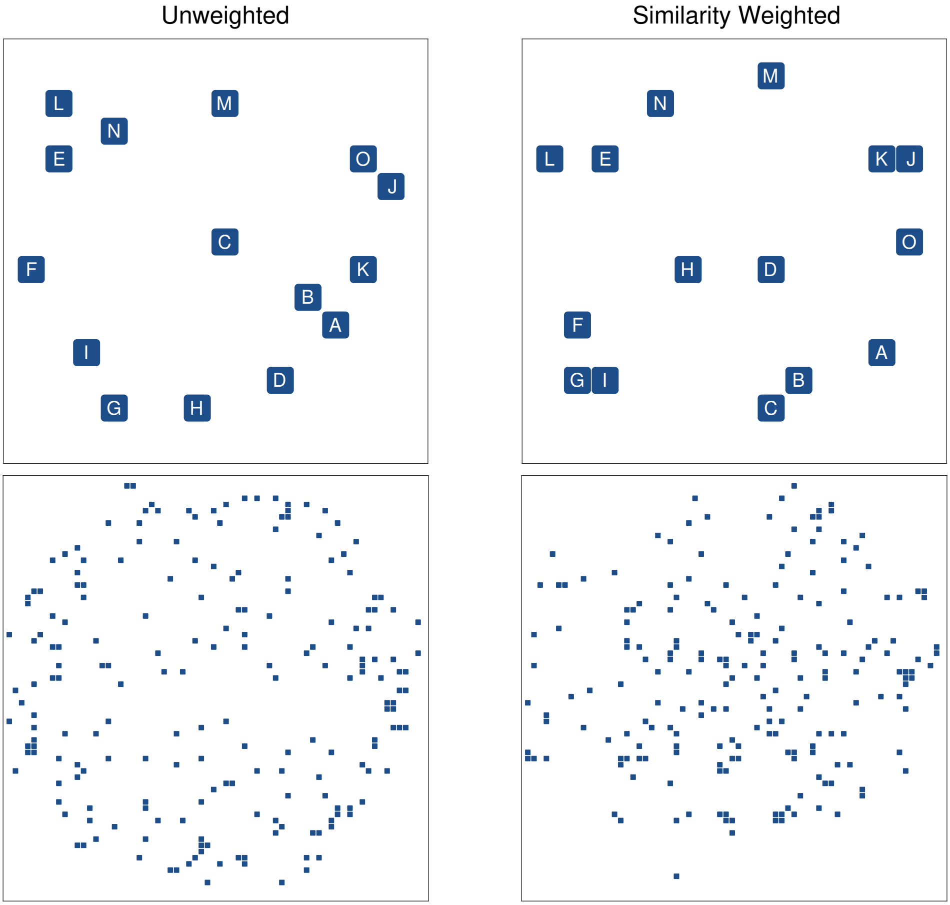


Figure 10.3: MDS visualization of the similarities between recall patterns of animal names based on the unweighted Kendall's distance (left panels) and similarity-weighted Kendall's distance (right panels). The top panels show 15 labeled people, while the bottom panels show all 200 people.

in the top panel for 15 labeled people, including A–C above, and in the bottom panel for all 200 people. The right-hand panels show the corresponding visualizations for the similarity-weighted metric. It is clear that the inclusion of similarity information leads to more clustering between the recall patterns, suggesting the presence of different recall patterns that can be understood in terms of the semantic relationships between the stimuli being recalled.

10.3.4 Top-k: Expert Sporting Predictions

Our last application involves predictions about player performance for the 2017 American Football season by experts from the fantasy football website `fantasypros.com`. On the website, experts provide rankings each week for each playing position commonly used in fantasy football. These rankings serve as advice for players as to which players they should place in their fantasy teams each week. We focus on the rankings of all 85 experts, but just for week 10 of the season, and just for the "kicker" position. We chose the kicker position because it is the one for which different experts often rank different numbers of players. In week 10, experts ranked between 13 and 20 kickers, with a median of 19.

Table 10.3 shows the actual points earned by each kicker[3], as well as the ranking provided by two of the experts. Kendall's distance provides a natural way of measuring the performance of the experts, by quantifying how close their predictions are to the truth. Some players scored the same number of points, which leads to ties in the true ranking. This can be accommodated using similarity weights, assigning a weight of 0 to any pair of kickers who are tied, and 1 to any pair of players who are not tied.

In addition, because the experts ranked different numbers of players, their Kendall's distance depends on the setting of the missingness parameter. For the optimistic setting ($\theta = 0$), expert A has accuracy 103 and expert B has accuracy 133, so that expert A is measured as having made better predictions. For the neutral setting ($\theta = \frac{1}{2}$), expert A has accuracy 153.5 and expert B has accuracy 149.5, so they are very similar. For the pessimistic setting ($\theta = 1$), expert A has accuracy 204 and expert B has accuracy 167, so now expert B is measured as having made better predictions.

In this application, the optimistic setting seems inappropriate. Expert A included only 14 players in their ranking, whereas Expert B included 20 players. Setting $\theta = 0$ means that whenever two players (e.g., Mason Crosby and Blair Walsh) are not ranked by an expert, this expert is given the benefit of the doubt and is not penalized. Expert B does include these two players, but predicts their ranking incorrectly, and is penalized for it. This property makes it appealing for an expert to only include the few players that they are very sure about, which is not what is sought from a good prediction. Both the neutral and pessimistic settings seem more appropriate, since they penalize experts who fail to make predictions about players.

Figure 10.4 shows the change in Kendall's distance for optimistic, neutral, and pessimistic top-k measures for all of the experts. Experts are represented

[3] Players included in expert predictions but not listed did not score any points.

Points	True ranking	Expert A	Expert B
15	Greg Zuerlein	Greg Zuerlein	Stephen Gostkowski
12	Nick Rose	Matt Bryant	Greg Zuerlein
11	Mason Crosby	Stephen Gostkowski	Matt Bryant
11	Stephen Gostkowski	Matt Prater	Matt Prater
11	Wil Lutz	Josh Lambo	Mike Nugent
10	Connor Barth	Graham Gano	Ryan Succop
10	Brandon McManus	Chris Boswell	Chris Boswell
9	Matt Bryant	Kai Forbath	Chandler Catanzaro
9	Graham Gano	Blair Walsh	Kai Forbath
9	Patrick Murray	Ryan Succop	Steven Hauschka
8	Kai Forbath	Wil Lutz	Wil Lutz
8	Matt Prater	Steven Hauschka	Brandon McManus
8	Blair Walsh	Mike Nugent	Josh Lambo
7	Robbie Gould	Chandler Catanzaro	Adam Vinatieri
7	Aldrick Rosas		Graham Gano
6	Chris Boswell		Blair Walsh
6	Zane Gonzalez		Robbie Gould
6	Josh Lambo		Mason Crosby
6	Ryan Succop		Connor Barth
5	Nick Novak		Nick Rose

Table 10.3: The number of fantasy points scored by kickers in week 10 of the 2017 American National Football League season, the true ranking of the players according to these point totals, and the ranked predictions of two experts from fantasypros.com.

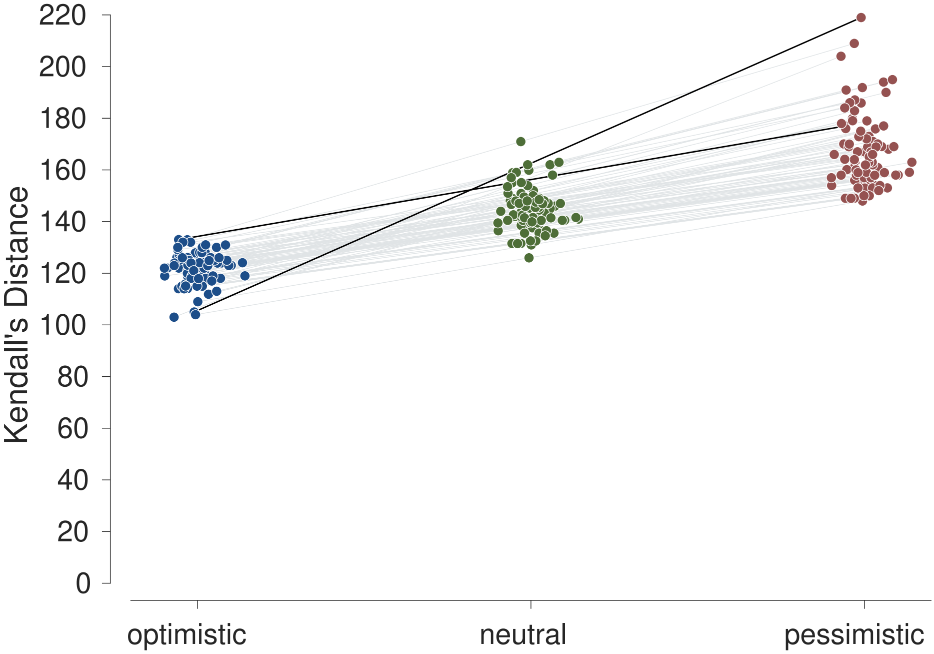


Figure 10.4: Optimistic, neutral, and pessimistic Kendall's distance measures of accuracy for 85 experts from `fantasypros.com`. Each expert predicted the fantasy football performance of kickers in week 10 of the American National Football League 2017 season. Each point corresponds to a participant, jittered around the Kendall's distance measure. The same participant for each measure is connected by a gray line. Experts A and B from Table 10.3 are highlighted by black lines.

by jittered markers with lines connecting the same expert under each measure. Increasing pessimism leads to the experts who ranked fewer players being penalized more heavily for these missing data. Thus, while the distance measure increases for all of the experts as pessimism increases, it increases more quickly for some experts.

