

Bayesian Inference for Psychology

A free book

Edited by Joachim Vandekerckhove, Dora Matzke, and Eric-Jan Wagenmakers

Draft of September 24, 2017

0.1.0

Bayesian Inference for Psychology

Edited by Joachim Vandekerckhove,
Dora Matzke, and Eric-Jan Wagenmakers

September 24, 2017

With most excellent contributions by

Beth Baribault	Annelies Bartlema	Udo Boehm
Bruno Boutin	Scott D. Brown	Pete Cassey
Tim de Jong	Fabian Dablander	Koen Derkx
Zoltan Dienes	Damian Dropmann	Peter A. Edelsbrunner
Sacha Epskamp	Alexander Etz	Quentin F. Gronau
Julia Haaf	Tahira Jamil	Patrick Knight
Michael D. Lee	Jonathon Love	Alexander Ly
Maarten Marsman	Dora Matzke	Neil McLatchie
Frans Meerhoff	Edgar C. Merkle	Richard D. Morey
Aakash Raj	Jeffrey N. Rouder	Isa Rutten
Alexandra Sarafoglou	Ravi Selker	Helen Steingroever
Martin Šmíra	Francis Tuerlinckx	Don van den Bergh
Johnny van Doorn	Erik-Jan van Kesteren	Don van Ravenzwaaij
Joachim Vandekerckhove	Wolf Vanpaemel	Josine Verhagen
Wouter Voorspoels	Eric-Jan Wagenmakers	Ting Wang

This volume is produced by the *Cognition and Individual Differences Lab* at the University of California, Irvine (cidlab.com). The editors make no claim to any intellectual property that may belong to the contributors or their institutions.

Contents

I Bayesian inference for psychology	1
1 Introduction to Bayesian inference for psychology	3
1.1 Introduction	3
1.1.1 What is probability?	4
1.1.2 The Product and Sum Rules of Probability	7
1.2 What is Bayesian inference?	9
1.2.1 Bayes' Rule	10
1.2.2 The prior predictive probability $P(X)$	11
1.2.3 Quantifying evidence	11
1.3 Probability theory in the continuous case	19
1.3.1 Estimating the mean of a normal distribution	26
1.4 Model comparison	30
1.5 Broader appeal and advantages of Bayesian inference	46
1.6 Conclusion	48
References	49
2 Theoretical advantages and practical ramifications	55
2.1 Bayesian Inference and its Benefits	58
2.1.1 Bayesian Parameter Estimation	59
2.1.2 Benefits of Bayesian Parameter Estimation	60
2.1.3 Bayesian Hypothesis Testing	68
2.1.4 Benefits of Bayesian Hypothesis Testing	71
2.2 Ten Objections to the Bayes Factor Hypothesis Test	77
2.2.1 Objection 1: Estimation is Always Superior to Testing	78
2.2.2 Objection 2: Bayesian Hypothesis Tests Can Indicate Evidence for Small Effects That Are Practically Meaningless	80
2.2.3 Objection 3: Bayesian Hypothesis Tests Promote Binary Decisions	81

2.2.4	Objection 4: Bayesian Hypothesis Tests Are Meaningless Under Mis-specification	81
2.2.5	Objection 5: Vague Priors are Preferable over Informed Priors	82
2.2.6	Objection 6: Default Priors are not Sufficiently Subjective	82
2.2.7	Objection 7: Subjective Priors are not Sufficiently Objective	84
2.2.8	Objection 8: Default Priors are Prejudiced Against Small Effects	84
2.2.9	Objection 9: Increasing Sample Size Solves All Statistical Problems	85
2.2.10	Objection 10: Bayesian Procedures Can be Hacked Too	85
2.3	Concluding Comments	86
	References	87
3	Example applications with JASP	99
3.1	The JASP Philosophy	100
3.2	Future Directions for Bayesian Analyses in JASP	121
3.3	Concluding Comments	122
	References	123
4	Parameter estimation in nonstandard models	129
4.1	Introduction	129
4.2	An Introduction with Linear Regression	130
4.2.1	Specification of Models as Generative Processes	130
4.2.2	A Toy Data Set	131
4.2.3	Implementing a Generative Model	131
4.3	WinBUGS Graphical User Interface	133
4.4	JAGS and Stan Command-Line Interface	137
4.5	Working from MATLAB	138
4.6	Working from R	143
4.6.1	Interacting with WinBUGS: R2WinBUGS	143
4.6.2	Interacting with JAGS: R2jags	146
4.6.3	Interacting with Stan: rstan	148
4.7	Example: Multinomial Processing Tree for Modeling False-Memory Data	151
4.7.1	Multinomial Processing Tree Models	152
4.7.2	Working from R using R2WinBUGS	154
4.7.3	Working from R using R2jags	156
4.7.4	Working from R using rstan	157
4.7.5	Working from MATLAB using Trinity	159
4.8	Conclusion	162
	References	164
5	Parameter estimation and Bayes factors	167
5.1	Posterior Estimation	169
5.2	Bayes Factors	171
5.3	Unification	173
5.4	Which Model Specification To Use?	176

5.5	The Potential of Spike-And-Slab Models In Psychology	178
5.6	Conclusions	179
	References	179
II	Teaching resources	181
6	Four reasons to prefer Bayesian over orthodox statistical analyses	183
6.1	Introduction	183
6.1.1	The nature of hypothesis testing	183
6.1.2	The anatomy of a Bayes factor	185
6.1.3	The model of \mathcal{H}_1	185
6.1.4	Putting it together: the meaning of a Bayes factor	189
6.2	Case Studies	190
6.2.1	Often significance testing will provide adequate answers	190
6.2.2	A high powered non-significant result is not necessarily sensitive . . .	191
6.2.3	A low-powered non-significant result is not necessarily insensitive . .	191
6.2.4	A high-powered significant result is not necessarily evidence for a theory	192
6.2.5	The answer to the question should depend on the question	194
6.3	Discussion	195
	References	197
7	How to become a Bayesian in eight easy steps: An annotated reading list	203
7.1	Introduction	203
7.2	Theoretical sources	204
7.2.1	Conceptual introduction: What is Bayesian inference?	205
7.2.2	Bayesian credibility assessments	207
7.2.3	Implications of Bayesian statistics for experimental psychology . . .	208
7.2.4	Structure and motivation of Bayes factors	210
7.3	Applied sources	213
7.3.1	Bayesian model comparison methods	214
7.3.2	Bayesian estimation	215
7.3.3	Prior elicitation	216
7.3.4	Bayesian cognitive modeling	217
7.4	Conclusion	218
	References	219
III	Advanced topics	231
8	Determining informative priors for cognitive models	233
8.1	Introduction	233
8.2	Three illustrative cognitive models	235
8.2.1	Exponential decay model of memory retention	235

8.2.2	Generalized Context Model of categorization	236
8.2.3	Wiener diffusion model of decision making	237
8.3	Sources for determining informative priors	239
8.3.1	Psychological and other scientific theory	239
8.3.2	Logic and invariances	241
8.3.3	Previous data and modeling	242
8.3.4	Elicitation	243
8.4	Methods for determining informative priors	244
8.4.1	Constraint satisfaction	244
8.4.2	Prior prediction	246
8.4.3	Hierarchical extension	248
8.5	Benefits of informative priors	250
8.6	Discussion	251
	References	253
9	Introduction to Markov Chain Monte–Carlo Sampling	259
9.1	Introduction	259
9.2	Example: In-Class Test	261
9.2.1	Limitations	262
9.3	MCMC Applied to a Cognitive Model	265
9.4	Sampling Beyond Basic Metropolis–Hastings	266
9.4.1	Gibbs Sampling	266
9.4.2	Differential Evolution	267
9.5	Summary	271
	References	271
10	Bayesian latent variable models for the analysis of experimental psychology data	279
10.1	Models	280
10.1.1	Factor Model	280
10.1.2	Structural Equation Models	284
10.2	Model Comparison: Bayes Factor Computation	286
10.2.1	Laplace Approximation	286
10.2.2	Savage-Dickey Density Ratio	287
10.3	Application: Risky Choice	288
10.3.1	Methods	289
10.3.2	Results	292
10.3.3	Discussion	295
10.4	General Discussion	297
	References	299

Prologue

The present volume is a collection of previously published papers on the topic of Bayesian inference. The volume is intended to be a low-threshold introduction, requiring only basic knowledge of mathematics and statistics, but also a practical guide, providing relevant background knowledge and pointers to tools for use in everyday research.

Each chapter of this volume is independently published as a peer-reviewed article in a special issue of *Psychonomic Bulletin and Review*, a Springer journal. The contents of these articles are reproduced here according to the terms of the license under which they were published. The content of each article is reproduced exactly, with cosmetic changes only.

It's the 21st century, and excuses for failing to use Bayesian inference grow scant. It is hoped that after reading *Bayesian Inference for Psychology*, there will be none.

Joachim Vandekerckhove
Dora Matzke
Eric-Jan Wagenmakers

I

Bayesian inference for psychology

1

Introduction to Bayesian inference for psychology

Alexander Etz and Joachim Vandekerckhove

Dark and difficult times lie ahead. Soon we must all face the choice between what is right and what is easy.

A. P. W. B. Dumbledore

Introduction

Bayesian methods by themselves are neither dark nor, we believe, particularly difficult. In some ways, however, they are radically different from classical statistical methods and as such, rely on a slightly different way of thinking that may appear unusual at first. Bayesian estimation of parameters will usually not result in a single estimate, but will yield a range of estimates with varying plausibilities associated with them; and Bayesian hypothesis testing will rarely result in the falsification of a theory but rather in a redistribution of probability between competing accounts.

Bayesian methods are also not new, with their first use dating back to the 18th century. Nor are they new to psychology: They were introduced to the field over 50 years ago, in what today remains a remarkably insightful exposition by Ward Edwards, Harold Lindman, and L. J. Savage (1963).

Nonetheless, until recently Bayesian methods have not been particularly mainstream in the social sciences, so the recent increase in their adoption means they are new to most practitioners – and for many psychologists, learning about new statistical techniques can evoke understandable feelings of anxiety or trepidation. At the same time, recent revelations regarding the reproducibility of psychological science (e.g. Open Science Collaboration, 2015; Etz & Vandekerckhove, 2016) have spurred interest in the statistical methods that find use in the field.

In the present article, we provide a gentle technical introduction to the rest of the special issue, starting from first principles. We will first provide a short overview involving the definition of probability, the basic laws of probability theory (the *product* and *sum* rules of probability), and how Bayes' rule and its applications emerge from these two simple laws. We will then illustrate how the laws of probability can and should be used for *inference*: to draw conclusions from observed data. We do not shy away from showing formulas and mathematical exposition, but where possible we connect them to a visual aid, either in a figure or a table, to make the concepts they represent more tangible. We also provide examples after each main section to illustrate how these ideas can be put into practice. Most of the key ideas outlined in this paper only require mathematical competence at the level of college algebra; as will be seen, many of the formulas are obtained by rearranging equations in creative ways such that the quantity of interest is on the left hand side of an equality.

At any point, readers more interested in the bigger picture than the technical details can safely skip the equations and focus on the examples and discussion. However, the use of verbal explanations only suffices to gain a superficial understanding of the underlying ideas and implications, so we provide mathematical formulas for those readers who are interested in a deeper appreciation. Throughout the text, we occasionally use footnotes to provide extra notational clarification for readers who may not be as well-versed with mathematical exposition.

While we maintain that the mathematical underpinnings serve understanding of these methods in important ways, we should also point out that recent developments regarding Bayesian statistical software packages (e.g., Wagenmakers, Love, et al., this volume; Matzke, Boehm, & Vandekerckhove, this volume; van Ravenzwaaij, Cassey, & Brown, this volume; Wagenmakers, Marsman, et al., this volume) have made it possible to perform many kinds of Bayesian analyses without the need to carry out any of the technical mathematical derivations. The mathematical basis we present here remains, of course, more general.

First, however, we will take some time to discuss a subtle semantic confusion between two interpretations of the key concept “probability.” The hurried reader may safely skip the section that follows (and advance to “The Product and Sum Rules of Probability”), knowing only that we use the word “probability” to mean “a degree of belief”: a quantity that indicates how strongly we believe something to be true.

What is probability?

Throughout this text, we will be dealing with the concept of *probability*. This presents an immediate philosophical problem, because the word “probability” is in some sense ambiguous: it will occasionally switch from one meaning to another and this difference in meaning is sometimes consequential.

In one meaning—sometimes called the *epistemic* interpretation—probability is a *degree of belief*: it is a number between zero and one that quantifies how strongly we should think something to be true based on the relevant information we have. In other words, probability is a mathematical language for expressing our uncertainty. This kind of probability is inherently subjective—because it depends on the information that *you* have available—and reasonable people may reasonably differ in the probabilities that they assign to events (or propositions).

Under the epistemic interpretation, there is hence no such thing as *the* probability—there is only *your* probability (Lindley, 2000). Your probability can be thought of as characterizing your state of incomplete knowledge, and in that sense probability does not exist beyond your mind.

We may for example say “There is a 60% probability that the United Kingdom will be outside the European Union on December 31, 2018.” Someone who believes there is a 60% probability this event will occur should be willing to wager *up to* \$6 against \$4 on the event, because their expected gain would be $at\ least\ 60\% \times (+4\$) + 40\% \times (-6\$)$, which is zero. In other words, betting more than \$6 would be unsound because they would expect to lose money, and to take such an action would not *cohere* with what they believe. Of course, in scientific practice one is rarely forced to actually make such bets, but it would be unfortunate if our probabilities (and hence our inferences) could not be acted on with confidence if such an occasion were to arise (Hill, 1974).

The fact that epistemic probabilities of events are subjective does not mean that they are *arbitrary*. Probabilities are not acts of will; they are subjective merely in the sense that they may differ from one individual to the next. That is just to say that different people bring different information to a given problem. Moreover, if different people update their beliefs in a rational way, then as data accumulate they will gradually approach agreement (unless they have a priori ruled out the point of agreement entirely; see, e.g., Jern, Chang, & Kemp, 2014). In fact, it can be shown that the only way that our pre-data beliefs (whatever those may be) will cohere with our post-data beliefs is to use probability to represent our uncertainty and update our beliefs according to the laws of probability (Lindley, 2000).

In another meaning—the *physical* or *aleatory*¹ interpretation—probability is a statement of an *expected frequency over many repetitions of a procedure*. A statement of aleatory probability might be “If I flip a fair coin very many times, the ratio of flips on which the coin will come up heads is 50%. Thus, the probability that a fair coin will come up heads is 50%.” These statements express properties of the *long-run behavior* of well-defined processes, but they can not speak to singular events; they require assumptions about physical repeatability and independence among repetitions. It is important to grasp that these frequencies are seen as being a real part of the physical world, in that “the relative frequencies of a die falling this way or that way are ‘persistent’ and constitute this die’s measurable properties, comparable to its size and weight” (Neyman, 1977, p. 99). Neyman’s quote provides an interesting contrast to the epistemic interpretation. Italian probabilist and influential Bayesian statistician Bruno de Finetti famously began his treatise *Theory of Probability* by stating “Probability does not exist” and that “the abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, . . . or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs” (de Finetti, 1974, p. x). This is not to say that we cannot build models that assign probabilities to the outcomes of physical processes, only that they are necessarily abstractions.

It is clear that these two interpretations of probability are not the same. There are many

¹From Latin *alea*, meaning dice.

situations to which the aleatory definition does not apply and thus probabilities could not be determined: we will not see repeated instances of December 31, 2018, in which the UK could be inside or outside the EU, we will only see one such event. Similarly, “what is the probability that *this coin, on the very next flip, will come up heads?*” is not something to which an aleatory probability applies: there are no long-run frequencies to consider *if there is only one flip that matters*.