10.4 Concluding Comments

In this chapter, we aimed to introduce three extensions of the Kendall's distance metric that are useful for analyzing ranking data in psychological research, as well as demonstrating the ability of the metric to accommodate top-k lists. Our applications gave worked examples of how the extensions can help improve the measurement of key properties of ranking data in the context of specific research goals. Two of the applications focused on measuring people's accuracy, and two focused on measuring the extent and nature of the individual differences between people. Measuring performance and individual differences are among the most common and basic goals of data analysis in psychology.

While we mostly applied the extensions separately, the final application showed that multiple extensions can be used simultaneously. There is nothing preventing Kendall's distance measures being designed to be sensitive to items, their positions, and their similarities in top-k lists where different people have different k. This underscores the flexibility and generality of the metric, and its ability to be adapted to answer specific questions in specific research contexts. While this flexibility should help improve data analysis, it may be important to use pre-registration to make a clear whether and how the extensions to the metric are used in an exploratory way (Lee et al., 2019). An OSF project page associated with this chapter is available at `https://osf.io/6k9t8/` It includes the R script used for calculating weighted Kendall's distance and example applications of the toy example described in this article.

Part III

Conclusion

Chapter 11

Summary and Future Directions

In the first part of my dissertation, I provided guidelines and reflection on Bayesian inference in general, aimed at familiarizing researchers with the core concepts of the Bayesian framework. Moreover, in the first part I explored the extent to which different researchers can approach a research question, both in simple and complex scenarios, such as in the comparison of mixed effects models. In the second part of my dissertation, I applied the Bayesian philosophy to rank-based tests, in order to combine the benefits of Bayesian inference with the benefits of rank-based tests. In this part I demonstrated both the use of ranks in hypothesis testing and the practical relevance of Kendall's τ in the modeling of psychological data. Below, I first summarize each chapter and its main conclusions, and then explore potential directions for future research for each part of the dissertation. This chapter ends with a general conclusion.

11.1 Part I: For Researchers

11.1.1 Chapter Summaries

Chapter 2 introduced the core concepts of Bayesian inference, and provided practical guidelines for the four stages of Bayesian inference: planning, conducting, interpreting, and reporting. Each stage was demonstrated with the running example featuring a Bayesian t-test. The aim of this chapter was to cover a broad spectrum of statistical analyses (i.e., the analyses offered in JASP), where the research question at hand concerns hypothesis testing, parameter estimation, or both. Especially the planning stage is relevant for both Bayesian and non-Bayesian analyses. There is a large emphasis on solidifying the choices made in the planning stage by using a preregistration format (i.e., Nosek et al., 2018; C. D. Chambers, 2013) and to share data and analysis code, in order to promote transparent and reproducible science.

Chapter 3 illustrated that statistical inference features inherently subjective components. Although the two scenario's presented to the four teams of statisticians were relatively simple (an association between two continuous variables and a cross table), each team chose a distinct approach in answering the two research questions. While two teams opted for a Bayesian analysis, their specific paths still differed (e.g., one team applied a log-transformation of the data first, while the other team did not). The other two teams applied other methods such as equivalence testing and p-values. However, all teams were in agreement about both answers to the research questions at hand. The general conclusion for both research questions was that the data were inconclusive, which was remarkable, considering both scenario's were based on the findings of published articles. The disagreements in method, but agreement in conclusion, underscores the idea that careful consideration (i.e., being aware of the capabilities and limitations of the analysis framework of choice) and planning of the analysis is paramount, and that the exact details of the statistical approach come second. Of course, this conclusion holds in the scenario where simple scenario's are deemed inconclusive, and it remains to be seen whether this finding generalizes to more complex scenarios with small effects.

Chapter 4 provided a teaching tool for introducing Bayesian inference at the beginner level. The binomial test provides an excellent starting point for Bayesian statistics education, since it is simple in its use due to its confined parameter space and conjugate prior/posterior, while also applicable to interesting research settings. While the chapter demonstrates the test using a beer tasting experiment conducted at the University of Amsterdam, the research question can of course vary. I have used this chapter and demonstration in my own teaching. Particularly in the first-year bachelor course "Research Methods and Statistics", using the binomial test to estimate and test the ratio of colored chocolate "kruidnoten" has been a fun and educational adventure.[1]

Chapter 5 reported results from a questionnaire sent to lead authors of empirical articles published in the journal *Nature Human Behavior*, in an effort to gauge how applied researchers reason about the concept of evidence for a claim. The chapter reported the results of two questions that asked the researcher to assess the plausibility of the claim in their article before, and after observing the data. The responses to this question enabled the computation of an informal Bayes factor for each researcher's claim, which often was merely supported by "anecdotal" evidence. Since these claims were published in a top tier journal, this result was rather shocking. The discrepancy between the private conviction of the researcher and how the claim is publicly reported, uncovers a flaw in the academic reporting system, where transparency about uncertainty is penalized, rather than rewarded.

Chapter 6 outlined several choices that emerge in the Bayesian comparison of mixed effects models. Specifically, it demonstrated how the choice of null and alternative model affect the definition of "an effect" in the context of random effects. Additionally, it explored the effect of aggregating the data, the choice of prior distributions, and the role that measurement error plays in mixed model comparisons. The aim of this chapter was to provide a common starting ground for discussion among experts about best practices in Bayesian comparison of mixed effects models, in order to draft a set of practical guidelines. By demonstrating the behavior of three possible model comparisons (i.e., com-

[1] The Bayesian lectures were given around the time of Sinterklaas, a Dutch celebration, where kruidnoten are a popular sweet.

bination of null and alternative model) in three different example scenario's, the differences between the model comparisons were underscored. Rather than electing the "best" model comparison, the aim was to inform the reader about these choices, such that they can choose the option that best suits their research question.

11.1.2 Discussion and Future Directions

The first part of this dissertation was contemplative in nature, and presented the perspectives of both statisticians and applied researchers on statistical inference. In this section, I explore avenues for future development that will focus on bridging the gap between the two groups of researchers. While expert opinions might differ about which specific statistical framework is the most appropriate, even in simple research settings, there is considerable agreement that the analyst should be well-informed and transparent about the tools they are using. When each step of the analysis is documented, and the data made openly available, critical readers of an article may reproduce the analysis, and explore to what extent the findings are robust to alternative analysis options. For example, a researcher first aggregates their observations, conducts a Bayesian repeated measures ANOVA (see Example 3 in Chapter 6), and reports evidence for an interaction effect. Another researcher ought to be able to repeat the analysis, but instead of aggregating the data, conduct a full mixed effects model comparison to assess how well the conclusion generalizes across different analysis setups.

Transparent inference is one part (and is easy to preach from our methodological ivory tower), but the careful documentation and dissemination of the various analysis choices that exist for a specific experimental design is a crucial second part. I believe starting centralized discussions on best practices, as demonstrated in Chapters 3 and 6, are an important area of psychological methods. Improving statistical practice in the field is arguably the most important goal of psychological methods, but the statistical literature can easily feel overwhelming for researchers in psychology. In other areas of psychology, research findings in the literature are typically aggregated using meta-analysis. While unsuited for quantitative aggregation, statistical methods can be aggregated through qualitative meta-*synthesis*. Currently, my colleagues and I are experimenting with this format, where Chapter 6 will serve as the starting point for discussion. If the discussion on mixed models in the near future proves fruitful, a set of guiding principles can be drafted that will be both practical and informative to the novice practitioner in mixed model comparison.[2] In doing so, it will be similar to the guidelines discussed in Chapter 2.

The idea behind this methodological meta-synthesis is to carefully select several example data sets (either synthetic or real) that are sensitive to the specific issues at hand, such that different approaches lead to qualitatively different conclusions and will therefore foster discussion. A potential shortcoming of Chapter 3 is that both example data sets were deemed inconclusive by all the analysis teams. Although the exercise itself still proved insightful, it limited the extent

[2] Of course, caution is warranted not to repeat iconic xkcd comic 927 (https://xkcd.com/927/).

of the discussion afterwards. It would therefore be valuable to conduct similar studies of experts analyzing identical data sets, but for more complex and slightly more conclusive data sets, in similar fashion to the Many Analysts setup (Silberzahn et al., 2018). The goal of the Many Analysts setup is to underscore statistical inference as a subjective exercise and to use the setup to create a robust inferential process. In contrast, the goal of this new line of studies is to use the differences between experts to demonstrate the practical relevance and implication of each analysis choice, as an educational endeavor, and to draft a set of guidelines based on expert consensus. Concretely, the first avenue of future development is therefore to continue the mixed modeling synthesis, and to apply this general framework to other analyses that lack a clear set of guiding principles, such as meta-analysis, Bayesian post-hoc testing in ANOVA, and accommodating violated assumptions.

Whereas the future development of methodological synthesis focuses on orchestrating discussion among experts, I also propose to further explore how applied researchers view the process of statistical inference. The questionnaire and data set presented in Chapter 5 offer a wealth of information on this subject, and can inform statisticians about how to best provide methodological and statistical advice. For example, proponents of the Bayes factor view the updating of prior odds to posterior odds of two competing hypotheses to be the holy grail of statistical inference. However, 15 out of the 31 respondents felt that a decrease in plausibility is not problematic, as long as the posterior odds are still in favor of the alternative hypothesis.[3] Bayes factor tutorial articles can use this information and emphasize why knowledge updating is important, or focus on the interplay between Bayes factor and posterior odds. Concretely, the second avenue for future development will focus on exploration and dissemination of the current data set, as well as expand the data set by approaching more leading authors of articles in high impact journals.

11.2 Part II: For Ranks

11.2.1 Chapter Summaries

Chapter 7 introduced a Bayesian framework for hypothesis testing and parameter estimation for the rank correlation Kendall's τ. The framework was based on work by Johnson (2005), who obtained an upper bound on BF_{10} (i.e., a lower bound on BF_{01}) by modeling the asymptotic sampling distribution of the test statistic. In order to create a default prior distribution for Kendall's τ, the default prior distribution for Pearson's ρ was transformed using Greiner's relation (Greiner, 1909). Additionally, the alternative hypothesis was adjusted such that, in combination with the new default prior distribution, a Bayes factor and posterior distribution were obtained.

[3] Question 3a of the questionnaire: "Suppose another researcher conducted a study in their field and then answered the previous two questions. Before seeing the data, the researcher was 80% confident that the claim is true. After seeing the data, the researcher was 60% confident that the claim is true . The researcher now argues that the data support their claim. Do you think this is reasonable?"

Chapter 8 provided an alternative method for Bayesian inference for Kendall's τ using data augmentation. In this setup, rank data are seen as impoverished manifestations of a latent (i.e., unobserved), normally distributed construct. Using Markov chain Monte Carlo sampling, the latent construct can be approximated in terms of a posterior distribution for the latent values, while the ordinal information in the data is preserved. In this Gibbs sampling algorithm, the posterior distribution of the parametric correlation Pearson's ρ is also approximated. The resulting posterior distribution of Pearson's ρ can then be transformed using Greiner's relation, to obtain the posterior distribution of Kendall's τ. This approach leads to highly similar inference as the approach outlined in Chapter 7, except for small sample sizes and high values for Kendall's τ, where the asymptotic approximation might not hold.

Chapter 9 applied the same latent normal algorithm introduced in Chapter 8, but presented this framework as a general method for constructing Bayesian versions of rank-based tests. The framework was applied to create Bayesian equivalents of the Wilcoxon rank sum test (i.e., the Mann-Whitney U test), the Wilcoxon signed rank test, and Spearman's ρ. The main idea of the data augmentation method is to use the ordinal information in the data to approximate the latent construct, and to conduct the parametric test (e.g., the Bayesian t-test, in the case of the Wilcoxon rank sum test) on these latent scores. This respects the uncertainty inherent in rank data, while creating a test that is easy to understand, since it is based on the parameterization of the parametric equivalent. For example, the Bayesian Wilcoxon rank sum test yields a posterior distribution for effect size δ, but on the latent level. Additionally, the discussed tests have an asymptotic relative efficiency close to 1, compared to their parametric equivalents, when the data are normally distributed. When the data depart from normality, the power of the rank-based tests surpasses their parametric equivalents.

Chapter 10 did not focus on hypothesis tests or parameter estimation, but instead presented Kendall's distance as a modeling tool for rank data in psychology. While Kendall's τ is often used for capturing a rank-based association between two variables, its unstandardized version is a highly versatile statistic that can be used to aggregate observed data. The chapter presented four extensions of Kendall's distance that can be used to model psychological phenomena, such as item similarity, item importance, and item position. The metric was demonstrated by applying it to four research scenario's. First, the serial recall of the events of 9/11, where incorrectly recalling some events is penalized more heavily than incorrectly recalling other events, using item weights. Second, participants' ranking of all twelve months, from most liked to most hated, where the most hated months were weighted more heavily with position weights to detect clusters of participants with similar aversions. Third, a free recall memory task where response styles were modeled, and animal similarities were accounted for by similarity weights. Fourth, experts' predictions of American football player performance, where varying lengths of the experts' responses (i.e., top-k lists) were accommodated using the missingness penalty parameter.

11.2.2 Discussion and Future Directions

The second part of this dissertation was more technical in nature, and presented Bayesian versions of several rank-based tests. While the asymptotic framework in Chapter 7 is limited to a few tests, the latent normal framework in Chapters 8 and 9 can be applied more generally. I believe that a major advantage of this framework is the ease of its interpretation, since it is closely related to the Bayesian framework for the parametric tests. For instance, the Bayesian Wilcoxon rank sum test applies the Bayesian t-test of Rouder et al. (2009), but on the latent level. The benefit of the latent normal framework is that the interpretation, prior specification, and parameterization are very similar between the rank-based and parametric methods, but where the rank-based method accounts for the extra uncertainty inherent in ordinal observations. Additionally, the rank-based tests eliminate certain arbitrary decisions in the analytic process (e.g., whether to apply a monotonic transformation, how to handle outliers, or whether the data are normally distributed).

In recent years, two alternatives to the method outlined in Chapters 8 and 9 have been developed. First, a similar latent normal method is implemented in the `brms` package (Bürkner, 2017; Bürkner & Vuorre, 2019) that allows rank-based parameter estimation for linear models. Although easily applied and quite versatile, the method lacks Bayes factor hypothesis tests. Second, Chechile (2020) outlines various Bayesian rank-based tests, such as the Wilcoxon signed rank test and Kendall's τ. This method uses the binomial likelihood to model the probability of observing a positive difference (in the case of the Wilcoxon signed rank test), or a concordant pair (in the case of Kendall's τ). However, it seems that the Bayes factors obtained through these methods are overly optimistic. For instance, for the data example in Chapter 7, a data set with only 20 observations and an observed Kendall's τ of 0.28, Chechile obtains $\text{BF}_{10} = 186$. This is in stark contrast to the $\text{BF}_{10} = 2.17$ presented in Chapter 7, particularly considering $\text{BF}_{10} = 5.2$ for Pearson's ρ. An example of the Wilcoxon signed rank test is available at https://osf.io/2wgtc/.