Aleatory probability may—in some cases—be a valid *conceptual* interpretation of probability, but it is rarely ever an *operational* interpretation (see Jaynes, 1984; Winkler, 1972; Wrinch & Jeffreys, 1919): it cannot apply to singular events such as the truth or falsity of a scientific theory, so we simply cannot speak of aleatory probabilities when wrestling with the uncertainty we face in scientific practice. That is to say, we may validly use aleatory probability to *think about* probability in an abstract way, but not to make statements about real-world observed events such as experimental outcomes.

In contrast, epistemic probability applies to any event that we care to consider—be it singular or repetitive—and if we have relevant information about real-world frequencies then we can choose to use that information to inform our beliefs. If repetition is possible and we find it reasonable to assume that the chance a coin comes up heads on a given toss does not change based on the outcome of previous tosses, then a Bayesian could reasonably believe both (a) that on the next toss there is a 50% chance it comes up heads; *and* (b) 50% of tosses will result in heads in a very long series of flips. Hence, epistemic probability is both a *conceptual* interpretation of probability and an *operational* interpretation. Epistemic probability can be seen as an extension of aleatory probability that applies to all the cases where the latter would apply and to countless cases where it could not.

Why this matters

We argue that the distinction above is directly relevant for empirical psychology. In the overwhelming majority of cases, psychologists are interested in making probabilistic statements about singular events: *this* theory is either true or not; *this* effect is either positive or negative; *this* effect size is probably between x and y ; and either *this* model or the other is more likely given the data. Seldom are we merely interested in the frequency with which a well-defined process will achieve a certain outcome. Even arbitrarily long sequences of faithful replications of empirical studies serve to address a *singular* question: “*is this* theory correct?” We might reasonably define a certain behavioral model and assign parameters (even parameters that are probabilities) to it, and then examine its long-run behavior. This is a valid aleatory question. However, it is not an inferential procedure: it is describing behavior of an idealized model but not making inferences with regard to that model. We might also wonder how frequently a researcher will make errors of inference (however defined) under certain conditions, but this is a purely academic exercise; unless the proportion of errors is 0 or 1, such a long-run frequency alone does not allow us to determine the probability the researcher actually made an error regarding any *singular* finding – regarding *this* coin, *this* effect, or *this* hypothesis. By contrast, epistemic probability expresses degrees of belief regarding specific, individual, *singular* events, and for that reason should be the default for scientific inference.

In the next section, we will introduce the basic rules of probability theory. These rules are agnostic to our conception of probability—they hold equally for epistemic and aleatory probability—but throughout the rest of this paper and particularly in the examples, we will, unless otherwise noted, use an epistemic interpretation of the word “probability.”

The Product and Sum Rules of Probability

Here we will introduce the two cardinal rules of probability theory from which essentially all of Bayesian inference derives. However, before we venture into the laws of probability, there are notational conventions to draw. First, we will use $P(A)$ to denote the probability of some event A , where A is a statement that can be true or false (e.g., A could be “it will rain today”, “the UK will be outside the EU on December 31, 2018”, or “the 20th digit of π is 3”). Next, we will use $(B|A)$ to denote the *conditional* event: the probability that B is true *given that A is true* (e.g., B could be “it will rain tomorrow”) is $P(B|A)$: the probability that it will rain tomorrow given that it rained today. Third, we will use (A, B) to denote a *joint* event: the probability that A and B are both true is $P(A, B)$. The joint probability $P(A, B)$ is of course equal to that of the joint probability $P(B, A)$: the event “it rains tomorrow and today” is logically the same as “it rains today and tomorrow.” Finally, we will use $(\neg A)$ to refer to the negation of A : the probability A is false is $P(\neg A)$. These notations can be combined: if C and D represent the events “it is hurricane season” and “it rained yesterday,” respectively, then $P(A, B|\neg C, \neg D)$ is the probability that it rains today and tomorrow, given that ($\neg C$) it is not hurricane season and that ($\neg D$) it did not rain yesterday (i.e., both C and D are not true).

With this notation in mind, we introduce the **Product Rule of Probability**:

$$\begin{aligned} P(A, B) &= P(B)P(A|B) \\ &= P(A)P(B|A). \end{aligned} \tag{1.1}$$

In words: the probability that A and B are both true is equal to the probability of B multiplied by the conditional probability of A *assuming B is true*. Due to symmetry, this is also equal to the probability of A multiplied by the conditional probability of B *assuming A is true*. The probability it rains today and tomorrow is the probability it first rains today multiplied by the probability it rains tomorrow *given that we know it rained today*.

If we assume A and B are statistically independent then $P(B)$ equals $P(B|A)$, since knowing A happens tells us nothing about the chance B happens. In such cases, the product rule simplifies as follows:

$$P(A, B) = P(A)P(B|A) = P(A)P(B). \tag{1.2}$$

Keeping with our example, this would mean calculating the probability it rains both today and tomorrow in such a way that knowledge of whether or not it rained today has no bearing on how strongly we should believe it will rain tomorrow.

Understanding the **Sum Rule of Probability** requires one further concept: the *disjoint set*. A disjoint set is nothing more than a collection of mutually exclusive events. To simplify the exposition, we will also assume that exactly one of these events must be true although

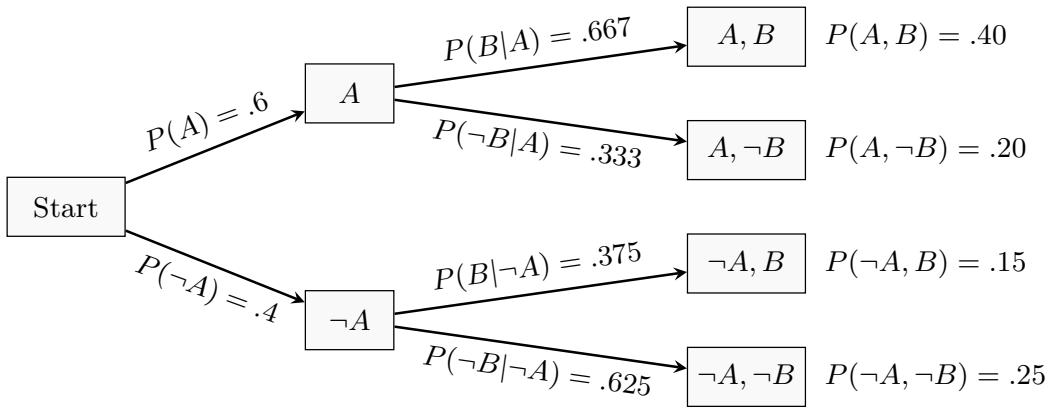


Figure 1.1: An illustration of the Product Rule of probability: The probability of the joint events on the right end of the diagram is obtained by multiplying the probabilities along the path that leads to it. The paths indicate where and how we are progressively splitting the initial probability into smaller subsets. A suggested exercise to test understanding and gain familiarity with the rules is to construct the equivalent path diagram (i.e., that in which the joint probabilities are identical) starting on the left with a fork that depends on the event B instead of A .

that is not part of the common definition of such a set. The simplest example of a disjoint set is some event and its denial:² $\{B, \neg B\}$. If B represents the event “It will rain tomorrow,” then $\neg B$ represents the event “It will not rain tomorrow.” One and only one of these events must occur, so together they form a disjoint set. If A represents the event “It will rain today,” and $\neg A$ represents “It will not rain today” (another disjoint set), then there are four possible pairs of these events, one of which must be true: (A, B) , $(A, \neg B)$, $(\neg A, B)$, and $(\neg A, \neg B)$. The probability of a single one of the singular events, say B , can be found by adding up the probabilities of all of the joint events that contain B as follows:

$$P(B) = P(A, B) + P(\neg A, B).$$

In words, the probability that it rains tomorrow is the sum of two joint probabilities: (1) the probability it rains today and tomorrow, and (2) the probability it does not rain today but does rain tomorrow.

In general, if $\{A_1, A_2, \dots, A_K\}$ is a disjoint set, the Sum Rule of Probability states:

$$\begin{aligned} P(B) &= P(A_1, B) + P(A_2, B) + \dots + P(A_K, B) \\ &= \sum_{k=1}^K P(A_k, B). \end{aligned} \tag{1.3}$$

That is, to find the probability of event B alone you add up all the joint probabilities that involve both B and one element of a disjoint set. Intuitively, it is clear that if one of

²We use curly braces $\{\dots\}$ to indicate a set of events. Other common examples of disjoint sets are the possible outcomes of a coin flip: {heads, tails}, or the possible outcomes of a roll of a six-sided die: {1, 2, 3, 4, 5, 6}. A particularly useful example is the truth of some model \mathcal{M} , which must be either true or false: $\{\mathcal{M}, \neg \mathcal{M}\}$.

Table 1.1: The event A is that it rains today. The event B is that it rains tomorrow. Sum across rows to find $P(A)$, sum down columns to find $P(B)$. One can also divide $P(A, B)$ by $P(A)$ to find $P(B|A)$, as shown in the next section.

	B	$\neg B$	$B \text{ or } \neg B$
A	$P(A, B) = .40$	$P(A, \neg B) = .20$	$\Rightarrow P(A) = .60$
$\neg A$	$P(\neg A, B) = .15$	$P(\neg A, \neg B) = .25$	$\Rightarrow P(\neg A) = .40$
$A \text{ or } \neg A$	$P(B) = .55$	$P(\neg B) = .45$	1.00

$\{A_1, A_2, \dots, A_K\}$ must be true, then the probability that *one of these and B* is true is equal to the base probability that B is true.

An illustration of the Product Rule of Probability is shown by the path diagram in Figure 1.1. Every fork indicates the start of a disjoint set, with each of the elements of that set represented by the branches extending out. The lines indicate the probability of selecting each element from within the set. Starting from the left, one can trace this diagram to find the joint probability of, say, A and B . At the *Start* fork there is a probability of .6 of going along the top arrow to event A (a similar diagram could of course be drawn that starts with B): The probability it rains today is .6. Then there is a probability of .667 of going along the next top fork to event (A, B) : The probability it rains tomorrow given it rained today is .667. Hence, of the initial .6 probability assigned to A , two-thirds of it forks into (A, B) , so the probability of (A, B) is $.6 \times .667 = .40$: Given that it rained today, the probability it rains tomorrow is .667, so the probability it rains both today and tomorrow is .4. The probability of any joint event at the end of a path can be found by multiplying the probabilities of all the forks it takes to get there.

An illustration of the Sum Rule of Probability is shown in Table 1.1, which tabulates the probabilities of all the joint events found through Figure 1.1 in the main cells. For example, adding up all of the joint probabilities across the row denoted A gives $P(A)$. Adding up all of the joint probabilities down the column denoted B gives $P(B)$. This can also be seen by noting that in Figure 1.1, the probabilities of the two child forks leaving from A , namely (A, B) and $(A, \neg B)$, add up to the probability indicated in the initial fork leading to A . This is true for any value of $P(B|A)$ (and $P(\neg B|A) = 1 - P(B|A)$).

What is Bayesian inference?

Together [the Sum and Product Rules] solve the problem of inference, or, better, they provide a framework for its solution.

D. V. Lindley (2000)

Bayesian inference is the application of the product and sum rules to real problems of inference. Applications of Bayesian inference are creative ways of looking at a problem through the lens of these two rules. The rules form the basis of a mature philosophy

of scientific learning proposed by Dorothy Wrinch and Sir Harold Jeffreys (Jeffreys, 1961, 1973; Wrinch & Jeffreys, 1921; see also Ly, Verhagen, & Wagenmakers, 2016). Together, the two rules allow us to calculate probabilities and perform scientific inference in an incredible variety of circumstances. We begin by illustrating one combination of the two rules that is especially useful for scientific inference: Bayesian hypothesis testing.

Bayes' Rule

Call event \mathcal{M} (the truth of) an hypothesis that a researcher holds and call $\neg\mathcal{M}$ a competing hypothesis. Together these can form a disjoint set: $\{\mathcal{M}, \neg\mathcal{M}\}$. The set $\{\mathcal{M}, \neg\mathcal{M}\}$ is necessarily disjoint if $\neg\mathcal{M}$ is simply the denial of \mathcal{M} , but in practice the set of hypotheses can contain any number of models spanning a wide range of theoretical accounts. In such a scenario, it is important to keep in mind that we cannot make inferential statements about any model not included in the set.

Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(\mathcal{M})$ and $P(\neg\mathcal{M})$. The hypotheses are well-defined if they make a specific prediction about the probability of each experimental outcome X through the *likelihood functions* $P(X|\mathcal{M})$ and $P(X|\neg\mathcal{M})$. Likelihoods can be thought of as how strongly the data X are implied by an hypothesis. *Conditional* on the truth of an hypothesis, likelihood functions specify the probability of a given outcome and are usually easiest to interpret in relation to other hypotheses' likelihoods. Of interest, of course, is the probability that \mathcal{M} is true, given the data X , or $P(\mathcal{M}|X)$.

By simple rearrangement of the factors of the Product Rule shown in the first line of Equation 1.1, $P(\mathcal{M}, X) = P(X)P(\mathcal{M}|X)$, we can derive that

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M}, X)}{P(X)}.$$

Due to the symmetric nature of the Product Rule, we can reformulate the joint event in the numerator above by applying the product rule again as in the second line in Equation 1.1, $P(\mathcal{M}, X) = P(\mathcal{M})P(X|\mathcal{M})$, and we see that this is equivalent to

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(X)}. \quad (1.4)$$

Equation 1.4 is one common formulation of **Bayes' Rule**, and analogous versions can be written for each of the other competing hypotheses; for example, Bayes' Rule for $\neg\mathcal{M}$ is

$$P(\neg\mathcal{M}|X) = \frac{P(\neg\mathcal{M})P(X|\neg\mathcal{M})}{P(X)}.$$

The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the prior predictive probability of the observed data (see below). In the way that $P(\mathcal{M})$ and $P(\neg\mathcal{M})$ are called prior probabilities because they capture our knowledge prior to seeing the data X , so $P(\mathcal{M}|X)$ and $P(\neg\mathcal{M}|X)$ are called the *posterior probabilities*.

The prior predictive probability $P(X)$

Many of the quantities in Equation 1.4 we know: we must have some prior probability (belief or prior information) that the hypothesis is true if we are even considering the hypothesis at all, and if the hypothesis is well-described it will attach a particular probability to the observed data. What remains is the denominator: the prior predictive probability $P(X)$ —the probability of observing a given outcome in the experiment, which can be thought of as the average probability of the outcome implied by the hypotheses, weighted by the prior probability of each hypothesis. $P(X)$ can be obtained through the sum rule by adding the probabilities of the joint events $P(X, \mathcal{M})$ and $P(X, \neg\mathcal{M})$, as in Equation 1.3, each of which is obtained through an application of the product rule, so we obtain the following expression:

$$\begin{aligned} P(X) &= P(X, \mathcal{M}) + P(X, \neg\mathcal{M}) \\ &= P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M}), \end{aligned} \quad (1.5)$$

which amounts to adding up the right-hand side numerator of Bayes' Rule for all competing hypotheses, giving a weighted-average probability of observing the outcome X .

Now that we have a way to compute $P(X)$ in Equation 1.5, we can plug the result into the denominator of Equation 1.4 as follows:

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}. \quad (1.6)$$

Equation 1.6 is for the case where we are only considering one hypothesis and its complement. More generally,

$$P(\mathcal{M}_i|X) = \frac{P(\mathcal{M}_i)P(X|\mathcal{M}_i)}{\sum_{k=1}^K P(\mathcal{M}_k)P(X|\mathcal{M}_k)}, \quad (1.7)$$

for the case where we are considering K competing and mutually-exclusive hypotheses (i.e., hypotheses that form a disjoint set), one of which is \mathcal{M}_i .