An obvious future direction of this part of my research is to explore applications of the latent normal framework to other rank-based tests. Some preliminary work has already been done to realize Bayesian versions of the Kruskal-Wallis test and the Friedman test (i.e., rank-based ANOVA), and the latent normal framework can also be extended to create Bayes factor hypothesis tests for partial rank correlations (e.g., Kendall, 1942; Q. Liu et al., 2018) and ordinally constrained parameters (J. M. Haaf et al., 2018). However, a more adventurous endeavor would be to explore the different models for rank data presented by Marden (1995). These models provide likelihood functions for rank data in certain settings. For instance, Mallows's ϕ model (Mallows, 1957) is a distance-based model that expresses the likelihood of the observed ranks as a function of the "modal" ranking (i.e., a ground truth, or the consensus ranking) and a dispersion parameter that governs how close the responses are to the modal ranking. The work presented in Chapter 10 is particularly suited to Mallows's ϕ model, since this model uses Kendall's distance to quantify the distance between observed rankings and the modal ranking. It would therefore be worthwhile to explore the distance-based models and incorporate the three weighting extensions outlined in Chapter 10. Mallows's ϕ model can be applied to analyze general agreement in rankings (e.g., preference rankings) or accuracy of participants' responses (e.g., in a memory recall experiment), and by incorporating the three weighting extensions, the flexibility of the model is greatly enhanced. Concretely, the third avenue for future development will focus on documenting the models presented by Marden, illustrating their usefulness in psychological science, and expanding the distance-based models with the weighted Kendall's distance.

11.3 General Conclusion

This dissertation has discussed Bayes factor hypothesis testing on two levels. First, a contemplative series of guidelines, reflections, and educational tools was provided, in order to increase the understanding, transparency, and rigor in the application of Bayesian inference. Second, the focus was shifted to providing tools for conducting rank-based tests and modeling rank data. Together, the two parts of this dissertation will enable any researcher to properly conduct a Bayes factor hypothesis test in a scenario with ordinal, or non-normal, measurements and a research question that pertains to testing for a difference between two independent groups, two dependent groups, or an association between two variables.

In this concluding chapter I have outlined three potential avenues for future development. First, to develop a wider array of methodological meta-syntheses by demonstrating different options in a specific statistical framework, documenting the approaches and opinions by experts in the field, and drafting a set of guiding principles. Second, to increase the effectiveness of statistical recommendations by studying the practical and conceptual challenges faced by applied researchers. Third, to extend distance-based models for rank data with the weighted partial Kendall's distance.

The final path for the future is more general. Most chapters in this dissertation are related to JASP, either by using JASP as an educational or illustrative tool, providing guidelines on how to best use it, or by developing new analyses and making these available through JASP. Because JASP is open source, features a graphical user interface, and has an active user community (e.g., through Github, the JASP forum, and social media), it not only makes statistics more accessible, but also lowers the threshold for users to be informed about best practices. In my work for JASP, I have therefore not only focused on adding analyses to the platform, but also on expanding the documentation for the existing analyses, and advising users on specific analysis issues. At the start of this dissertation, I stated my belief that a successful statistical method is not only functional and accurate, but also features proper dissemination and documentation. With over $100,000$ downloads in the past three months, I believe that JASP is instrumental in lowering the threshold for researchers to conduct (rank-based) Bayes factor hypothesis tests. In the future, I aim to further expand the available analyses, write tutorial blogs and articles, and teach workshops.

Exploring each avenue presented above aims to bridge the gap between applied and statistics researchers, and to ensure prudent statistical practice across the board.

Part IV

Appendices

References

Albert, J. H. (1992a). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics, 17*(3), 251–269.

Albert, J. H. (1992b). Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation, 44*, 47–61.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669–679.

Altmann, E. M. (2003). Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 17*, 1067–1080.

Alvo, M., & Yu, P. (2014). *Statistical methods for ranking data*. New York: Springer New York.

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology, 66*, 1–7.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*, 17–21.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*, 3–25.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4:328*.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

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

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19*, 58–80.

Beg, M. S., & Ahmad, N. (2003). Soft computing techniques for rank aggregation on the world wide web. *World Wide Web, 6*, 5–22.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.

Bennett, J. H. (Ed.). (1990). *Statistical inference and analysis: Selected correspondence of R. A. Fisher*. Oxford: Clarendon Press.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18*, 1–32.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O., & Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics, 36*, 963–982.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics, 29*, 201–218.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.

Borgwardt, K. M., & Ghahramani, Z. (2009). Bayesian two-sample tests. *arXiv preprint arXiv:0906.4032*.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology, 49*, 229–240.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology, 30*, 149–165.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of computational social choice*. Cambridge University Press.

Brooks, R. J. (1974). Bayesian analysis of the two-sample problem under the Lehmann alternatives. *Biometrika, 61*, 501-507.

Brooks, R. J. (1978). Bayesian analysis of a two-sample problem based on the rank order statistic. *Journal of the Royal Statistical Society: Series B*, *40*, 50-57.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*, 77–101.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

Chambers, J., & Hastie, T. (1992). *Statistical models in S*. Wadsworth & Brooks/-Cole Advanced Books & Software.

Chechile, R. A. (2020). *Bayesian statistics for experimental scientists: A general introduction using distribution-free methods*. MIT Press.

Chen, J. J., Tsong, Y., & Kang, S.-H. (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, *26*, 569–578.

Chen, Y., & Hanson, T. E. (2014). Bayesian nonparametric k-sample tests for censored and uncensored data. *Computational Statistics & Data Analysis*, *71*, 335–346.

Chernoff, H., & Savage, R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *The Annals of Statistics*, *29*, 972–994.

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*, 80–101.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*, 385–396.

Colonius, H. (2016). An invitation to coupling and copulas: With applications to multisensory modeling. *Journal of Mathematical Psychology*, *74*, 2–10.

Conover, W. (1999). *Practical nonparametric statistics* (3rd ed.). Wiley.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of 5th annual future business technology conference* (pp. 5–12). EUROSIS.

Cureton, E. (1956). Rank-biserial correlation. *Psychometrika*, *21*, 287–290.

De Morgan, A. (1847/2003). *Formal logic: The calculus of inference, necessary and probable*. Honolulu: University Press of the Pacific.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*, 240–261.

Diamond, G. A., & Kaul, S. (2004). Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. *Journal of the American College of Cardiology*, *43*, 1929–1939.

Dickey, J. M., & 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.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychololog*y, *5:781*.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*, 207–218.

Draper, N. R., & Cox, D. R. (1969). On distributions and their transformation to normality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*, 472–476.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*, 1–5.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*, 60–69.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1*(2), 281–295.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, *17*, 134–160.

Ferguson, T. S., Genest, C., & Hallin, M. (2000). Kendall's tau for serial dependence. *Canadian Journal of Statistics*, *28*, 587–604.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 507–521.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fligner, M. A., & Verducci, J. S. (1986). Distance based ranking models. *Royal Statistical Society . Series B*, *48*, 359–369.

Fligner, M. A., & Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, *83*, 892–901.

Frisby, J. P., & Clatworthy, J. L. (1975). Learning to see complex random-dot stereograms. *Perception*, *4*, 173–178.

Gelman, A. (2013). *p* values and statistical practice. *Epidemiology*, *24*, 69–72.

Gelman, A., & Vehtari, A. (2020). What are the most important statistical ideas of the past 50 years? *arXiv preprint arXiv:2012.00174*.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Genest, C., & 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.

Goodman, S. N. (2018). How sure are you of your result? Put a number on it. *Nature*, *564*, 7–8.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from https://mc-stan.org/rstanarm (R package version 2.21.1)

Greiner, R. (1909). Über das Fehlersystem der Kollektivmasslehre. *Zeitschift für Mathematik und Physik*, *57*, 121–158.

Griffin, H. (1958). Graphic computation of tau as a coefficient of disarray. *Journal of the American Statistical Association*, *53*, 441–447.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2018). Informed Bayesian t-tests. Manuscript submitted for publication. *arXiv preprint arXiv:1704.02479*.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, *92*(10).

Haaf, J., Ly, A., & Wagenmakers, E. (2019). Retire significance, but still test hypotheses. *Nature*, *567*(7749), 461.

Haaf, J. M., Klaassen, F., & Rouder, J. (2018). *Capturing ordinal theoretical constraint in psychological science.* PsyArXiv. Retrieved from psyarxiv.com/a4xu9

Hastings, W. (1970). Monte Carlo samplings methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, *35*, 258–279.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*, 200–215.

Hodges, J., & Lehmann, E. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, *27*, 324–335.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293–325.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382–401.

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, *1*, 265–283.

Hoff, P. (Ed.). (2009). *A first course in Bayesian statistical methods*. Dordrecht, The Netherlands: Springer.

Hollander, M., & Wolfe, D. (1973). *Nonparametric statistical methods* (3rd ed.). New York: Wiley.