Quantifying evidence

Now that we have, in one equation, factors that correspond to our knowledge before— $P(\mathcal{M})$ —and after— $P(\mathcal{M}|X)$ —seeing the data, we can address a slightly alternative question: *How much did we learn due to the data X ?* Consider that every quantity in Equation 1.7 is either a prior belief in an hypothesis, or the probability that the data would occur under a certain hypothesis—all known quantities. If we divide both sides of Equation 1.7 by $P(\mathcal{M}_i)$,

$$\frac{P(\mathcal{M}_i|X)}{P(\mathcal{M}_i)} = \frac{P(X|\mathcal{M}_i)}{\sum_{k=1}^K P(\mathcal{M}_k)P(X|\mathcal{M}_k)}, \quad (1.8)$$

we see that after observing outcome X , the ratio of an hypothesis's posterior probability to its prior probability is larger than 1 (i.e., its probability goes up) if the probability it attaches to the observed outcome is greater than a weighted-average of all such probabilities – averaged across all candidate hypotheses, using the respective prior probabilities as weights.

If we are concerned with only two hypotheses, a particularly interesting application of Bayes' Rule becomes possible. After collecting data we are left with the posterior probability of two hypotheses, $P(\mathcal{M}|X)$ and $P(\neg\mathcal{M}|X)$. If we form a ratio of these probabilities we can quantify our *relative belief* in one hypothesis vis-à-vis the other, or what is known as the posterior odds: $P(\mathcal{M}|X)/P(\neg\mathcal{M}|X)$. If $P(\mathcal{M}|X) = .75$ and $P(\neg\mathcal{M}|X) = .25$, the posterior odds are $.75/.25 = 3$, or 3:1 ("three to one") in favor of \mathcal{M} over $\neg\mathcal{M}$. Since the posterior probability of an hypothesis is equal to the fraction in the right-hand side of Equation 1.6, we can calculate the posterior odds as a ratio of two right-hand sides of Bayes' Rule as follows:

$$\frac{P(\mathcal{M}|X)}{P(\neg\mathcal{M}|X)} = \frac{\frac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}}{\frac{P(\neg\mathcal{M})P(X|\neg\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}},$$

which can be reduced to a simple expression (since the denominators cancel out),

$$\underbrace{\frac{P(\mathcal{M}|X)}{P(\neg\mathcal{M}|X)}}_{\text{Posterior odds}} = \underbrace{\frac{P(\mathcal{M})}{P(\neg\mathcal{M})}}_{\text{Prior odds}} \times \underbrace{\frac{P(X|\mathcal{M})}{P(X|\neg\mathcal{M})}}_{\text{Bayes factor}}. \quad (1.9)$$

The final factor—the Bayes factor—can be interpreted as *the extent to which the data sway our relative belief from one hypothesis to the other*, which is determined by comparing the hypotheses' abilities to predict the observed data. If the data are more probable under \mathcal{M} than under $\neg\mathcal{M}$ (i.e., if $P(X|\mathcal{M})$ is larger than $P(X|\neg\mathcal{M})$) then \mathcal{M} does the better job predicting the data, and the posterior odds will favor \mathcal{M} more strongly than the prior odds.

It is important to distinguish Bayes factors from posterior probabilities. Both are useful in their own role – posterior probabilities to determine our total belief after taking into account the data and to draw conclusions, and Bayes factors as a learning factor that tells us how much evidence the data have delivered. It is often the case that a Bayes factor favors \mathcal{M} over $\neg\mathcal{M}$ while at the same time the posterior probability of $\neg\mathcal{M}$ remains greater than \mathcal{M} . As Jeffreys, in his seminal paper introducing the Bayes factor as a method of inference, explains: "If ... the [effect] examined is one that previous considerations make unlikely to exist, then we are entitled to ask for a greater increase of the probability before we accept it," and moreover, "To raise the probability of a proposition from 0.01 to 0.1 does not make it the most likely alternative" (Jeffreys, 1935, p. 221). This distinction is especially relevant to today's publishing environment, where there exists an incentive to publish counterintuitive results – whose very description as counterintuitive implies most researchers would not have expected them to be true. Consider as an extreme example Bem (2011) who presented data consistent with the hypothesis that some humans can predict future random events. While Bem's data may indeed provide positive evidence for that hypothesis (Rouder & Morey, 2011), it is staggeringly improbable a priori and the evidence in the data does not stack up to the strong priors many of us will have regarding extrasensory perception – extraordinary claims require extraordinary evidence.

Since Bayes factors quantify statistical evidence, they can serve two (closely related) purposes. First, evidence can be applied to defeat prior odds: supposing that prior to the

data we believe that $\neg\mathcal{M}$ is three times more likely than \mathcal{M} (i.e., the prior ratio favoring $\neg\mathcal{M}$ is 3, or its prior probability is 75%), we need a Bayes factor favoring \mathcal{M} that is greater than 3 so that \mathcal{M} will end up the more likely hypothesis. Second, evidence can be applied to achieve a desired level of certainty: supposing that we desire a high degree of certainty before making any practical decision (say, at least 95% certainty or a posterior ratio of at least 19) and supposing the same prior ratio as before, then we would require a Bayes factor of $19 \times 3 = 57$ to defeat the prior odds and obtain this high degree of certainty. These practical considerations (often left implicit) are formalized by utility (loss) functions in *Bayesian decision theory*. We will not go into Bayesian decision theory in depth here; introductions can be found in Lindley (1985) or Winkler (1972), and an advanced introduction is available in Robert (2007).

In this section, we have derived Bayes' Rule as a necessary consequence of the laws of probability. The rule allows us to update our belief regarding an hypothesis in response to data. Our beliefs after taking account the data are captured in the *posterior probability*, and the amount of updating is given by the *Bayes factor*. We now move to some applied examples that illustrate how this simple rule pertains to cases of inference.

Example 1: “The happy herbologist”

At Hogwarts School of Witchcraft and Wizardry,³ professor Pomona Sprout leads the Herbology Department (see Illustration). In the Department’s greenhouses, she cultivates crops of a magical plant called *green codacle* – a flowering plant that when consumed causes a witch or wizard to feel euphoric and relaxed. Professor Sybill Trelawney, the professor of Divination, is an avid user of green codacle and frequently visits Professor Sprout’s laboratory to sample the latest harvest.

However, it has turned out that one in a thousand codacle plants is afflicted with a mutation that changes its effects: Consuming those rare plants causes unpleasant side effects such as paranoia, anxiety, and spontaneous levitation. In order to evaluate the quality of her crops, Professor Sprout has developed a mutation-detecting spell. The new spell has a 99% chance to accurately detect an existing mutation, but also has a 2% chance to falsely indicate that a healthy plant is a mutant. When Professor Sprout presents her results at a School colloquium, Trelawney asks two questions: What is the probability that a codacle plant is a mutant, when your spell says that it is? And what is the probability the plant is a mutant, when your spell says that it is healthy? Trelawney’s interest is in knowing how much trust to put into Professor Sprout’s spell.

Call the event that a specific plant is a mutant \mathcal{M} , and that it is healthy $\neg\mathcal{M}$. Call the event that Professor Sprout’s spell diagnoses a plant as a mutant D , and that it diagnoses it healthy $\neg D$. Professor Trelawney’s interest is in the probability that the plant is indeed a mutant given that it has been diagnosed as a mutant, or $P(\mathcal{M}|D)$, and the probability the plant is a mutant given it has been diagnosed healthy, or $P(\mathcal{M}|\neg D)$. Professor Trelawney, who is an accomplished statistician, has all the relevant information to apply Bayes’ Rule (Equation 1.7 above) to find these probabilities. She knows the prior probability that a plant

³With our apologies to J. K. Rowling.

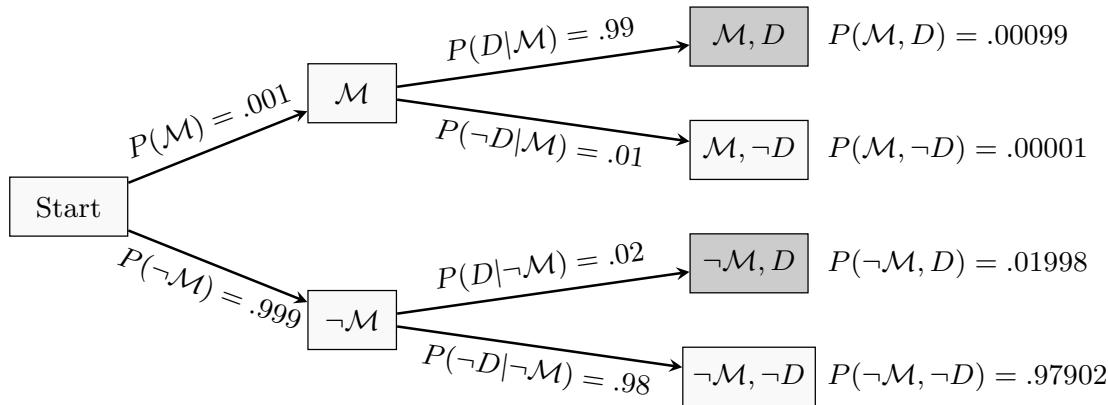


Figure 1.2: The event \mathcal{M} is that a given codacle plant is a mutant. The event D is that Professor Sprout’s spell returns a mutant diagnosis. A mutant diagnosis D is in fact observed, so the only paths that remain relevant are those that lead to a mutant diagnosis (events (\mathcal{M}, D) and $(\neg\mathcal{M}, D)$, shaded). Professor Trelawney takes the following steps to find the posterior probability the plant is a mutant given the mutant diagnosis: Multiply $P(\mathcal{M})$ by $P(D|\mathcal{M})$ to find $P(\mathcal{M}, D)$; multiply $P(\neg\mathcal{M})$ by $P(D|\neg\mathcal{M})$ to find $P(\neg\mathcal{M}, D)$; add $P(\mathcal{M}, D)$ and $P(\neg\mathcal{M}, D)$ to find $P(D)$; divide $P(\mathcal{M}, D)$ by $P(D)$ to find $P(\mathcal{M}|D)$. Professor Trelawney’s question can be rephrased as: of the total probability remaining in the diagram after D is observed – which is equal to $P(D)$ – what proportion of it originated at the \mathcal{M} node? The results of Professor Trelawney’s calculations are given in the text.

is a mutant is $P(\mathcal{M}) = .001$, and thus the prior probability that a plant is not a mutant is $P(\neg\mathcal{M}) = 1 - P(\mathcal{M}) = .999$. The probability of a correct mutant diagnosis given the plant is a mutant is $P(D|\mathcal{M}) = .99$, and the probability of an erroneous healthy diagnosis given the plant is a mutant is thus $P(\neg D|\mathcal{M}) = 1 - P(D|\mathcal{M}) = .01$. When the plant is healthy, the spell incorrectly diagnoses it as a mutant with probability $P(D|\neg\mathcal{M}) = .02$, and correctly diagnoses the plant as healthy with probability $P(\neg D|\neg\mathcal{M}) = 1 - P(D|\neg\mathcal{M}) = .98$.

When Professor Sprout’s spell gives a mutant diagnosis, the posterior probability that the plant is really a mutant is given by Bayes’ Rule:

$$P(\mathcal{M}|D) = \frac{P(\mathcal{M})P(D|\mathcal{M})}{P(\mathcal{M})P(D|\mathcal{M}) + P(\neg\mathcal{M})P(D|\neg\mathcal{M})}.$$

Professor Trelawney can now consult Figure 1.2 to find that the posterior probability the plant is a mutant given a mutant diagnosis is:

$$P(\mathcal{M}|D) = \frac{.001 \times .99}{.001 \times .99 + .999 \times .02} \approx .047.$$

A mutant diagnosis from Professor Sprout’s spell raises the probability the plant is a mutant from $.001$ to roughly $.047$. This means that when a plant is diagnosed as a mutant, the posterior probability the plant is *not* a mutant is $P(\neg\mathcal{M}|D) \approx 1 - .047 = .953$. The low prior probability that a plant is a mutant means that, even with the spell having 99%

accuracy to correctly diagnose a mutant plant as such, a plant diagnosed as a mutant is still probably safe to eat – nevertheless, Professor Trelawney will think twice.

Analogous calculations show that the posterior probability that a plant is a dangerous mutant, given it is diagnosed as healthy, is:

$$P(\mathcal{M}|\neg D) = \frac{.001 \times .01}{.001 \times .01 + .999 \times .98} \approx .000010.$$

The posterior probability that a plant is a dangerous mutant despite being diagnosed as healthy is quite small, so Trelawney can be relatively confident she is eating a healthy plant after professor Sprout's spell returns a healthy diagnosis.

A major advantage of using Bayes' Rule in this way is that it gracefully extends to more complex scenarios. Consider the perhaps disappointing value of $P(\mathcal{M}|D)$: a mutant diagnosis only raises the posterior probability to just under 5%. Suppose, however, that Trelawney knows that Professor Sprout's diagnosis (D_S) is statistically independent from the diagnosis of her talented research associate Neville Longbottom (D_L) – meaning that for any given state of nature \mathcal{M} or $\neg\mathcal{M}$, Longbottom's diagnosis does not depend on Sprout's. Further suppose that both Sprout and Longbottom return the mutant diagnosis (and for simplicity we also assume Longbottom's spells are equally as accurate as Sprout's). To find the posterior probability the plant is a mutant after two independent mutant diagnoses, $P(\mathcal{M}|D_S, D_L)$, Trelawney can apply a fundamental principle in Bayesian inference: **Yesterday's posterior is today's prior** (Lindley, 2000).

Since we take diagnosis D_S and diagnosis D_L as conditionally independent, we know that $P(D_L|\mathcal{M}, D_S) = P(D_L|\mathcal{M})$ and $P(D_L|\neg\mathcal{M}, D_S) = P(D_L|\neg\mathcal{M})$, giving

$$\begin{aligned} P(\mathcal{M}|D_S, D_L) &= \frac{P(\mathcal{M}|D_S)P(D_L|\mathcal{M})}{P(\mathcal{M}|D_S)P(D_L|\mathcal{M}) + P(\neg\mathcal{M}|D_S)P(D_L|\neg\mathcal{M})} \\ &= \frac{.047 \times .99}{.047 \times .99 + .953 \times .02} \approx .71, \end{aligned}$$

where the probability the plant is a mutant *prior to Longbottom's diagnosis* D_L , $P(\mathcal{M}|D_S)$, is the probability it is a mutant *posterior to Sprout's diagnosis* D_S . This illustrates the value of multiple independent sources of evidence: a plant that has twice been independently diagnosed as a mutant is quite likely to be one. A third independent diagnosis would put the posterior probability over 99%. Note that, crucially, we would have obtained precisely the same final probability of .71 had we updated $P(\mathcal{M})$ to $P(\mathcal{M}|D_S, D_L)$ all at once. This is easily confirmed when we consider the two diagnoses as a joint event (D_S, D_L) and use the conditional probability $P(D_S, D_L|\mathcal{M}) = P(D_S|\mathcal{M}) \times P(D_L|\mathcal{M})$ (as in Equation 1.2) to update $P(\mathcal{M})$ to $P(\mathcal{M}|D_S, D_L)$ in a single step.

Discussion It is instructive to consider some parallels of this (admittedly fictional) example to current practices in social science. The scenario is similar in setup to a null-hypothesis significance testing scenario in which one defines the null hypothesis \mathcal{H}_0 (e.g., that there is no effect of some manipulation) and its negation \mathcal{H}_1 (that there is an effect), and the end



Illustration. Professor Pomona Sprout is Chair of the Herbology Department at Hogwarts School of Witchcraft and Wizardry. (c) Brian Clayton, used with permission.

goal is to make a choice between two possible decisions $\{D, \neg D\}$; D means deciding to reject \mathcal{H}_0 , and $\neg D$ means deciding not to reject \mathcal{H}_0 . In the example above the rate at which we falsely reject the null hypothesis (i.e., deciding to reject it when in fact it is true) is given by $P(D|\neg\mathcal{M}) = .02$ – this is what is commonly called the false alarm rate. The rate at which we correctly reject the null hypothesis (i.e., rejecting it if it is false) is $P(D|\mathcal{M}) = .99$. However, even with a low false alarm rate and a very high correct rejection rate, a null hypothesis rejection may not necessarily provide enough evidence to overcome the low prior probability an alternative hypothesis might have.

Example 2: “A curse on your hat”

At the start of every school year, new Hogwarts students participate in the centuries-old Sorting ceremony, during which they are assigned to one of the four Houses of the School: Gryffindor, Hufflepuff, Ravenclaw, or Slytherin. The assignment is performed by the Sorting Hat, a pointy hat which, when placed on a student’s head, analyzes their abilities and personality before loudly calling out the House that it determines as the best fit for the student. For hundreds of years the Sorting Hat has assigned students to houses with perfect accuracy and in perfect balance (one-quarter to each House).

Unfortunately, the Hat was damaged by a stray curse during a violent episode at the School. As a result of the dark spell, the Hat will now occasionally blurt out “Slytherin!” even when the student’s proper alliance is elsewhere. Now, the Hat places exactly 40% of

first-years in Slytherin instead of the usual 25%, and each of the other Houses get only 20% of the cohort.

To attempt to correct the House assignment, Professor Cuthbert Binns has developed a written test—the *Placement Accuracy Remedy for Students Erroneously Labeled* or P.A.R.S.E.L. test—on which true Slytherins will tend to score *Excellent* (S_E), while Ravenclaws will tend to score *Outstanding* (S_O), Gryffindors *Acceptable* (S_A), and Hufflepuffs *Poor* (S_P). Benchmark tests on students who were Sorted before the Hat was damaged have revealed the approximate distribution of P.A.R.S.E.L. scores within each House (see Table 1.2). The test is administered to all students who are sorted into Slytherin House by the damaged Sorting Hat, and their score determines the House to which they are assigned. Headmistress Minerva McGonagall, who is a Gryffindor, asks Professor Binns to determine the probability that a student who was sorted into Slytherin and scored *Excellent* on the P.A.R.S.E.L. test actually belongs in Gryffindor.

Table 1.2: Probability of each P.A.R.S.E.L. score by true House affiliation. Each value indicates the conditional probability $P(S|\mathcal{M})$, that is, the probability that a student from house \mathcal{M} obtains score S .

	Excellent (S_E)	Outstanding (S_O)	Acceptable (S_A)	Poor (S_P)
Slytherin (\mathcal{M}_S)	0.80	0.10	0.05	0.05
Gryffindor (\mathcal{M}_G)	0.05	0.20	0.70	0.05
Ravenclaw (\mathcal{M}_R)	0.05	0.80	0.15	0.00
Hufflepuff (\mathcal{M}_H)	0.00	0.10	0.25	0.65

The solution relies on the repeated and judicious application of the Sum and Product Rules, until an expression appears with the desired quantity on the left hand side and only known quantities on the right hand side. To begin, Professor Binns writes down Bayes' Rule (remembering that a joint event like (D_S, S_E) can be treated like any other event):

$$P(\mathcal{M}_G|D_S, S_E) = \frac{P(\mathcal{M}_G)P(D_S, S_E|\mathcal{M}_G)}{P(D_S, S_E)}$$

Here, \mathcal{M}_G means that the true House assignment is Gryffindor, D_S means that the Sorting Hat placed them in Slytherin, and S_E means the student scored *Excellent* on the P.A.R.S.E.L. test.

In most simple cases, we often have knowledge of simple probabilities, of the form $P(A)$ and $P(B|A)$, while the probabilities of joint events (A, B) are harder to obtain. For Professor Binns' problem, we can overcome this difficulty by using the Product Rule to unpack the joint event in the numerator:⁴

$$P(\mathcal{M}_G|D_S, S_E) = \frac{P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|S_E, \mathcal{M}_G)}{P(D_S, S_E)}.$$

⁴Note that this is an application of the Product Rule to the scenario where both events are conditional on \mathcal{M}_G : $P(D_S, S_E|\mathcal{M}_G) = P(S_E|\mathcal{M}_G)P(D_S|S_E, \mathcal{M}_G)$.

Now we discover the probability $P(D_S|S_E, \mathcal{M}_G)$ in the numerator. Since the cursed hat's recommendation does not add any information about the P.A.R.S.E.L. score above and beyond the student's true House affiliation (i.e., it is *conditionally independent*; the test score is not entirely independent of the hat's recommendation since the hat is often right about the student's correct affiliation and the affiliation influences the test score), we can simplify this conditional probability: $P(D_S|S_E, \mathcal{M}_G) = P(D_S|\mathcal{M}_G)$. Note that the numerator now only contains known quantities: $P(S_E|\mathcal{M}_G)$ can be read off as 0.05 from Table 1.2; $P(D_S|\mathcal{M}_G)$ is the probability that a true Gryffindor is erroneously sorted into Slytherin, and since that happens to one in five true Gryffindors (because the proportion sorted into Gryffindor went down from 25% to 20%), $P(D_S|\mathcal{M}_G)$ must be 0.20; and $P(\mathcal{M}_G)$ is the base probability that a student is a Gryffindor, which we know to be one in four. Thus,

$$\begin{aligned} P(\mathcal{M}_G|D_S, S_E) &= \frac{P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|\mathcal{M}_G)}{P(D_S, S_E)} \\ &= \frac{0.25 \times 0.05 \times 0.20}{P(D_S, S_E)}. \end{aligned}$$

This leaves us having to find $P(D_S, S_E)$, the prior predictive probability that a student would be Sorted into Slytherin and score *Excellent* on the P.A.R.S.E.L. test. Here, the Sum Rule will help us out, because we can find the right hand side numerator for each type of student in the same way we did for true Gryffindors above – we can find $P(D_S, S_E|\mathcal{M}_i)$ for any House $i = S, G, R, H$. Hence (from Equation 1.3),

$$\begin{aligned} P(D_S, S_E) &= \sum_i P(\mathcal{M}_i)P(S_E|\mathcal{M}_i)P(D_S|\mathcal{M}_i) \\ &= P(\mathcal{M}_S)P(S_E|\mathcal{M}_S)P(D_S|\mathcal{M}_S) \\ &\quad + P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|\mathcal{M}_G) \\ &\quad + P(\mathcal{M}_R)P(S_E|\mathcal{M}_R)P(D_S|\mathcal{M}_R) \\ &\quad + P(\mathcal{M}_H)P(S_E|\mathcal{M}_H)P(D_S|\mathcal{M}_H) \\ &= 0.25 \times 0.80 \times 1.00 \\ &\quad + 0.25 \times 0.05 \times 0.20 \\ &\quad + 0.25 \times 0.05 \times 0.20 \\ &\quad + 0.25 \times 0.00 \times 0.20 \\ &= 0.2050. \end{aligned}$$

So finally, we arrive at:

$$P(\mathcal{M}_G|D_S, S_E) = \frac{0.0025}{0.2050} = 0.0122,$$

which allows Professor Binns to return to the Headmistress with good news: There is only around a 1% probability that a student who is Sorted into Slytherin and scores *Excellent* on the P.A.R.S.E.L. test is actually a Gryffindor. Binns further claims that the probability that such a student is a true Slytherin is over 95%, and that the combined procedure—that consists of first letting the Sorting Hat judge and then giving Slytherin-placed students a P.A.R.S.E.L. test and rehousing them by their score—will correctly place students

of any House with at least 90% probability. For example, he explains, a true Ravenclaw would be sorted into their correct House by the Hat with 80% ($P(D_R|\mathcal{M}_R)$) probability, and would be placed into Slytherin with 20% probability. In the second case, the student would be given the P.A.R.S.E.L. test, in which they would obtain an *Outstanding* with 80% ($P(S_O|\mathcal{M}_R)$) probability. Hence, they would be placed in their correct House with probability $P(D_R|\mathcal{M}_R) + P(D_S|\mathcal{M}_R) \times P(S_O|\mathcal{M}_R) = 0.80 + 0.20 \times 0.80 = 0.96$.

Discussion The Sorting Hat example introduces two extensions from the first. Here, there are not two but four possible “models” – whereas statistical inference is often seen as a choice problem between two alternatives, probabilistic inference naturally extends to any number of alternative hypotheses. The extension that allows for the evaluation of multiple hypotheses did not require the ad hoc formulation of any new rules, but relied entirely on the same basic rules of probability.

The example additionally underscores an inferential facility that we believe is vastly underused in social science: we selected between models making use of two *qualitatively different* sources of information. The two sources of information were individually insufficient but jointly powerful: the Hat placement is only 80% accurate in most cases, and the written test was only 50% accurate for the Ravenclaw case, but together they are 90% accurate. Again, this extension is novel only in that we had not yet considered it – the fact that information from multiple sources can be so combined requires no new facts and is merely a consequence of the two fundamental rules of probability.

Probability theory in the continuous case

In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the [posterior] distribution... indicates the range of values that are consistent with our prior information and data, and which honesty therefore compels us to admit as possible values.

E. T. Jaynes (1986)

The full power of probabilistic inference will come to light when we generalize from discrete events A with probabilities $P(A)$, to continuous parameters a with probability densities $p(a)$.⁵ Probability densities are different from probabilities in many ways. Densities express how much probability exists “near” a particular value of a , while the probability of any particular value of a in a continuous range is zero. Probability densities cannot be negative but they can be larger than 1, and they translate to probabilities through the mathematical operation of integration (i.e., calculating the area under a function over a certain interval). Possibly the most well-known distribution in psychology is the theoretical distribution of IQ in the population, which is shown in Figure 1.3.

⁵When we say a parameter is “continuous” we mean it could take any one of the infinite number of values comprising some continuum. For example, this would apply to values that follow a normal distribution.

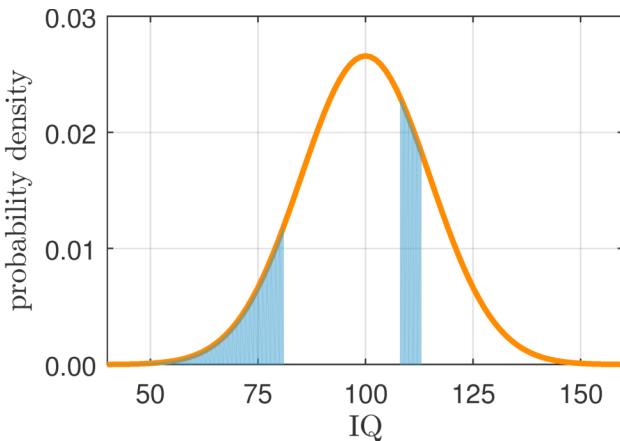


Figure 1.3: An example of a probability density function (PDF). PDFs express the relative plausibility of different values and can be used to determine the probability that a value lies in any interval. The PDF shown here is the theoretical distribution of IQ in the population: a normal distribution (a.k.a. Gaussian distribution) with mean 100 and standard deviation 15. In this distribution, the filled region to the left of 81 has an area of approximately 0.10, indicating that for a random member of the population, there is a 10% chance their IQ is below 81. Similarly, the narrow shaded region on the right extends from 108 to 113 and also has an area of 0.10, meaning that a random member has a 10% probability of falling in that region.

By definition, the total area under a probability density function is 1:

$$1 = \int_A p(a)da,$$

where capitalized A indicates that the integration is over the entire range of possible values for the parameter that appears at the end – in this case a . The range A is hence a disjoint set of possible values for a . For instance, if a is the mean of a normal distribution, A indicates the range of real numbers from $-\infty$ to ∞ ; if a is the rate parameter for a binomial distribution, A indicates the range of real numbers between 0 and 1. The symbol da is called the *differential* and the function that appears between the integration sign and the differential is called the *integrand* – in this case $p(a)$.

We can consider how much probability is contained within smaller sets of values within the range A ; for example, we could consider the integral over only the values of a that are less than 81, which would equal the probability that a is less than 81.⁶

$$P(a < 81) = \int_{-\infty}^{81} p(a)da.$$

In Figure 1.3, the shaded area on the left indicates the probability density over the region $(-\infty, 81)$.

⁶Strictly speaking, this integral is the probability that a is less than *or equal to* 81, but the probability of any single point in a continuous distribution is 0. By the sum rule, $P(a \leq 81) = P(a < 81) + P(a = 81)$, which simplifies to $P(a \leq 81) = P(a < 81) + 0$.

The fundamental rules of probability theory in the discrete case—the sum and product rules—have continuous analogues. The continuous form of the product rule is essentially the same as in the discrete case: $p(a, b) = p(a)p(b|a)$, where $p(a)$ is the density of the continuous parameter a and $p(b|a)$ denotes the *conditional density* of b (i.e., the density of b assuming a particular value of a). As in the discrete case of Equation 1.1, it is true that $p(a, b) = p(a)p(b|a) = p(b)p(a|b)$, and that $p(a, b) = p(a)p(b)$ if we consider a and b to be statistically independent. For the continuous sum rule, the summation in Equation 1.3 is replaced by an integration over the entire parameter space B :

$$p(a) = \int_B p(a, b) db.$$

Because this operation can be visualized as a function over two dimensions ($p(a, b)$ is a function that varies over a and b simultaneously) that is being collapsed into the one-dimensional margin ($p(a)$ varies only over a), this operation is alternatively called *marginalization*, *integrating over b* , or *integrating out b* .

Using these continuous forms of the sum and product rules, we can derive a continuous form of Bayes' Rule by successively applying the continuous sum and product rules to the numerator and denominator (analogously to Equation 1.7):

$$\begin{aligned} p(a|b) &= \frac{p(a, b)}{p(b)} = \frac{p(a)p(b|a)}{p(b)} \\ &= \frac{p(a)p(b|a)}{\int_A p(a)p(b|a) da}. \end{aligned} \tag{1.10}$$

Since the product in the numerator is divided by its own integral, the total area under the posterior distribution always equals 1; this guarantees that the posterior is always a proper distribution if the prior and likelihood are proper distributions.

One application of Bayesian methods to continuous parameters is *estimation*. If θ (theta) is a parameter of interest (say, the success probability of a participant in a task), then information about the relative plausibility of different values of θ is given by the probability density $p(\theta)$. If new information becomes available, for example in the form of new data x , the density can be updated and made conditional on x : It should be noted that by “continuous form of Bayes’ Rule” we mean that the prior and posterior distributions for the model parameter(s) are continuous – the sample data can still be discrete, as in Example 3 below.

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = \frac{p(\theta)p(x|\theta)}{\int_{\Theta} p(\theta)p(x|\theta)d\theta}. \tag{1.11}$$

Since in the context of scientific learning these two densities typically represent our knowledge of a parameter θ before and after taking into account the new data x , $p(\theta)$ is often called the *prior density* and $p(\theta|x)$ the *posterior density*. Obtaining the posterior density involves the evaluation of Equation 1.11 and requires one to define a likelihood function $p(x|\theta)$, which indicates how strongly the data x are implied by every possible value of the parameter θ . It should be noted that by “continuous form of Bayes’ Rule” we mean that the prior and

posterior distributions for the model parameter(s) are continuous – the sample data can still be discrete, as in Example 3 below.

The numerator on the right hand side of Equation 1.11, $p(\theta)p(x|\theta)$, is a product of the prior distribution and the likelihood function, and it completely determines the shape of the posterior distribution (note that the denominator in that equation is not a function of the parameter θ ; even though the parameter seems to feature in the integrand, it is in fact “integrated out” so that the denominator depends only on the data x). For this reason, many authors prefer to ignore the denominator of Equation 1.11 and simply write the posterior density as proportional to the numerator, as in $p(\theta|x) \propto p(\theta)p(x|\theta)$. We do not, because this conceals the critical role the denominator plays in a predictive interpretation of Bayesian inference.