Holmes, C. C., Caron, F., Griffin, J. E., & Stephens, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Analysis*, *10*, 297–320.

Hotelling, H., & Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, *7*, 29–43.

Howie, D. (2002). *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Jackson, R. W. (1939). Reliability of mental tests. *British Journal of Psychology*, *29*, 267–287.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*, 2-9.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*, 422–446.

JASP Team. (2020). *JASP (Version 0.14.1)[Computer software]*. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1937). On the relation between direct and inverse methods in statistics. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *160*, 325–348.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54–69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

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

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, *54*, 129–144.

Kempthorne, O. (1975). Fixed and mixed models in the analysis of variance. *Biometrics*, *31*, 473–486.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93.

Kendall, M. (1942). Partial rank correlation. *Biometrika*, *32*, 277–283.

Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods*. New York: Oxford University Press.

Kerby, D. (2014). The simple difference formula: an approach to teaching nonparametric correlation. *Innovative Teaching*, *3*, 1–9.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799.

Kousta, S. (Ed.). (2020). Editorial: Tell it like it is. *Nature Human Behavior*, *4*, 1.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

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

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, *53*, 814–861.

Kumar, R., & Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th international conference on world wide web* (pp. 571–580).

Labadi, L. A., Masuadi, E., & Zarepour, M. (2014). Two-sample Bayesian nonparametric goodness-of-fit test. *arXiv preprint arXiv:1411.3427*.

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

Lee, M. D., Abramyan, M., & Shankle., W. R. (2015). New methods, measures, and models for analyzing memory impairment using triadic comparisons. *Behavior Research Methods*.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... others (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*, 141–153.

Lee, M. D., Steyvers, M., & Miller, B. J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE*, *9*, 1–9.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lehmann, E. (1975). *Nonparametrics: Statistical methods based on ranks* (1st ed.). London ; New York: Holden-Day, Inc.

Lehmann, E. (1999). *Elements of large sample theory*. Springer.

Liang, F., German, R. P., Clyde, A., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–424.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia (PA): SIAM.

Lindley, D. V. (1986). Comment on "Why isn't everyone a Bayesian?" by Bradley Efron. *The American Statistician*, *40*, 6–7.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*, 673–687.

Liu, Q., Li, C., Wanga, V., & Shepherd, B. E. (2018). Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics*, *74*, 595–605.

Liu, S., & Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, *87*, 353–369.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Lukács, G., Kleinberg, B., Kunzi, M., & Ansorge, U. (2020). Response time concealed information test on smartphones. *Collabra: Psychology*, *6*, 1–14.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 1–11.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*, 4–13.

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.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, *44*, 114–130.

Mann, H., & Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.

Marden, J. I. (1995). *Analyzing and modeling rank data* (1st ed ed.). London ; New York: Chapman & Hall.

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*, 545–555.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Morey, R. D., & Rouder, J. N. (2018). Bayesfactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-4.2)

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)*, *140*, 48–63.

Nelsen, R. (2006). *An introduction to copulas* (second ed.). Springer-Verlag New York.

Noether, G. E. (1955). On a theorem of Pitman. *Annals of Mathematical Statistics*, *26*, 64–68.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 2600–2606.

Olssen, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Olssen, U., Drasgow, F., & Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 443–460.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *195*, 1–405.

Pearson, K., & Pearson, E. (1922). On polychoric coefficients of correlation. *Biometrika*, *14*, 127–156.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society Series B*, *44*, 234–243.

Pinheiro, J. C., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=nlme` (R package version 3.1-150)

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed–effects models in S and S–PLUS*. New York: Springer.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.* Vienna, Austria.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

R Development Core Team. (2004). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3–900051–00–3)

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: Wiley.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*, 309–338.

Röhmel, J. (2001). Statistical considerations of FDA and CPMP rules for the investigation of new anti–bacterial products. *Statistics in Medicine*, *20*, 2561–2571.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*, 28–34.

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

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, *23*, 1779–1786.

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

Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*, 186-190.

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2013). Hierarchical Bayesian models. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology: Volume 1, foundations and methodology.* London, United Kingdom: Cambridge University Press.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

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

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Savage, I. (1956). Contributions to the theory of rank order statistics-the two-sample case. *The Annals of Mathematical Statistics*, *27*, 590–615.

Scheffe, H. (1956). Alternative models for the analysis of variance. *The Annals of Mathematical Statistics*, *27*, 251–271.

Schielzeth, H., & Forstmeier, W. (2008). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

Schramm, P., & Rouder, J. N. (2019). Are reaction time transformations really beneficial? *PsyArXiv. March*, *5*.

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-*n* lists. *Decision*, *4*, 87–101.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shankle, W. R., Mangrola, T., Chan, T., & Hara, J. (2009). Development and validation of the Memory Performance Index: Reducing measurement error in recall tests. *Alzheimer's & Dementia*, *5*, 295–306.

Shaw, C. J., & Trimble, T. (1963). Algorithm 175: shuttle sort. *Communications of the ACM*, *6*, 312–313.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=afex (R package version 0.26-0)

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. *New Methods in Cognitive Psychology*, *28*, 4–31.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*, 175–200.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*, 1–130.

Stan Development Team. (2016). *rstan: The R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.14.1)

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*, 1042–1058.

Sung, L., Hayden, J., Greenberg, M. L., Koren, G., Feldman, B. M., & Tomlinson, G. A. (2005). Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *Journal of Clinical Epidemiology*, *58*, 261–268.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Tawakol, A., Ishai, A., Takx, R. A. P., Figueroa, A. L., Ali, A., Kaiser, Y., ... Pitman, R. K. (2017). Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, *389*, 834–845.

Thall, P. F., & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 657–671.

Thalmann, M., & Niklaus, M. (2018). BayesRS: Bayes factors for hierarchical linear models with continuous predictors [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=BayesRS (R package version 0.1.3)

The BaSiS group. (2001). *Bayesian standards in science: Standards for reporting of Bayesian analyses in the scientific literature.* Internet. Retrieved from http://lib.stat.cmu.edu/bayesworkshop/2001/BaSis.html

Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests does not suffice: A plea for plausibility. *Psychonomic Bulletin & Review*, *25*, 548–559.

Tukey, J. W. (1995). Controlling the proportion of false discoveries for multiple comparisons: Future directions. In V. S. L. Williams, L. V. Jones, & I. Olkin (Eds.), *Perspectives on statistics for educational research: Proceedings of a workshop* (pp. 6–9). Research Triangle Park, NC: National Institute of Statistical Sciences.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Beyond the new statistics: Bayesian inference for psychology [special issue]. *Psychonomic Bulletin & Review*, *25*.

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2019). *A cautionary note on estimating effect size.* PsyArXiv. Retrieved from psyarxiv.com/h6pr8

van der Vaart, A. (2000). *Asymptotic statistics*. Cambridge University Press.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, *72*, 303-308.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2019). Bayesian estimation of Kendall's tau using a latent normal approach. *Statistics & Probability Letters*, *145*, 268–272.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's rho. *Journal of Applied Statistics*, *47*, 2984-3006.

van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*, 1-50.

van Rossum, G. (1995). *Python tutorial* (Tech. Rep. No. CS-R9526). Amsterdam: Centrum voor Wiskunde en Informatica (CWI).

Varnum, M. E., & Grossmann, I. (2016). Pathogen prevalence is associated with cultural changes in gender equality. *Nature Human Behaviour*, *1*, 1–4.

Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., ... Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology: Cognition*, *6:494*.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., ... Morey, R. D. (2018b). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018a). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

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 S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). John Wiley and Sons.

Wasserman, L. (2006). *All of nonparametric statistics*. New York: Springer Science and Business Media.

Wasserstein, R., & Lazar, N. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Weber-Schoendorfer, C., & Schaefer, C. (2008). The safety of cetirizine during pregnancy: A prospective observational cohort study. *Reproductive Toxicology*, *26*, 19–23.

Westfall, H. A., & Lee, M. D. (2020). A model-based analysis of the impairment of semantic memory. *Manuscript submitted for publication.*

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & 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*, 752–760.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Willerman, L., Schultz, R., Rutledge, J. N., & Bigler, E. D. (1991). In vivo brain size and intelligence. *Intelligence*, *15*, 223–228.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, *2*, 181–247.

Yuan, Y., & Johnson, V. E. (2008). Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, *18*, 1185–1200.

Zhou, L., Deng, C., Ooi, T. L., & He, Z. J. (2016). Attention modulates perception of visual space. *Nature Human Behaviour*, *1*, 1–5.