The denominator $p(x)$ is the weighted-average probability of the data x , where the form of the prior distribution determines the weights. This normalizing constant is the continuous analogue of the prior predictive distribution, often alternatively referred to as the *marginal likelihood* or the Bayesian *evidence*.⁷ Consider that, in a similar fashion to the discrete case, we can rearrange Equation 1.11 as follows—dividing each side by $p(\theta)$ —to illuminate in an alternative way how Bayes’ rule operates in updating the prior distribution $p(\theta)$ to a posterior distribution $p(\theta|x)$:

$$\frac{p(\theta|x)}{p(\theta)} = \frac{p(x|\theta)}{p(x)} = \frac{p(x|\theta)}{\int_{\Theta} p(\theta)p(x|\theta)d\theta}. \quad (1.12)$$

On the left hand side, we see the ratio of the posterior to the prior density. Effectively, this tells us for each value of θ how much more or less plausible that value became due to seeing the data x . The equation shows that this ratio is determined by how well that specific value of θ predicted the data, in comparison to the weighted-average predictive accuracy across all values in the range Θ . In other words, **parameter values that exceed the average predictive accuracy across all values in Θ have their densities increased, while parameter values that predict worse than the average have their densities decreased** (see Morey, Romeijn, & Rouder, 2016; Wagenmakers, Morey, & Lee, in press).

While the discrete form of Bayes’ rule has natural applications in hypothesis testing, the continuous form more naturally lends itself to parameter estimation. Examples of such questions are: “What is the probability that the regression weight β is positive?” and “What is the probability that the difference between these means is between $\delta = -.3$ and $\delta = .3$?”. These questions can be addressed in a straightforward way, using only the product and sum rules of probability.

Example 3: “Perfection of the puking pastille”

In the secretive research and development laboratory of *Weasley’s Wizarding Wheezes*, George Weasley works to develop gag toys and prank foods for the entertainment of young witches and wizards. In a recent project, Weasley is studying the effects of his store’s famous *puking*

⁷We particularly like Evans’s take on the term Bayesian *evidence*: “For evidence, as expressed by observed data in statistical problems, is what causes beliefs to change and so we can measure evidence by measuring change in belief” (Evans, 2014, p. 243).

pastilles, which cause immediate vomiting when consumed. The target audience is Hogwarts students who need an excuse to leave class and enjoy making terrible messes.

Shortly after the pastilles hit Weasley's store shelves, customers began to report that puking pastilles cause not one, but multiple “expulsion events.” To learn more about this unknown behavior, George turns to his sister Ginny and together they decide to set up an exploratory study. From scattered customer reports, George believes the expulsion rate to be between three to five events per hour, but he intends to collect data to determine the rate more precisely. At the start of this project, George has no distinct hypotheses to compare – he is interested only in estimating the expulsion rate.

Since the data x are counts of the number of expulsion events within an interval of time, Ginny decides that the appropriate model for the data (i.e., likelihood function) is a Poisson distribution (see top panel of Figure 1.4):

$$p(x|\lambda) = \frac{1}{x!} \exp(-\lambda) \lambda^x, \quad (1.13)$$

with the λ (lambda) parameter representing the expected number of events within the time interval (note $\exp(-\lambda)$ is simply a clearer way to write $e^{-\lambda}$).

A useful prior distribution for Poisson rates is the Gamma distribution (Gelman, Carlin, Stern, & Rubin, 2004, Appendix A):⁸

$$p(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \exp(-\lambda b) \lambda^{a+1}, \quad (1.14)$$

A visual representation of the Gamma distribution is given in the second panel of Figure 1.4. A Gamma distribution has two parameters that determine its form, namely shape (a) and scale (b).⁹ The Gamma distribution is useful here for two reasons: first, it has the right *support*, meaning that it provides nonzero density for all possible values for the rate (in this case all positive real numbers); and second, it is *conjugate* with the Poisson distribution, a technical property to be explained below.

Before collecting further data, the Weasleys make sure to specify what they believe to be reasonable values based on the reports George has heard. In the second panel of Figure 1.4, Ginny set the prior parameters to $a = 2$ and $b = 0.2$ by drawing the shape of the distribution for many parameter combinations and selecting a curve that closely resembles George’s prior information: Values between three and five are most likely, but the true value of the expulsion rate could conceivably be much higher.

Three volunteers are easily found, administered one puking pastille each, and monitored for one hour. The observed event frequencies are $x_1 = 7$, $x_2 = 8$, and $x_3 = 19$.

⁸Recall that $x! = x \times (x - 1) \times \cdots \times 1$ (where $x!$ is read as “the factorial of x ,” or simply “ x factorial”). Similarly, the Gamma function $\Gamma(a)$ is equal to $(a - 1)! = (a - 1) \times (a - 2) \times \cdots \times 1$ when a is an integer. Unlike a factorial, however, the Gamma function is more flexible in that it can be applied to non-integers.

⁹To ease readability we use Greek letters for the parameters of a likelihood function and Roman letters for the parameters of prior (posterior) distributions. The parameters that characterize a distribution can be found on the right side of the conditional bar; for instance, the likelihood function $p(x|\lambda)$ has parameter λ , whereas the prior distribution $p(\lambda|a, b)$ has parameters (a, b) .

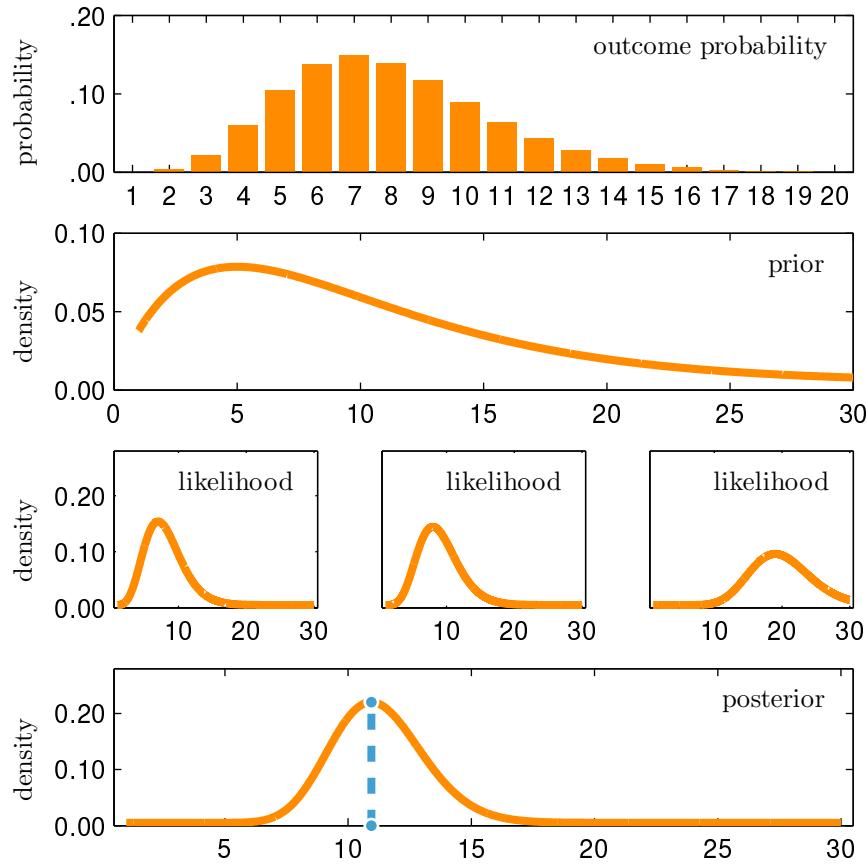


Figure 1.4: Progress of the Weasley's knowledge of the expulsion rate λ (the expected number of expulsion events per hour). **Top row:** An example Poisson distribution. The function is $p(x|\lambda = 7)$ as defined in Equation 1.13. The height of each bar indicates the probability of that particular outcome (e.g., number of expulsion events). **Second row:** The prior distribution of λ ; a Gamma distribution with parameters $a = 2$ and $b = 0.2$. **Third row:** The likelihood functions associated with $x_1 = 7$ (left), $x_2 = 8$ (center), and $x_3 = 19$ (right). **Bottom row:** The posterior distribution of λ ; a Gamma distribution with parameters $a = 36$ and $b = 3.2$.

With the prior density (Eq. 1.14) and the likelihood (Eq. 1.13) known, Ginny can use Bayes' rule as in Equation 1.10 to derive the posterior distribution of λ , conditional on the new data points $X_n = (x_1, x_2, x_3)$. She will assume the $n = 3$ data points are independent given λ , so that their likelihoods may be multiplied.¹⁰ This leaves her with the following expression for the posterior density of $(\lambda|X_n, a, b)$:

$$p(\lambda|X_n, a, b) = \frac{\frac{b^a}{\Gamma(a)} \exp(-\lambda b) \lambda^{a+1} \times \prod_{i=1}^{n=3} \frac{1}{x_i!} \exp(-\lambda) \lambda^{x_i}}{\int_{\Lambda} \frac{b^a}{\Gamma(a)} \exp(-\lambda b) \lambda^{a+1} \times \prod_{i=1}^{n=3} \frac{1}{x_i!} \exp(-\lambda) \lambda^{x_i} d\lambda}. \quad (1.15)$$

This expression may look daunting, but Ginny Weasley is not easily intimidated. She

¹⁰The likelihood function of the combined data is $p(X_n|\lambda) = p(x_1|\lambda) \times p(x_2|\lambda) \times p(x_3|\lambda)$, which we write using the product notation, $\prod_{i=1}^{n=3} p(x_i|\lambda)$, in the following equations to save space. Similarly, $\prod_{i=1}^3 \exp(-\lambda) \lambda^{x_i} = \exp(-3\lambda) \lambda^{(x_1+x_2+x_3)}$.

goes through the following algebraic steps to simplify the expression: (1) collect all factors that do not depend on λ (which, notably, includes the entire denominator) and call them $Q(X_n)$, and (2) combine exponents with like bases:

$$\begin{aligned} p(\lambda|X_n, a, b) &= Q(X_n) \exp(-\lambda b) \lambda^{a+1} \times \prod_{i=1}^{n=3} \exp(-\lambda) \lambda^{x_i} \\ &= Q(X_n) \exp[-\lambda(b + n)] \lambda^{(a + \sum_{i=1}^{n=3} x_i) + 1}. \end{aligned}$$

Note the most magical result that is obtained here! Comparing the last equation to Equation 1.14, it turns out that these have *exactly the same form*. Renaming $(b + n)$ to \hat{b} and $(a + \sum_i^n x_i)$ to \hat{a} makes this especially clear:

$$p(\lambda|X_n, a, b) = \frac{\hat{b}^{\hat{a}}}{\Gamma(\hat{a})} \exp(-\lambda \hat{b}) \lambda^{\hat{a}+1} = p(\lambda|\hat{a}, \hat{b}).$$

Here, Ginny has completed the distribution by replacing the scaling constant $Q(X_n)$ with the scaling constant of the Gamma distribution – after all, we know that the outcome must be a probability density, and each density has a unique scaling constant that ensures the total area under it is 1.

The posterior distribution $p(\lambda|X_n, a, b)$ thus turns out to be equal to the prior distribution with updated parameters $\hat{b} = b + n$ and $\hat{a} = a + \sum_i^n x_i$. Differently put,

$$p(\lambda|X_n, a, b) = p\left(\lambda | a + \sum_i^n x_i, b + n\right). \quad (1.16)$$

This amazing property, where the prior and posterior distributions have the same form, results from the special relationship between the Gamma distribution and the Poisson distribution: *conjugacy*. The bottom panel of Figure 1.4 shows the much more concentrated posterior density for λ : a Gamma distribution with parameters $\hat{a} = 36$ and $\hat{b} = 3.2$.

When priors and likelihoods are conjugate, three main advantages follow. First, it is easy to express the posterior density because it has the same form as the prior density (as seen in Equation 1.16). Second, it is straightforward to calculate means and other summary statistics of the posterior density. For example, the mean of a Gamma distribution has a simple formula: a/b . Thus, George and Ginny's prior density for λ has a mean of $a/b = 2/.2 = 10$, and their posterior density for λ has a mean of $\hat{a}/\hat{b} = 36/3.2 = 11.25$. The prior and posterior densities' respective modes are $(a - 1)/b = 5$ and $(\hat{a} - 1)/\hat{b} = 35/3.2 \approx 11$, as can be seen from Figure 1.4. Third, it is straightforward to update the posterior distribution sequentially as more data become available.

Discussion Social scientists estimate model parameters in a wide variety of settings. Indeed, a focus on estimation is the core of the *New Statistics* (Cumming, 2014). The *puking pastilles* example illustrates how Bayesian parameter estimation is a direct consequence of the rules of probability theory, and this relationship licenses a number of interpretations that the *New Statistics* does not allow. Specifically, the basis in probability theory allows George

and Ginny to (1) point at the most plausible values for the rate of expulsion events and (2) provide an interval that contains the expulsion rate with a certain probability (e.g., a Gamma distribution calculator shows that λ is between 8.3 and 14.5 with 90% probability).

The applications of parameter estimation often involve exploratory settings: no theories are being tested and a distributional model of the data is assumed for descriptive convenience. Nevertheless, parameter estimation can be used to adjudicate between theories under certain special circumstances: if a theory or hypothesis makes a particular prediction about a parameter's value or range, then estimation can take a dual role of hypothesis testing. In the social sciences most measurements have a natural reference point of zero, so this type of hypothesis will usually be in the form of a directional prediction for an effect. In our example, suppose that George was specifically interested in whether λ was less than 10. Under his prior distribution for λ , the probability of that being the case was 59.4%. After seeing the data, the probability λ is less than 10 decreased to 26.2%.

Estimating the mean of a normal distribution

By far the most common distribution used in statistical testing in social science, the normal distribution deserves discussion of its own. The normal distribution has a number of interesting properties—some of them rather unique—but we discuss it here because it is a particularly appropriate choice for modeling unconstrained, continuous data. The mathematical form of the normal distribution is

$$\begin{aligned} p(x|\mu, \sigma^2) &= N(x|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \end{aligned}$$

with the μ (mu) parameter representing the average (*mean*) of the population from which we are sampling and σ (sigma) the amount of dispersion (*standard deviation*) in the population. We will follow the convention that the normal distribution is parameterized with the *variance* σ^2 . An example normal distribution is drawn in Figure 1.3.

One property that makes the normal distribution useful is that it is *self-conjugate*: The combination of a normal prior density and normal likelihood function is itself a normal distribution, which greatly simplifies the derivation of posterior densities. Using Equation 1.10, and given some data set $X_n = (x_1, x_2, \dots, x_n)$, we can derive the following expression for the posterior density ($\mu|X_n, a, b$):

$$p(\mu|X_n, a, b) = \frac{N(\mu|a, b^2) \times \prod_i^n N(x_i|\mu, \sigma^2)}{\int_M N(\mu|a, b^2) \times \prod_i^n N(x_i|\mu, \sigma^2) d\mu}$$

Knowing that the product of normal distributions is also a normal distribution (up to a scaling factor), it is only a matter of tedious algebra to derive the posterior distribution of μ . We do not reproduce the algebraic steps here – the detailed derivation can be found in (Gelman et al., 2004), among many other places. The posterior is

$$p(\mu|X_n, a, b) = N\left(\mu|\hat{a}, \hat{b}^2\right),$$

where

$$\hat{b}^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{b^2}}$$

and

$$\begin{aligned}\hat{a} &= \left(\frac{\hat{b}^2}{b^2}\right)a + \left(\frac{\hat{b}^2}{\sigma^2/n}\right)\bar{x} \\ &= W^2a + (1 - W^2)\bar{x},\end{aligned}$$

where \bar{x} refers to the mean of the sample.

Carefully inspecting these equations can be instructive. To find \hat{b} , the standard deviation (i.e., spread) of the posterior distribution of μ , we must compare the spread of the prior distribution, b , to the *standard error of the sample*, σ/\sqrt{n} . The formula for \hat{b} represents how our uncertainty about the value of μ is reduced due to the information gained in the sample. If the sample is noisy, such that the standard error of the sample is large compared to the spread of the prior, then relatively little is learned from the data compared to what we already knew before, so the difference between \hat{b} and b will be small. Conversely, if the data are relatively precise, such that the standard error of the sample is small when compared to the spread of the prior, then much will be learned about μ from the data and \hat{b} will be much smaller than b .