Appendix A

Online Supplementary Material

	Summary
Chapter 2	
https://osf.io/25ekj/	Annotated .jasp file of the running example (*t*-test and Wilcoxon rank sum test)
https://osf.io/wae57/	Annotated .jasp file of online Example 2 (Mixed ANOVA)
https://osf.io/q38da/	Annotated .jasp file of online Example 3 (Correlation)
https://osf.io/ybszx/	Annotated .jasp file of online Example 4 (Informed *t*-test)
Chapter 3	
https://osf.io/hykmz/	Repository of the two data sets that were analyzed by the teams
https://osf.io/f4z7x/	The complete email discussion between the teams, after the analyses were conducted
Chapter 4	
https://osf.io/cf3a2/	Annotated .jasp file of data and analyses of the beer tasting experiment
https://osf.io/428pb/	Repository of the data and videos of data collection

	Summary
Chapter 5	
https://osf.io/zjnpm/	The questionnaire and text based answers of respondents (“Why (not) is it reasonable to publish a claim after seeing a decrease in plausibility?” and “What do you believe constitutes statistical evidence?”)
https://osf.io/kd4ps/	Annotated .jasp file of the analysis of the difference in researchers’ assessments of prior and posterior plausibility of their claims
Chapter 6	
https://tinyurl.com/y7nlelyy	Shiny app for simulating mixed effects data for various parameter/sampling settings
https://tinyurl.com/ycamajfw	Shiny app for exploring simulation study results, comparing the different model comparisons for various parameter settings
https://osf.io/tjgc8/	R-code to generate and analyze the data set used in Example 1 (the effect of aggregation)
https://osf.io/xpk85/	R-code to generate and analyze the data sets used in Example 2 (the effect of measurement error)
https://osf.io/cw5jd/	R-code to analyze the data set used in Example 3 (a random interaction effect)
Chapter 7	
https://osf.io/bg4vw/	R-code to compute the posterior distribution and Bayes factor for Kendall’s τ using the asymptotic method
https://osf.io/es5ag/	Illustration of prior distribution and simulation study for verifying the asymptotic normality of Kendall’s τ
Chapter 8	
https://osf.io/87zqx/	R-code to compute the posterior distribution and Bayes factor for Kendall’s τ using the latent normal method
https://osf.io/b54mp/	Simulation study for comparing the behaviors of the different methods for Bayesian inference of Kendall’s τ

	Summary
Chapter 9	
https://osf.io/gny35/	OSF repository of R-code for the rank sum, signed rank, and Spearman's ρ
https://osf.io/j5wud/	R-code for reproducing all examples in paper
https://tinyurl.com/y9ogaa6d	Shiny app for exploring the simulation study results for the ranks sum and signed rank test
https://tinyurl.com/y9oewhss	Shiny app for exploring the simulation study results for Spearman's ρ
Chapter 10	
https://osf.io/5agdv/	R code for the computation of the weighted partial Kendall's distance
https://osf.io/4ej6s/	Example applications of the weighted partial Kendall's distance for the toy soda data

Appendix B

Publications

B.1 Under Review

1. **van Doorn, J.B.**, Aust, F., Haaf, J.M., Stefan, A., & Wagenmakers, E.–J. (under review). Bayes Factors for Mixed Models.
2. **van Doorn, J.B.**, Westfall, H., & Lee, M.D. (under review). Using the Weighted Kendall's Distance to Analyze Psychological Data and Models.
3. Scheepstra, K.W.F., **van Doorn, J.B.**, Scheepens, D.S., de Haan, A., Schukking, N., Zantvoord, J.B., & Lok, A. (under review). Rapid speed of response to ECT treatment in bipolar depression: A retrospective chart review.

B.2 Published

1. **van Doorn, J.B.**, van den Bergh, D., Dablander, F., Derks, K., van Dongen, N.N.N., Evans, N. J., Gronau, Q. F., Haaf, J.M., Kunisato, Y., Ly, A., Marsman, M., Sarafoglou, A., Stefan, A., & Wagenmakers, E.–J. (in press). Strong Public Claims May Not Reflect Researchers' Private Convictions. *Significance.*
2. **van Doorn, J.B.**, van den Bergh, D., Boehm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharsky, S., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.–J. (in press). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*.
3. **van Doorn, J.B.**, Ly, A., Marsman, M., & Wagenmakers, E.–J. (2020). Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's ρ. *Journal of Applied Statistics, 47*, 2984–3006.
4. **van Doorn, J.B.**, Matzke, D., & Wagenmakers, E.–J. (2020). An In-Class Demonstration of Bayesian Inference. *Psychology Learning and Teaching, 19*, 36–45.

5. Ly, A., Stefan, A., **van Doorn, J.B.**, Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharsky, S., Derks, K., Gronau, Q. F., Raj, A., Boehm, U., van Kesteren, E.–J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.–J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the *p* value hypothesis test. *Computational Brain & Behavior, 3*, 153–161.

6. van den Bergh, D., **van Doorn, J.B.**, Marsman, M., Draws, T., van Kesteren, E.–J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.–J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology, 120*, 73–96.

7. **van Doorn, J.B.**, Ly, A., Marsman, M., & Wagenmakers, E.–J. (2019). Bayesian Estimation of Kendall's tau Using a Latent Normal Approach. *Statistics and Probability Letters, 145*, 268–272.

8. Crüwell, S., **van Doorn, J.B.**, Etz, A., Makel, M.C., Moshontz, H., Niebaum, J., Orben, A., Parsons, S. and Schulte-Mecklenbeck, M. (2019). 7 Easy Steps to Open Science: An Annotated Reading List. *Zeitschrift für Psychologie, 227*, 237–248.

9. van Dongen, N. N. N., **van Doorn, J. B.**, Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.–J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician, 73*, 328–339.

10. **van Doorn, J.B.**, Ly, A., Marsman, M., & Wagenmakers, E-J. (2018). Bayesian Inference for Kendall's Rank Correlation Coefficient. *The American Statistician, 72*, 303–308.

11. Jepma, M., Koban, L., **van Doorn, J.B.**, Jones, M., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature Human Behaviour, 2*, 838–855.

12. 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.B.**, Smira, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58–76.

13. Derks, K., Burger, J., **van Doorn, J.B.**, Kossakowski, J. J., Matzke, D., Atticciati, L., Beitner, J., Benzesin, V., de Bruijn, A. L., Cohen, T. R. H., Cordesius, E. P. A., van Dekken, M., Delvendahl, N., Dobbelaar, S., Groenendijk, E. R., Hermans, M. E., Hiekkaranta, A. P., Hoekstra, R. H. A., Hoffmann, A. M.,

Hogenboom, S. A. M., Kahveci, S., Karaban, I. J., Kevenaar, S. T., te Koppele, J. L., Kramer, A.-W., Kroon, E., Kucharsky, S., Lieuw-On, R., Lunansky, G., Matzen, T. P., Meijer, A., Nieper, A., de Nooij, L., Poelstra, L., van der Putten, W. J., Sarafoglou, A., Schaaf, J. V., van de Schraaf, S. A. J., van Schuppen, S. and Schutte, M. H. M. and Seibold, M. and Slagter, S. K. and Snoek, A. C. and Stracke, S. and Tamimy, Z. and Timmers, B. and Tran, H. and Uduwa-Vidanalage, E. S. and Vergeer, L. and Vossoughi, L. and Yucel, D. E., & Wagenmakers, E.–J. (2018). Network models to organize a dispersed literature: The case of misunderstanding analysis of covariance. *Journal of European Psychology Students*, *9*, 48–57.

14. **Open Science Collaboration**. (2015). Estimating the reproducibility of psychological science. *Science*, *349*.

Appendix C

Dutch summary

Dit proefschrift, getiteld “Bayes Factor Hypothesis Tests for Ranks and Researchers”, behandelt verschillende methoden voor gedragsmatig onderzoek. Het eerste deel van dit proefschrift biedt richtlijnen voor, en beschouwing van, Bayesiaanse statistiek in het algemeen, met als doel het introduceren van de centrale concepten van Bayesiaanse statistiek voor toegepaste onderzoekers. Dit deel verkent ook de verschillende routes die onderzoekers kunnen bewandelen wanneer zij geconfronteerd worden met een onderzoeksvraag en dataset. Het tweede deel van dit proefschrift past de Bayesiaanse filosofie toe op hypothesetoetsing voor rangorde data, zodat de voordelen van rangorde data gecombineerd kunnen worden met de voordelen van Bayesiaanse statistiek. Dit deel laat zien hoe rangordes gebruikt kunnen worden in hypothesetoetsing, en hoe Kendall’s τ gebruikt kan worden om psychologische data te modelleren.

C.1 Deel I: For Researchers

Hoofdstuk 2 introduceert centrale concepten in Bayesiaanse statistiek, en geeft praktische richtlijnen voor de vier fases van Bayesiaanse gevolgtrekking: plannen, uitvoeren, interpreteren, en rapporteren. Elke fase wordt uitgelegd aan de hand van een voorbeeld dataset. Het doel van dit hoofdstuk is om een breed spectrum te dekken van statistische analyses waarbij hypothesetoetsing, parameterschatting, of beiden het doel is/zijn. Vooral de planningfase is relevant voor zowel Bayesiaanse als niet-Bayesiaanse analyses. Er is een grote nadruk op het vastleggen van de keuzes die gemaakt worden tijdens het plannen door middel van preregistratie (i.e., Nosek et al., 2018; C. D. Chambers, 2013), om transparante en reproduceerbare wetenschap te faciliteren.

Hoofdstuk 3 demonstreert hoe statistische gevolgtrekking intrinsiek subjectieve componenten bevat. Er werden twee relatief simpele onderzoeks scenario’s (een associatie tussen twee continue variabelen, en een kruistabel) voorgelegd aan vier teams van statistici, waarbij elk team de opdracht kreeg om antwoord te geven op de centrale onderzoeksvraag. Zelfs in deze simpele scenario’s werden vrij uiteenlopende statistische methoden gekozen om antwoord te geven. Ook de twee teams die kozen voor een Bayesiaanse analyse, verschilden in hun specifieke aanpak (bijvoorbeeld het toepassen van een log-transformatie op de data). De andere twee teams kozen voor andere methoden, zoals *equivalence testing* en p-waarden. Ondanks de verschillen in aanpak, kwamen alle teams tot de conclusie dat de data niet eenduidig genoeg waren. Deze conclusie was verrassend, omdat beide scenario's uit gepubliceerde artikelen afkomstig waren. De verschillen in methode, maar overeenkomst in conclusie, benadrukt het centrale idee dat een goed doordachte statistische aanpak (zoals het van tevoren plannen welke analysestappen nodig zijn) centraal staat in statistische gevolgtrekking, en dat de specifieke details van de methode op de tweede plaats komen.

Hoofdstuk 4 beschrijft een onderwijsmethode voor het introduceren van Bayesiaanse statistiek op het beginnersniveau. De binomiale toets is een uitstekend beginpunt voor een Bayesiaanse introductie, omdat deze simpel te gebruiken is dankzij de beperkte parameterruimte, de conjugate prior- en posterior verdeling, en de brede toepassing in psychologisch onderzoek. Het hoofdstuk demonstreert de toets aan de hand van een bierproefexperiment aan de Universiteit van Amsterdam, maar de onderzoeksvraag en data kunnen gemakkelijk anders ingevuld worden. Ik heb deze methode gebruikt in mijn eigen onderwijs, in de bachelorcursus "Research Methods and Statistics", waar het schatten en toetsen van de proportie van chocolade pepernoten (de cursus vond plaats rond Sinterklaas) een leuke en leerzame ervaring was.

Hoofdstuk 5 rapporteert een vragenlijst die gestuurd is naar eerste auteurs van empirische artikelen die zijn gepubliceerd in het academisch tijdschrift *Nature Human Behavior*, om in kaart te te brengen hoe toegepaste onderzoekers denken over het begrip van bewijs voor een bepaalde bewering. Dit hoofdstuk beschreef de resultaten van twee vragen, die bij de onderzoekers peilden hoe plausibel zij hun bewering achtten, voordat ze hun data bekeken, en nadat ze hun data bekeken. De antwoorden op deze vraag maken het mogelijk om een informele Bayes factor te berekenen voor de empirische bewering van elke deelnemer, waarbij de Bayes factors vooral duidden op "anekdotisch" bewijs voor de bewering. Omdat deze beweringen allemaal gepubliceerd zijn in een hoogstaand tijdschrift, is dit resultaat nogal verrassend. Het verschil tussen de zelfgerapporteerde overtuiging van de onderzoeker, en de overtuigdheid waarmee een artikel geschreven wordt, toont een groot probleem aan in het wetenschappelijke systeem, waar transparantie over onzekerheid gestraft wordt, in plaats van beloond.

Hoofdstuk 6 beschrijft verschillende keuzes die naar voren komen wanneer een Bayesiaanse vergelijking wordt gedaan voor *mixed effects models*. Het hoofdstuk demonstreert hoe de keuze van nul- en alternatief model de definitie beïnvloedt van "een random effect". Ook worden het effect van het aggregeren van data, de keuze voor prior verdelingen, en de rol van meetfout besproken in de context van *mixed effects model* selectie. Het doel van dit hoofdstuk is het samenbrengen van verschillende vraagstukken, en om concrete vraagstellingen te presenteren, zodat er een set van leidende principes samengesteld kan worden op basis van de discussie die hieruit voortkomt. Het hoofdstuk bespreekt het gedrag van drie verschillende modelvergelijkingen (i.e., een combinatie van nul- en alternatief model), in drie verschillende voorbeeldscenario's, die elk de verschillen tussen de vergelijkingen onderstrepen. In plaats van een "beste" modelvergelijking voor te stellen, is het doel van dit hoofdstuk om de lezer te informeren over deze keuzes, zodat zij zelf in staat zijn de beste aanpak te kiezen voor hun onderzoeksvraag.

C.2 Deel II: For Ranks

Hoofdstuk 7 introduceert een Bayesiaans kader voor hypothesetoetsing en parameterschatting voor de rangcorrelatie Kendall's τ. Dit kader is gebaseerd op werk van Johnson (2005), waar een bovengrens kan worden berekend voor BF_{10} (i.e., een ondergrens voor BF_{01}), door de asymptotische steekproevenverdeling van de toetsstatistiek te modelleren. Om een standaard prior verdeling op te stellen voor Kendall's τ, werd de standaard prior verdeling voor Pearson's ρ te transformeren met Greiner's relation (Greiner, 1909). Bovendien wordt de alternatieve hypothese zo aangepast dat het, gecombineerd met de standaard prior verdeling, mogelijk is om een Bayes factor en posterior verdeling voor Kendall's τ te berekenen.

Hoofdstuk 8 introduceert een tweede kader voor Bayesiaanse gevolgtrekking voor Kendall's τ. Dit nieuwe kader gebruikt *data augmentation*, en houdt in dat de rang data beschouwd worden als ordinale manifestaties van een latent (i.e., niet geobserveerd), normaal verdeeld construct. Door middel van Markov chain Monte Carlo sampling, kan dit latente construct benaderd worden in de vorm van een posterior verdeling voor de latente waarden, waarbij de ordinale informatie van de oorspronkelijke data behouden blijft. Dit Gibbs sampling algoritme benadert ook de posterior verdeling van de parametrische correlatie coefficient Pearson's ρ, welke getransformeerd kan worden naar een posterior verdeling voor Kendall's τ middels Greiner's relation. Deze methode leidt tot erg vergelijkbare resultaten als de methode uit Hoofdstuk 7, behalve in scenario's met lage steekproefomvang en/of een hoge waarde voor de geobserveerde Kendall's τ. In deze gevallen werkt de huidige methode beter, omdat de asymptotische benadering van de eerste methode hier tekort schiet.

Hoofdstuk 9 past hetzelfde latent normale algoritme toe als in Hoofdstuk 8, maar presenteerde dit kader als een algemene methode voor het mogelijk maken van Bayesiaanse gevolgtrekking voor op rang gebaseerde statistische toetsen. Dit kader werd toegepast op de Mann-Whitney U toets, de Wilcoxon rangtekentoets, en Spearman's ρ. Het basisprincipe van de *data augmentation* methode is het gebruiken van de ordinale informatie in de data, om zo het latente construct te benaderen. Hierna kan de parametrische toets (bijvoorbeeld de Bayesiaanse t-toets, in het geval van de Mann-Whitney U toets), toegepast worden op de latente waarden. Deze procedure neemt de onzekerheid mee die inherent is aan rangorde data, terwijl het ook een gemakkelijk te begrijpen procedure is, omdat het dezelfde parameterisatie heeft als de parametrische toets. Bijvoorbeeld, de Bayesiaanse Mann-Whitney U toets geeft een posterior verdeling voor de effectgrootte δ, maar dan op het latente niveau. Bovendien hebben deze toetsen een asymptotische efficiëntie dichtbij de 1 wanneer de data normaal verdeeld is, wat betekent dat beide toetsen ongeveer even krachtig zijn. Naarmate de data minder normaal verdeeld zijn, is de rangordetoets krachtiger dan de parametrische toets.