To find \hat{a} , the mean of the posterior distribution for μ , we need to compute a weighted average of the prior mean and the sample mean. In the formula above, the weights attached to a and \bar{x} sum to 1 and are determined by how much each component contributes to the total precision of the posterior distribution. Naturally, the best guess for the value of μ splits the difference between what we knew of μ before seeing the sample and the estimate of μ obtained from the sample; whether the posterior mean is closer to the prior mean or the sample mean depends on a comparison of their relative precision. If the data are noisy compared to the prior (i.e., the difference between prior variance b^2 and posterior variance \hat{b}^2 is small, meaning W^2 is near 1), then the posterior mean will stay relatively close to the prior mean. If the data are relatively precise (i.e., W^2 is near zero), the posterior mean will move to be closer to the sample mean. If the precision of the prior and the precision of the data are approximately equal then W^2 will be near 1/2, so the posterior mean for μ will fall halfway between a and \bar{x} .

The above effect is often known as *shrinkage* because our sample estimates are pulled back toward prior estimates (i.e., shrunk). Shrinkage is generally a desirable effect, in that it will lead to more accurate parameter estimates and empirical predictions (see Efron & Morris, 1977). Since Bayesian estimates are automatically shrunk according to the relative precision of the prior and the data, incorporating prior information simultaneously improves our parameter estimates and protects us from being otherwise misled by noisy estimates in small samples. Quoting Gelman (2010, p. 163): “Bayesian inference is conservative in that it goes with what is already known, unless the new data force a change.”

Another way to interpret these weights is to think of the prior density as representing some amount of information that is available from an unspecified number of previous hypothetical observations, which are then added to the information from the real observations in

the sample. For example, if after collecting 20 data points the weights come to $W^2 = .5$ and $1 - W^2 = .5$, that implies that the prior density carried 20 data points' worth of information. In studies for which obtaining a large sample is difficult, the ability to inject outside information into the problem to come to more informed conclusions can be a valuable asset. A common source of outside information is estimates of effect sizes from previous studies in the literature. As the sample becomes more precise, usually through increasing sample size, W^2 will continually decrease, and eventually the amount of information added by the prior will become a negligible fraction of the total (see also the principle of stable estimation, described in Edwards et al., 1963).

Example 4: “Of Murtlaps and Muggles”

According to *Fantastic Beasts and Where to Find Them* (Scamander, 2001), a Murtlap is a “rat-like creature found in coastal areas of Britain” (p. 56). While typically not very aggressive, a startled Murtlap might bite a human, causing a mild rash, discomfort in the affected area, profuse sweating, and some more unusual symptoms.

Anecdotal reports dating back to the 1920s indicate that Muggles (non-magical folk) suffer a stronger immunohistological reaction to Murtlap bites. This example of physiological differences between wizards and Muggles caught the interest of famed magizoologist Newton (“Newt”) Scamander, who decided to investigate the issue: When bitten by a Murtlap, do symptoms persist longer in the average Muggle than in the average wizard?

The Ministry of Magic keeps meticulous historical records of encounters between wizards and magical creatures that go back over a thousand years, so Scamander has a great deal of information on wizard reactions to Murtlap bites. Specifically, the average duration of the ensuing sweating episode is 42 hours, with a standard deviation of 2. Due to the large amount of data available, the standard error of measurement is negligible. Scamander’s question can now be rephrased: What is the probability a Murtlap bite on a Muggle results in an average sweating episode longer than 42 hours?

Scamander has two parameters of interest: the population mean—episode duration μ —and its corresponding population standard deviation σ . He has no reason to believe there is a difference in dispersion between the magical and non-magical populations, so he will assume for convenience that σ is known and does not differ between Muggles and wizards (i.e., $\sigma = 2$; ideally, σ would be estimated as well, but for ease of exposition we will take the standard deviation as known). To characterize his background information about the population mean μ , Scamander uses a prior density represented by a normal distribution, $p(\mu|a, b) = N(\mu|a, b^2)$, where a represents the location of the mean of the prior and b represents its standard deviation (i.e., the amount of uncertainty we have regarding μ).

Before collecting any data, Scamander must assign to μ a prior distribution that represents what he believes to be the range of plausible values for this parameter before collecting data. From his informal observations, Scamander believes that the mean difference between wizards and Muggles will probably not be larger than 15 hours. To reflect this information, Scamander centers the prior distribution $p(\mu|a, b)$ at $a = 42$ hours (the average among wizards) with a standard deviation of $b = 6$ hours, so that prior to running his study there is a 95% probability μ lies between (approximately) 27 and 57 hours. Thus,

$$p(\mu|a, b) = N(\mu|42, 6^2).$$

With these prior distributions in hand, Scamander can compute the prior probability that μ is less than 42 hours by finding the area under the prior distribution to the left of the benchmark value via integration. Integration from negative infinity to some constant is most conveniently calculated with the *cumulative distribution function* Φ :

$$\begin{aligned} p(\mu < 42|a, b) &= \int_{-\infty}^{42} N(\mu|a, b^2) d\mu \\ &= \Phi(42|a, b^2), \end{aligned}$$

which in this case is exactly 0.5 since the benchmark value is exactly the mean: Scamander centered his prior on 42 and specified that the Muggle sweating duration could be longer or shorter with equal probability.

Scamander covertly collects information on a representative sample of 30 Muggles by exposing them to an angry Murtlap.¹¹ He finds a sample mean of $\bar{x} = 43$ and standard error of $s = \sigma/\sqrt{n} = 2/\sqrt{30} = 0.3651$. Scamander can now use his data and the above formulas to update what he knows about μ .

Since the spread of the prior for μ is large compared to the standard error of the sample ($b = 6$ versus $s = 0.3651$), Scamander has learned much from the data and his posterior density for μ is much less diffuse than his prior:

$$\hat{b} = \sqrt{\frac{1}{\frac{1}{s^2} + \frac{1}{b^2}}} = \sqrt{\frac{1}{\frac{1}{0.3651^2} + \frac{1}{6^2}}} = 0.3645.$$

With \hat{b} in hand, Scamander can find the weights needed to average a and \bar{x} : $W^2 = (0.3645/6)^2 = 0.0037$ and $1 - W^2 = 0.9963$, thus $\hat{a} = 0.0037 \times 42 + 0.9963 \times 43 = 42.9963$ hours. In summary, Scamander's prior distribution for μ , $p(\mu|a, b) = N(\mu|42, 6^2)$, is updated into a much more informative posterior distribution, $p(\mu|\hat{a}, \hat{b}) = N(\mu|42.9963, 0.3645^2)$. This posterior distribution is shown in the left panel of Figure 1.5; note that the prior density looks nearly flat when compared to the much more peaked posterior density.

Now that the posterior distribution of μ is known, Scamander can revisit his original question: What is the probability that μ is greater than 42 hours? The answer is again obtained by finding the area under the posterior distribution to the right of the benchmark value via integration:

$$\begin{aligned} p(\mu < 42|\hat{a}, \hat{b}) &= \int_{42}^{\infty} N(\mu|\hat{a}, \hat{b}^2) d\mu \\ &= 1 - \int_{-\infty}^{42} N(\mu|\hat{a}, \hat{b}^2) d\mu \\ &= 1 - \Phi(42|\hat{a}, \hat{b}^2) \\ &= 1 - \Phi(42|42.9963, 0.3645^2) \approx 0.9970. \end{aligned}$$

¹¹In order to preserve the wizarding world's statutes of secrecy, Muggles who are exposed to magical creatures must be turned over to a team of specially-trained wizards called *Obliviators*, who will erase the Muggles' memories, return them to their homes, and gently steer them into the kitchen.

In summary, the probability that the reaction to Murtlap bites in the average Muggle is greater than in the average wizard increases from exactly 50% to 99.70%.

Discussion The conclusion of a Bayesian estimation problem is the full posterior density for the parameter(s). That is, once the posterior density is obtained then the estimation problem is complete. However, researchers often choose to report summaries of the posterior distribution that represent its content in a meaningful way. One common summary of the posterior density is a *posterior (credible) interval*. Credible intervals have a unique property: as Edwards et al. (1963) put it, “The Bayesian theory of interval estimation is simple. To name an interval that you feel 95% certain includes the true value of some parameter, simply inspect your posterior distribution of that parameter; any pair of points between which 95% of your posterior density lies defines such an interval” (p. 213). This property is made possible by the inclusion of a prior density in the statistical model (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). It is important not to confuse credible intervals with *confidence intervals*, which have no such property (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Thus, when Scamander reports that there is a 99.70% probability that μ lies between 42 and positive infinity hours, he is reporting a 99.70% credible interval. It is important to note that there is no unique interval for summarizing the posterior distribution; the choice depends on the context of the research question.

Model comparison

[M]ore attention [should] 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.

H. Jeffreys (1939)

Consider the following theoretical questions. Is participant performance different than chance? Does this gene affect IQ? Does stimulus orientation influence response latency? For each of these questions the researcher has a special interest in a particular parameter value and entertains it as a possibility. However, when we estimate a parameter using a continuous distribution the answers to each of these questions is necessarily “yes.” To see why, recall that a probability density function specifies how much probability exists *near*—not *at*—a particular value of the parameter. That is, with a continuous probability distribution, probability only exists within a given *range* of the parameter space; the probability of any *single point* within the distribution is zero. This is inconsistent with our belief that a specified parameter value might hold true. Moreover, this poses a problem for any research question that focuses on a single value of a continuous parameter, because if its prior probability is zero then no amount of data can cause its posterior probability to become anything other than zero.

A simple but brilliant solution to this problem was first executed by Haldane (1932) but is credited mostly to Jeffreys (1939; see Etz & Wagenmakers, in press). The solution

involves applying the sum and product rules *across multiple independent statistical models at once*. We can specify multiple separate models that have different implications about the parameter of interest, call it θ , and calculate the probability of each model after data are collected. One model, say \mathcal{M}_0 , says θ is equal to a single special value denoted θ_0 . A second model, say \mathcal{M}_1 , says θ is unknown and assigns it a continuous prior density, implying θ is not equal to θ_0 . After collecting data X , there are two main questions to answer: (1) What is $P(\mathcal{M}_0|X)$, the posterior probability that $\theta = \theta_0$? And (2) what is $p(\theta|X, \mathcal{M}_1)$, the posterior distribution¹² of θ under \mathcal{M}_1 (i.e., considering the new data X , if $\theta \neq \theta_0$ then what might θ be)?

As before, this scenario can be approached with the product and sum rules of probability. The setup of the problem is captured by Figure 1.7 (focusing for now on the left half). We start at the initial fork with two potential models: \mathcal{M}_0 and \mathcal{M}_1 . This layer of analysis is called the *model space*, since it deals with the probability of the models. Subsequently, each model implies some belief about the value of θ . This layer of analysis is called the *parameter space*, since it specifies what is known about the parameters within a model and each model has its own independent parameter space. Under \mathcal{M}_0 the value of θ is known to be equal to θ_0 , so all of its probability is packed into a “spike” (a *point mass*) at precisely θ_0 . Under \mathcal{M}_1 the value of θ is unknown and we place a probability distribution over the potential values of θ in the form of a *conditional prior density*. Each model also makes predictions about what data will occur in the experiment, information represented by each model’s respective *sample space*. We then condition on the data we observe, which allows us to update each layer of the analysis to account for the information gained. Below is a step-by-step account of how this is done.

We answer our questions in reverse order, first deriving the posterior distribution of θ under \mathcal{M}_1 , for a reason that will become clear in a moment. In this setup there are events that vary among three dimensions: X , θ , and \mathcal{M}_1 . When joint events have more than two components, the product rule decomposes $p(X, \theta, \mathcal{M}_1)$ one component at a time to create a chain of conditional probabilities and densities (for this reason the product rule is also known as the *chain rule*). This was seen above in Example 2. These chains can be thought of as moving from one layer of Figure 1.7 to the next. Thus, since we could choose any one of the three events to be factored out first, the product rule creates three possible initial chains with two probabilities per chain,

$$\begin{aligned} p(X, \theta, \mathcal{M}_1) &= P(\mathcal{M}_1)p(X, \theta|\mathcal{M}_1) \\ &= P(X)p(\theta, \mathcal{M}_1|X) \\ &= p(\theta)p(X, \mathcal{M}_1|\theta). \end{aligned}$$

(where the use of $P(X)$ or $p(X)$ depends on whether the data are discrete or continuous; we assume they are discrete here).

¹²Note that we will now be using probabilities and probability densities side-by-side. In general, if the event to which the measure applies (i.e., what is to the left of the vertical bar) has a finite number of possible values, we will consider probabilities and use uppercase $P(\cdot)$ to indicate that. If the event has an infinite number of possible values in a continuum, we will consider probability densities and use lowercase $p(\cdot)$. In the case of a joint event in which at least one component has an infinite set of possibilities, the joint event will also have an infinite set of possibilities and we will use probability densities there also.

A natural choice is to work with the first formulation, $p(X, \theta, \mathcal{M}_1) = P(\mathcal{M}_1)p(X, \theta | \mathcal{M}_1)$, since $P(\mathcal{M}_1)$, the prior probability of the model, is known to us (it corresponds to the probability we take the right fork at the start of Figure 1.7). The product rule can then be applied again to the remaining joint probability on the right hand side as follows:

$$P(\mathcal{M}_1) \times p(X, \theta | \mathcal{M}_1) = P(\mathcal{M}_1) \times P(X | \mathcal{M}_1)p(\theta | X, \mathcal{M}_1), \quad (1.17)$$

By symmetry of the product rule, we can also write

$$P(\mathcal{M}_1) \times P(X, \theta | \mathcal{M}_1) = P(\mathcal{M}_1) \times p(\theta | \mathcal{M}_1)P(X | \theta, \mathcal{M}_1). \quad (1.18)$$

If we now equate the right hand sides of Equations 1.17 and 1.18, we can divide out $P(\mathcal{M}_1)$ and $P(X | \mathcal{M}_1)$:

$$\begin{aligned} P(\mathcal{M}_1)P(X | \mathcal{M}_1)p(\theta | X, \mathcal{M}_1) &= P(\mathcal{M}_1)p(\theta | \mathcal{M}_1)P(X | \theta, \mathcal{M}_1) \\ p(\theta | X, \mathcal{M}_1) &= \frac{p(\theta | \mathcal{M}_1)P(X | \theta, \mathcal{M}_1)}{P(X | \mathcal{M}_1)} \end{aligned}$$

and by recognizing that $P(X | \mathcal{M}_1) = \int_{\Theta} p(\theta | \mathcal{M}_1)P(X | \mathcal{M}_1, \theta)d\theta$ by way of the sum rule, we are left with the following:

$$p(\theta | X, \mathcal{M}_1) = \frac{p(\theta | \mathcal{M}_1)P(X | \theta, \mathcal{M}_1)}{\int_{\Theta} p(\theta | \mathcal{M}_1)P(X | \theta, \mathcal{M}_1)d\theta}. \quad (1.19)$$

This last formula is identical to the continuous form of Bayes' Rule (Equation 1.10), where now each term is also conditional on \mathcal{M}_1 .

The implication of this finding is that it is possible to perform inference using the distribution of θ under \mathcal{M}_1 , $p(\theta | X, \mathcal{M}_1)$, *ignoring everything relating to other models*, since no other models (such as \mathcal{M}_0) feature in this calculation. As before, the denominator is known as the *marginal likelihood* for \mathcal{M}_1 , and represents a predictive distribution for potential future data, $P(X | \mathcal{M}_1)$. This predictive distribution is shown in the sample space under \mathcal{M}_1 in Figure 1.7, and can be thought of as the average prediction made across all possible parameter values in the model (weighted by the conditional prior density). Once the data are collected and the result is known, we can condition on the outcome and use it to update $p(\theta | \mathcal{M}_1)$ to obtain $p(\theta | X, \mathcal{M}_1)$.

To answer our first question—what is $P(\mathcal{M}_0 | X)$?—we need to find our way back to the discrete form of Bayes' Rule (Equation 1.7). Recall that for hypothesis testing the key terms to find are $P(X | \mathcal{M}_0)$ and $P(X | \mathcal{M}_1)$, which can be interpreted as how accurately each hypothesis predicts the observed data in relation to the other. Since the parameter space under \mathcal{M}_0 is simply $\theta = \theta_0$, we can write $P(X | \mathcal{M}_0) = P(X | \theta_0)$. However, since the parameter space under \mathcal{M}_1 includes a continuous distribution, we need to find \mathcal{M}_1 's average predictive success across the whole parameter space, $P(X | \mathcal{M}_1) = \int_{\Theta} p(\theta | \mathcal{M}_1)P(X | \mathcal{M}_1, \theta)d\theta$. Conveniently, as we just saw above in Equation 1.19, this is also the normalizing constant in the denominator of the posterior distribution of θ under \mathcal{M}_1 . Hence, the discrete form of

Bayes' Rule for hypothesis testing can be rewritten as

$$\begin{aligned} P(\mathcal{M}_1|X) &= \frac{P(\mathcal{M}_1)P(X|\mathcal{M}_1)}{P(\mathcal{M}_1)P(X|\mathcal{M}_1) + P(\mathcal{M}_0)P(X|\mathcal{M}_0)} \\ &= \frac{P(\mathcal{M}_1) \int_{\Theta} p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta}{P(\mathcal{M}_1) \int_{\Theta} p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta + P(\mathcal{M}_0)P(X|\theta_0)}. \end{aligned}$$

Furthermore, in cases of model comparison between a “point null” (i.e., an hypothesis that, like our \mathcal{M}_0 , involves a prior point mass on some parameter) and an alternative with a continuous prior for the parameter, one can rewrite the odds form of Bayes' Rule from Equation 1.9 as follows:

$$\begin{aligned} \underbrace{\frac{P(\mathcal{M}_1|X)}{P(\mathcal{M}_0|X)}}_{\text{Posterior odds}} &= \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{P(X|\mathcal{M}_1)}{P(X|\mathcal{M}_0)}}_{\text{Bayes factor } (BF_{10})} \\ &= \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{\int_{\Theta} p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta}{P(X|\theta_0)}}_{\text{Bayes factor } (BF_{10})}, \end{aligned}$$

where the Bayes factor is the ratio of the marginal likelihoods from the two models, and its subscript indicates which models are being compared (BF_{10} means \mathcal{M}_1 is in the numerator versus \mathcal{M}_0 in the denominator).

Finally, we point out one specific application of Bayes' rule that occurs when certain values of θ have a special theoretical status. For example, if θ represents the difference between two conditions in an experiment, then the case $\theta = 0$ will often be of special interest (see also Rouder & Vandekerckhove, this volume). Dividing each side of Equation 1.19 by $p(\theta|\mathcal{M}_1)$ allows one to quantify the change in the density at this point:

$$\frac{p(\theta = 0|X, \mathcal{M}_1)}{p(\theta = 0|\mathcal{M}_1)} = \frac{P(X|\theta = 0, \mathcal{M}_1)}{\int_{\Theta} p(\theta|\mathcal{M}_1)p(X|\theta, \mathcal{M}_1)d\theta} = BF_{01}$$

This change in density is known as the Savage–Dickey density ratio or the Savage–Dickey representation of the Bayes factor (Dickey, 1971; see also Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010, and Wagenmakers, Marsman, et al., this volume; and see also Marin & Robert, 2010, for some cautionary notes). When it applies, the Savage–Dickey ratio allows for an especially intuitive interpretation of the Bayes factor: If the point null value is lower on the alternative model's conditional posterior density than its prior density, the Bayes factor supports \mathcal{M}_1 over \mathcal{M}_0 by the ratio of their respective heights, and vice-versa.

The conditions under which the Savage–Dickey ratio applies are typically met in practice, since they correspond to the natural way one would build nested models for comparison (for a good discussion on the different types of nested models see Consonni & Veronese, 2008, Section 2). Namely, that all facets of the models are the same except that the smaller model fixes θ to be θ_0 . In our development above there is only one parameter so this

condition is automatically satisfied. If, however, we have additional parameters common to both models, say ϕ , then the Savage-Dickey ratio is obtained using the marginal prior and posterior densities, $p(\theta = \theta_0|X, \mathcal{M}_1)/p(\theta = \theta_0|\mathcal{M}_1)$, where the marginal distribution is found using the sum rule, $p(\theta|X, \mathcal{M}_1) = \int_{\Phi} p(\phi, \theta|X, \mathcal{M}_1)d\phi$. For this to be a proper representation of the Bayes factor, we must ensure that the conditional prior for ϕ under \mathcal{M}_1 , when $\theta = \theta_0$, equals the prior density for ϕ under \mathcal{M}_0 . In other terms, the Savage-Dickey representation holds only if the parameters are statistically independent a priori: $p(\phi|\theta = \theta_0, \mathcal{M}_1) = p(\phi|\mathcal{M}_0)$.

Above, our motivation for model comparison was that we wanted to test the hypothesis that a parameter took a single specified value. However, model comparison is not limited to cases where point nulls are tested. The above formulation allows us to compare any number of different types of models by finding the appropriate $P(X|\mathcal{M})$. Models do not need to be nested or even have similar functional forms; in fact, the models need not be related in any other way than that they make quantitative predictions about the data that have been observed. For example, a non-nested comparison might pit a model with a mostly positive prior distribution for θ against a model where the support of the prior distribution for θ is restricted to negative values only. Or rather than a precise point null we can easily adapt the null model such that we instead compare \mathcal{M}_1 against model \mathcal{M}_S , which says θ is “small.” Extending model comparison to the scenario where there are more than two (but finitely many) competing models \mathcal{M}_k is similar to before, in that

$$P(\mathcal{M}_i|X) = \frac{P(\mathcal{M}_i)p(X|\mathcal{M}_i)}{\sum_k P(\mathcal{M}_k)p(X|\mathcal{M}_k)}. \quad (1.20)$$

In practice, Bayes factors can be difficult to compute for more complicated models because one must integrate over possibly very many parameters to obtain the marginal likelihood (Kass & Raftery, 1995; Wasserman, 2000). Recent computational developments have made the computation of Bayes factors more tractable, especially for common scenarios (Wagenmakers, Love, et al., this volume; Wagenmakers, Marsman, et al., this volume). For uncommon or complex scenarios, one might resort to reporting a different model comparison metric that does not rely on the marginal likelihood, such as the various information criteria (AIC, BIC, DIC, WAIC) or leave-one-out cross validation (LOOCV; see Spiegelhalter, Best, Carlin, & van der Linde, 2002; Vandekerckhove, Matzke, & Wagenmakers, 2015; Vehtari & Ojanen, 2012). However, it should be emphasized that for the purposes of inference these alternative methods can be suboptimal.

Example 5: “The French correction”

Proud of his work on Murtlap bite sensitivity, Newt Scamander (from Example 4) decides to present his results at a conference on magical zoology held in Carcassonne, France. As required by the 1694 International Decree on the Right of Access to Magical Research Results, he has made all his data and methods publicly available ahead of time and he is confident that his findings will withstand the review of the audience at this annual meeting. He delivers a flawless presentation that culminates in his conclusion that Muggles are, indeed, slightly

more sensitive to Murtlap bites than magical folk are. The evidence, he claims, is right there in the data.

After his presentation, Scamander is approached by a member of the audience—the famously critical high-born wizard Jean-Marie le Cornichonesque—with a simple comment on the work: “*Monsieur*, you have not told us the evidence for your claim.”

“In fact,” continues le Cornichonesque, “given your prior distributions for the difference between Muggles and magical folk, you have not even *considered* the possibility that the true difference might be exactly zero, and your results merely noise. In other words, you are putting the cart before the horse because you estimate a population difference before establishing that evidence for one exists! Instead, if you please, let us ascertain how much more stock we should put in *your* claim over the *more parsimonious* claim of no difference between the respective population means.”

Scamander is unfazed by the nobleman’s challenge, and, with a flourish of his wand makes the following equations appear in the air between them:

$$\begin{aligned}\mathcal{M}_s : \mu &\sim N(42, 6) \\ \mathcal{M}_c : \mu &= 42\end{aligned}$$

“These,” Scamander says, “are our respective hypotheses. I claim that Muggles have different symptom durations on average than wizards and witches. I have prior information that completes my model. Your claim is that the population means may be exactly equal. In order to quantify the relative support for each of these hypotheses, we need a Bayes factor. Luckily, in this case the Bayes factor is quite easy to calculate with the Savage-Dickey density ratio, like so . . .”

$$\begin{aligned}\frac{p(\mu|X, \mathcal{M}_s)}{p(\mu|\mathcal{M}_s)} &= \frac{p(\mu|X, \mathcal{M}_s)}{p(\mu|\mathcal{M}_s)} \\ &= \frac{N(\mu|\hat{a}, \hat{b}^2)}{N(\mu|a, b^2)}\end{aligned}$$

“Now that we have derived the ratio of posterior to prior density, all that remains is to plug in the values of the parameters and to compute the ratio of Gaussian densities at the specified points . . .”

$$\begin{aligned}BF_{cs} &= \frac{N(42 | 42.9963, 0.3645^2)}{N(42 | 42, 6^2)} \\ &= \frac{0.0261}{0.0665} = 0.3925 = \frac{1}{2.5475}\end{aligned}$$

“*Tant pis.* A Bayes factor of not even three favors your hypothesis. You have essentially *no* evidence for your claim,” snorts le Cornichonesque, before turning his back and leaving Scamander alone in the conference room.

Discussion What has happened here? At first glance, it appears that at first Scamander had strong evidence that Muggles are more sensitive than magical folk to Murtlap bites, and

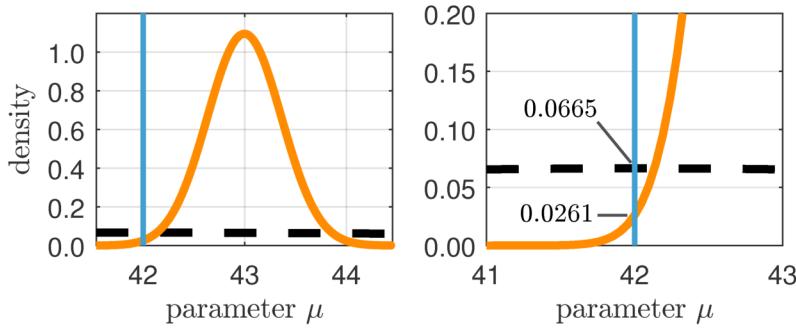


Figure 1.5: A closer look at the prior (dashed) and posterior (solid) densities involved in Newt Scamander’s study on the relative sensitivity of magical folk and Muggles to Murtlap bites. The left panel shows the location of the fixed value (42) in the body of the prior and posterior distributions. The right panel is zoomed in on the density in the area around the fixed value. Comparing the prior density to the posterior density at the fixed value reveals that very little was learned about this specific value: the density under the posterior is close to the density under the prior and amounts to a Bayes factor of approximately 3 supporting a deviation from the fixed value.

now through some sleight of hand his evidence appears to have vanished. To resolve the paradox of le Cornichonesque, it is important to appreciate a few facts. First, in Example 4, Scamander indeed did not consider the hypothesis \mathcal{M}_c that $\mu = 42$. In fact, because a continuous prior density was assigned to μ , the prior probability of it taking on any particular value is zero.

The paradox of le Cornichonesque occurs in part because of a confusion between the hypotheses being considered. While in our example, le Cornichonesque wishes to compare an “existence” and a “nonexistence” hypothesis, Scamander started out from an existence assumption and arrives at conclusions about *directionality* (see also Marsman & Wagenmakers, 2016).

Implicitly, there are four different models being considered in all. There is \mathcal{M}_c , which specifies no effect, and \mathcal{M}_z , which specifies *some* effect, but also \mathcal{M}_- , which specifies an effect in the negative direction, and \mathcal{M}_+ , which specifies an effect in the positive direction. These last two models are concealed by Scamander’s original analysis, but his model specification implies a certain probability for the events ($\mu < 42$) and ($\mu > 42$). Indeed, because we know that the probability that Muggles are more (vs. less) sensitive than their magical counterparts increased from $P(\mu < 42) = 50\%$ to $P(\mu < 42|X) = 99.70\%$, we can compute Bayes factors for this case as well. In odds notation, the prior odds were increased from 1 to 333; the Bayes factor, found by taking the ratio of posterior to prior odds, is in this case equal to the posterior odds. Scamander’s test for direction returns a much stronger result than le Cornichoneque’s test of existence.

As a rule, inference must be limited to the hypotheses under consideration: No method of inference can make claims about theories not considered or ruled out a priori. Moreover, the answer we get naturally depends on the question we ask. The example that follows involves a very similar situation, but the risk of the paradox of le Cornichonesque is avoided by making explicit all hypotheses under consideration.

Example 6: “The measure of an elf”

In the wizarding world, the Ministry of Magic distinguishes between two types of living creatures. *Beings*, such as witches, wizards, and vampires, are creatures who have the intelligence needed to understand laws and function in a peaceful society. By contrast, *Beasts* are creatures such as trolls, dragons, and grindylows, which do not have that capacity. Recently, the classification of house-elves has become a matter of contention. On one side of the debate is the populist wizard and radio personality Edward Runcorn, who claims that house-elves are so far beneath wizard intelligence that they should be classified as Beasts; on the other side is the famed elfish philosopher and acclaimed author Doc, who argues that elves are as intelligent as wizards and should be classified as Beings, with all the rights and responsibilities thereof. The Ministry of Magic decides to investigate and convene the *Wizengamot’s Internal Subcommittee on House Elf Status* (W.I.S.H.E.S.), an ad-hoc expert committee. W.I.S.H.E.S. in turn calls on psychometrician Dr. Karin Bones of the Magical Testing Service to decide whether house-elves are indeed as intelligent as wizards.

Bones knows she will be asked to testify before W.I.S.H.E.S. and takes note of the composition of the three-member committee. The committee’s chairperson is Griselda Marchbanks, a venerable and wise witch who is known for her impartiality and for being of open mind to all eventualities. However, the junior members of W.I.S.H.E.S. are not so impartial: one member is Edward Runcorn, the magical supremacist who believes that wizards and witches are more intelligent than house elves; the other is Hermione Granger, a strong egalitarian who believes that house elves are equal in intelligence to wizards and witches.

Bones begins her task by formalizing three basic hypotheses. She will call the population’s average wizarding intelligence quotient (WIQ) μ_w for wizards and witches and μ_e for elves. She can now call the difference between the population means $\delta = \mu_w - \mu_e$ so that δ captures how much *more* intelligent magical folk are. If wizards and elves are equally intelligent, $\delta = 0$. If they are not, δ can take on nonzero values. We can restate this as an hypothesis of approximately no difference (\mathcal{M}_0), an hypothesis of substantial positive difference (\mathcal{M}_+ ; magical folk much more intelligent than elves), and an hypothesis of substantial negative difference (\mathcal{M}_- ; elves much more intelligent than magical folk):

$$\mathcal{M}_0 : \delta \approx 0$$

$$\mathcal{M}_+ : \delta > 0$$

$$\mathcal{M}_- : \delta < 0.$$

However, it is not enough to state simply that $\delta < 0$ because as a model for data, it is underspecified: no quantitative predictions follow (i.e., the likelihood for a specific data set cannot be calculated). In order to be more specific, Bones consults with W.I.S.H.E.S. and together they decide on three concrete models:¹³

$$\begin{aligned} p(\delta|\mathcal{M}_0) &= I(-5 < \delta < 5)/10 && \text{if } -5 < \delta < 5 \\ p(\delta|\mathcal{M}_+) &= 2N(\delta|5, 15)I(\delta > 5) && \text{if } \delta > 5 \\ p(\delta|\mathcal{M}_-) &= 2N(\delta|-5, 15)I(\delta < -5) && \text{if } \delta < -5. \end{aligned}$$

¹³ $I(\cdot)$ is the *indicator function*, which takes the value 1 if its argument is true and 0 otherwise; here it takes the role of a truncation. Since these distributions are truncated, they must be multiplied by a suitable constant such that they integrate to 1 (i.e., we *renormalize* them to be proper distributions).