Hoofdstuk 10 behandelt geen hypothesetoetsing of parameterschatting, maar presenteert Kendall's τ juist als modelleergereedschap voor de psychologische wetenschap. Kendall's τ wordt vaak gebruikt om een associatie tussen twee variabelen te kwantificeren, maar kan ook gebruikt worden om geobserveerde rangorde data op veel verschillende manieren te aggregeren in de vorm van Kendall's *distance*. Dit hoofdstuk presenteert vier uitbreidingen voor Kendall's distance die gebruikt kunnen worden voor het modelleren van psychologische fenomenen, zoals item gelijkheid, item belangrijkheid, en item positie. Elke uitbreiding wordt gedemonstreerd aan de hand van een onderzoeksscenario uit de literatuur.

Appendix D

Acknowledgments

If we view this dissertation as the culmination of a crazy, 32 year long journey, I would have hopelessly stranded an approximate 3,471 times if it were not for all the wonderful people at the various stages of my life. Although I would much rather do this in person, and I cannot wait for that to happen, I want to take a brief moment to express my deep gratitude to those involved.

First off, I would like to thank my three supervisors, who have each raised me academically in their own way. EJ, your infinite compassion, enthusiasm (for both statistics and tiny dancing shoes), and academic playfulness have made this project the most educational and enjoyable academic endeavor of my life. On top of that you founded JASP, which will single-handedly save humanity.[citation needed]Maarten, the fact that I could come with you with questions about obscure MCMC-algorithms, as well as obscure 90s hip-hop always brought me infinite joy and comfort. Lexi, your guerrilla mathematics lectures and awful jokes have been sorely missed these past 2 years and will hopefully make a comeback soon. You are always ready to help – whether that is illuminating some statistical concept or saving me from homelessness.

It has been a great pleasure waddling around in the PML department and I am incredibly grateful for its exquisite collection of the sweetest nerds possible - even the non-Bayesians. The Friday drinks, coffee chats, and random acts of soccer/yoga/sunbathing provided much-needed warmth that the research could not always provide. I also thank the JASP developers for their infinite patience with me and my social science level programming skills. Denny, Han, and Dora, thank you for the excellent job of cultivating a liberating and stimulating environment, and for being part of the committee for my defense.

I would also like to express my gratitude to the academics outside of the UvA that have greatly enriched my academic journey. René Diekstra, your intriguing social psychology lectures and books have inspired me to pursue a degree in psychology, and some random question you once asked in class about Cohen's kappa might have awakened something statistical in me. Francis Tuerlinckx, Wolf Vanpaemel, and Dries Debeer, your wonderfully zen explanations of statistics and your kind Belgian charm have made me truly fall in love with statistics and teaching. Michael Lee and Alex Etz, every time you come visit, it feels like a little festival at the lab. You always seem to strike a perfect balance between research talk and literally anything else, which is both refreshing and inspiring. Irene Klugkist, Francis, and Michael, thank you for being part of the committee for my defense and taking your time to wrestle with this dissertation.

I have moved around from city to city quite a bit during my studies, but each time I thankfully stumbled into all sorts of wonderful people that made me feel right at home, guiding me to new experiences, and who made each stage a tremendously enjoyable adventure, from Eindhoven, to Middelburg, Leuven, and finally Amsterdam. As the final challenge of this thesis, I would like to thank the people of these various stages of my life, in more or less chronological order.

Thank you guys and girl of the J-crew, for being "onnozele ganzen" with me in our teenage years, introducing me to dungeons & dragons, long hair, and feelings of superiority derived from musical taste. Joris, you helped me see the real forests. You have been a cornerstone of my life for over 50% of it by now, and it has been an absolute pleasure evolving together, and staying exactly the same together. Karel, you taught me what it means to be a student – both in partying and eagerness to learn and explore. Even though our India trip nearly meant the end of me, it became a defining experience in my life, thanks to you. Diederik, Bart, and Luuk, thank you for accommodating the quintessential elements of student life - super smash bros, terrible movies, and great bromances. Kris, your adventurousness and lovingness were incredibly inspiring and I will always cherish our times together.

When I moved to exotic Leuven, I was very grateful to befriend Manouk. You packed enough Dutchness in you to make our whole clique feel right at home and I greatly enjoyed our wine & gossip nights. After a while I managed to even trick some Belgians into befriending me, and nothing since has made me prouder. Frits, Hendrik, Christine, Jan, Bram, and Kevin, you are all such incredible sweethearts. Thanks to you I experienced the real soul of Belgium. The countless nights at the Oude Markt, though a bit foggy, filled with delicious beer and even more delicious conversations have a very special place in my heart. Lastly, I want to thank the holy trinity of housemates. Niels, Ike, and Napoleon, our borderline hobo adventures at times were magical, and I have seldom met such skilled chillers and dreamers. Niels, I am glad you are still closeby and I hope we have many more Martian adventures.

Finally, there is Amsterdam and its people. As a new arrival, and being practically Belgian at that point, I was very fortunate to meet my new family among the international students. Ale, Çeren, Maien, Adela, and Leonie, it was amazing to discover this beautiful city (but mostly the inside of Ale's room) together, and I am very glad that we are still here together, six years later. Thank you to Tessa, Riet, Ravi, Boris, Marie, Ruben, Sacha, and Joost for easing me slowly back into Dutch culture. Sjoerd, your inferior Mario kart skills have truly helped me successfully complete this dissertation. Us being neighbors has been, "probabilistically speaking", the greatest coincidence I know. It always is a tremendous pleasure to dabble in varying states of waviness, melancholy, or just sheer idiocy with you. Haley, you are probably the only person that could have convinced me to go to the Keukenhof, and what a wonderful day that turned out to be. Meeting you, and our subsequent adventures have been a great blessing and broadening of horizons.

At the start of my PhD I had the amazing luck to meet and befriend what would later become my paranymphs, and I cannot imagine a better pair. I could not have completed this PhD without either one of you. Jonas, what a prime specimen you are – idiotic genius, ecstatic stoic, compassionate critic. The way you manage to mix extremes in such beautiful ways is straight-up magic and I feel very blessed to be your friend. Quentin, you are the sweetest, gentlest friend, which makes your sporadic sarcasm even more legendary and devastating. I absolutely cherish our soccer sessions and the hundredfold increase in your swagger they cause, and hope we will have many more in the future. Myrte, I am incredibly grateful for the times we have had and the great experiences we have shared, you have helped me broaden horizons I did not even know existed.

Next is a collection of clubs, that somehow always consist of the same marvelous idiots. Whether it is reading, hiking, or cooking, we find some way to enjoy it in some unconventional way. Giacomo, your seemingly endless supply of insane stories that you then shrug off with a marvelling acceptance never ceases to amaze and inspire me. Joris, your compassion knows no bounds, and you always seem to be ready to carry the world on your shoulders. Damiano, your infinite enthusiasm shines so bright in Tilburg I can still see it from my balcony in Amsterdam. Fabian, your ability to truly appreciate the finer things in life, whether a book or a party, is highly contagious. Javi, having some knowledge, but being able to bluff the rest, is the best recipe for adventure, and I will gladly join you for one any time. Lucia, you made me realize what it means to be with your soulmate/love of your life/whatever you want to call it. Knowing we will spend the rest of our lives together brings me true peace. Your sentence has been the hardest to write because every time I see you I already get the unstoppable urge to bombard you with all the sugary clichés I can come up with.

Als laatste wil ik mijn dank en liefde betuigen aan de mensen die mij daadwerkelijk groot gebracht hebben – mam, Bonnie, pap. Jullie zijn altijd een enorme bron van inspiratie, positieve energie, en motivatie geweest door me te laten zien hoe je hard kan werken en hard kan genieten. Hoe gek het leven soms ook kan lopen, en waar ik me ook heb bevonden, ik heb altijd geweten dat er in Eindhoven lekker slap geouwehoer, een boswandeling, en heel veel liefde te vinden is. Ik ben onzettend dankbaar met jullie als familie en ik had dit niet zonder jullie alledrie kunnen doen.