\mathcal{M}_0 is the assumption that the true difference δ is somewhere between -5 and 5 with all values equally likely – a uniform distribution. This is based on a consensus among W.I.S.H.E.S. that differences of only five WIQ points are negligible for the Ministry’s classification purposes: differences in this range are *practically equivalent to zero*. Under \mathcal{M}_+ , it is assumed that wizards score at least 5 points higher than elves on average ($\delta > 5$) but differences of 20 are not unexpected and differences of 40 possible, if unlikely. Under \mathcal{M}_- , it is assumed that wizards score at least 5 points *lower* than elves ($\delta < -5$).

After having determined the three hypotheses that W.I.S.H.E.S. wishes to consider, Bones decides to collect one more piece of information: how strongly each member of the committee believes in each of the three options. She provides each member with 100 tokens and three cups, and gives them the following instructions:

I would like you to distribute these 100 tokens over these three cups. The first cup represents \mathcal{M}_- , the second \mathcal{M}_0 , and the third \mathcal{M}_+ . You should distribute them proportionally to how strongly you believe in each hypothesis.

Marchbanks’ inferred prior probabilities of each of the three hypotheses are $(25, 50, 25)$, Granger’s are $(15, 70, 15)$, and Runcorn’s are $(5, 15, 80)$. For more in-depth discussion on prior elicitation, see Garthwaite, Kadane, and O’Hagan (2005) and Lee and Vanpaemel (this volume).

To summarize the different prior expectations, Bones constructs a figure to display the marginal distribution of the effect size δ for each committee member. This marginal prior density is easily obtained with the sum rule:

$$\begin{aligned} p(\delta) &= \sum_{h \in (\mathcal{M}_-, \mathcal{M}_0, \mathcal{M}_+)} p(h)p(\delta|h) \\ &= p(\mathcal{M}_-)p(\delta|\mathcal{M}_-) + p(\mathcal{M}_0)p(\delta|\mathcal{M}_0) + p(\mathcal{M}_+)p(\delta|\mathcal{M}_+). \end{aligned}$$

Figure 1.6 shows the resulting distribution for each of the committee members. These graphs serve to illustrate the relative support each committee member’s prior gives to each possible population difference.

Using a well-calibrated test, Bones sets out to gather a sample of $n_1 = 100$ magical folk and $n_2 = 100$ house-elves, and obtains WIQ scores of $M_w = 99.00$ for wizards and witches and $M_e = 101.00$ for elves, giving a sample difference of $d = -2.00$. The test is calibrated such that the standard deviation for magical folk and elves are both equal to 15: $\sigma_w = \sigma_e = 15.00$, which in turn gives a standard deviation for their difference δ of $\sigma_\delta = \sqrt{15^2 + 15^2} = 21.21$. Therefore, the standard error of measurement is $s_e = 21.21/\sqrt{n_1 + n_2} = 1.50$ and the likelihood function to use is now $N(d|\delta, s_e^2) = N(-2|\delta, 1.5^2)$.

To address the committee’s question, Bones can now use Equation 1.20 to obtain the posterior probability of each model:

$$P(\mathcal{M}_i|d) = \frac{p(\mathcal{M}_i)p(d|\mathcal{M}_i)}{P(\mathcal{M}_0)p(d|\mathcal{M}_0) + P(\mathcal{M}_-)p(d|\mathcal{M}_-) + P(\mathcal{M}_+)p(d|\mathcal{M}_+)}.$$

For this, she needs to compute the three marginal likelihoods $p(d|\mathcal{M}_0)$, $p(d|\mathcal{M}_-)$, and $p(d|\mathcal{M}_+)$, which are obtained with the continuous sum rule. For the case of \mathcal{M}_0 , the marginal

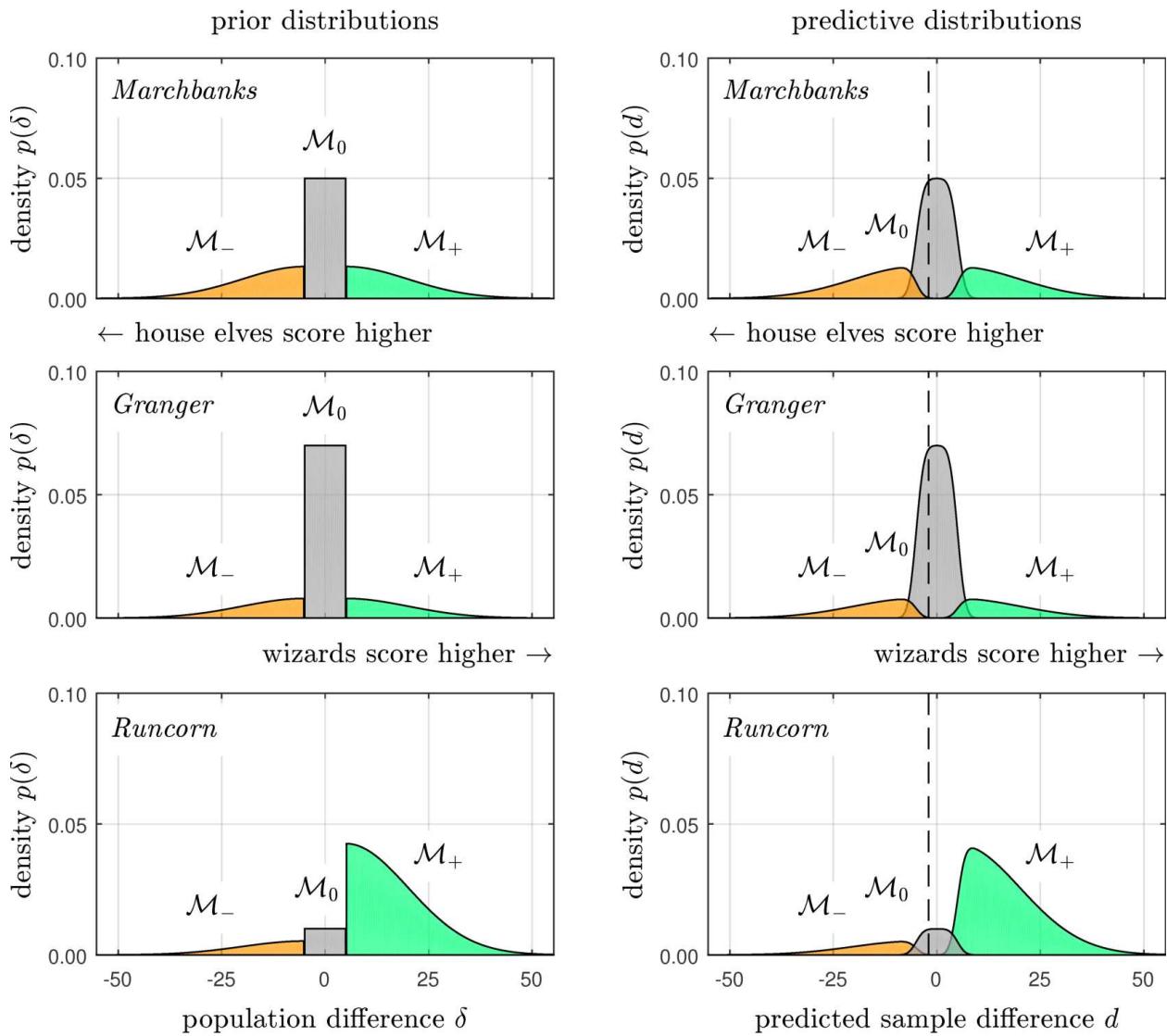


Figure 1.6: **Left:** Each of the three panel members has their own prior probability on each of the three possible models \mathcal{M}_- , \mathcal{M}_0 , and \mathcal{M}_+ . In this scenario, the three models do not overlap in the parameter space: no parameter value is supported by more than one model. However, this is merely a convenient feature of this example and not a requirement of Bayesian model selection – it is entirely possible (and common) for two different models to support the same parameter value. **Right:** The predicted observed difference in a sample with a standard error of estimation of 1.5. Here, the predictive distribution for each model has been multiplied by the prior probability for that model. This representation has the interesting property that the posterior ratio between two models, given some observed difference, can be read from the figure as the ratio between the heights of the two corresponding densities. Note, for example, that at the dashed vertical line (where $d = 2$), the posterior probability for \mathcal{M}_0 is higher than that for \mathcal{M}_- or \mathcal{M}_+ for every judge. If the distributions had not been scaled by the prior probability, these height ratios would give the Bayes factor.

likelihood can be worked out by hand in a few steps:¹⁴

$$\begin{aligned}
 p(d|\mathcal{M}_0) &= \int_{\Delta} p(\delta|\mathcal{M}_0) \times p(d|\delta, \mathcal{M}_0) d\delta \\
 &= \int_{\Delta} \frac{1}{10} I(-5 < \delta < 5) \times N(d|\delta, s_e^2) d\delta \\
 &= \frac{1}{10} \int_{-5}^5 N(d|\delta, s_e^2) d\delta \\
 &= \frac{1}{10} [\Phi(2| -5, 1.5^2) - \Phi(2| 5, 1.5^2)] \\
 &= 9.772 \times 10^{-2}
 \end{aligned}$$

For the cases of \mathcal{M}_+ and \mathcal{M}_- , the derivation is much more tedious. It can be done by hand by making use of the fact that the product of two normal distributions has a closed-form solution. However, a numerical approximation can be very conveniently performed with standard computational software or—at the Ministry of Magic—a simple numerical integration spell.¹⁵ For this particular task, Dr. Bones arrives at $p(d|\mathcal{M}_+) = 8.139 \times 10^{-8}$ and $p(d|\mathcal{M}_-) = 1.209 \times 10^{-3}$.

Bones now has all that she needs to compute the posterior probabilities of each hypothesis and for each committee member. The prior and posterior probabilities are given in Table 1.3. As it turns out, the data that Bones has available should effectively overwhelm each of the three members' prior probabilities and put the bulk of the posterior probability on \mathcal{M}_0 for each member. Counting on the ability of each committee member to rationally update their beliefs, she prepares a concise presentation in which she lays out a confident case for elf equality and “Being” status.

Discussion Probability theory allows model comparison in a wide variety of scenarios. In this example the psychometrician deals with a set of three distinct models, each of which was constructed ad hoc – custom-built to capture the psychological intuition of the researcher (and a review panel). Once the models were built, the researcher had only to “turn the crank” of probabilistic inference and posterior probabilities are obtained through standard mechanisms that rely on little other than the sum and product rules of probability. As this example illustrates, the practical computation of posterior probabilities will often rely on calculus or numerical integration methods; several papers in this special issue deal with computational software that is available (Wagenmakers, Love, et al., this volume; Matzke et al., this volume; van Ravenzwaaij et al., this volume; Wagenmakers, Marsman, et al., this volume).

¹⁴Bones' derivation makes use of the fact that the identity function $I(\cdot)$ can be factored out of the integrand if the integration bounds are accordingly limited to the region where the argument is true. This fact is used in moving from the second step to the third.

¹⁵Some popular non-magical options include MATLAB (The Mathworks, Inc., 2015) and R (R Development Core Team, 2004), or readers can use www.wolframalpha.com. MATLAB and R code for this example is available on the OSF repository and in the Appendix.

Table 1.3: Prior and posterior probabilities for each hypothesis and each committee member. Probabilities are updated with Equation 1.20. The fourth row in each half of the table serves to emphasize that, for the purposes of the committee, $P(\mathcal{M}_-)$ and $P(\mathcal{M}_0)$ constitute a single category since they both lead to the classification of “Being” rather than “Beast.” Thus, we consider $P(\text{“Being”}) = P(\mathcal{M}_-) + P(\mathcal{M}_0)$.

	Marchbanks	Granger	Runcorn
$P(\mathcal{M}_-)$.250	.150	.050
$P(\mathcal{M}_0)$.500	.700	.150
$P(\mathcal{M}_+)$.250	.150	.850
$P(\text{“Being”})$.750	.850	.200
$P(\mathcal{M}_- d)$.006	.003	.012
$P(\mathcal{M}_0 d)$.994	.997	.988
$P(\mathcal{M}_+ d)$.000	.000	.000
$P(\text{“Being”} d)$	1.000	1.000	1.000

An interesting aspect to this example is the fact that the analyst is asked to communicate to a diverse audience: three judges who hold different prior notions about the crucial hypotheses. That is, they hold different notions on the *prior probability that each hypothesis is true*. They happen to agree on the *prior distribution of the δ parameter* under each hypothesis (but we made that simplification only for ease of exposition; it is not a requirement of the method). This is comparable to the situation in which most researchers find themselves: there is one data set that brings evidence, but there are many—possibly diverse—prior notions. Given that prior probabilities must be subjective, how can researchers hope to reasonably communicate their results if they can only report their own subjective knowledge?

One potential strategy is the one employed by the psychometrician in the example. The strategy relies on the realization that we can compute posterior probabilities for *any* rational person as soon as we know their prior probabilities. Because the psychometrician had access to the prior probabilities held by each judge, she was able to determine whether her evidence would be compelling to this particular audience.

Social scientists who present evidence to a broad audience can take a similar approach by *formulating multiple prior distributions* – for example, some informative priors motivated by theory, some priors that are uninformative or indifferent in some ways, and some priors that might be held by a skeptic. Such a practice would be a form of *sensitivity analysis* or *robustness analysis*. If the data available are sufficiently strong that skeptics of all camps must rationally come to the same conclusion, then concerns regarding the choice of priors are largely alleviated. This was the case above, where Marchbanks, Granger, and Runcorn all were left with a greater than 98% posterior probability for the model specifying elf equality despite their wide-ranging prior probabilities.

Of course, data is often noisy and the evidence may in many cases not be sufficient to convince the strongest skeptics. In such cases, collecting further data may be useful. Otherwise, the researcher can transparently acknowledge that reasonable people could reasonably come to different conclusions.

An alternative option is to report the evidence in isolation. Especially when the ultimate

claim is binary—a discrimination between two models—one might report only the amount of discriminating evidence for or against a model. By reporting only the amount of evidence, in the form of a Bayes factor, every individual reader can combine that evidence with their own prior and form their own conclusions. This is now a widely-recommended approach (e.g., Wagenmakers, Marsman, et al., this volume; but see Robert, 2016, for words of caution; and see Kruschke & Liddell, this volume, for a discussion of scenarios in which the Bayes factor should not be the final step of an analysis) that is taken in the final example.

Example 7: “Luck of the Irish”

Every four years, the wizarding world organizes the most exhilarating sporting event on earth: the Quidditch World Cup. However, the Cup is often a source of controversy. In a recent edition, aspersions were cast on the uncommonly strong showing by the Irish team: An accusation was brought that the Irish players were dosed with a curious potion called *felix felicis*, which gives an individual an extraordinary amount of “dumb luck.”

At the Ministry of Magic’s Department for International Magical Cooperation—who oversee the event and have decided to investigate the doping claims—junior statistician Angelina Johnson noticed that the Irish team had *another* striking piece of good luck: in each of the four games, the Irish team captain won the coin toss that allows them to choose in which direction to play. From these data, Johnson reasons as follows.

If the coin is fair, and there is no cheating, then the Irish team captain should win the toss with 50% probability on each occasion ($\mathcal{M}_0 : \theta = \theta_0 = 0.5$). However, if the captain has taken *felix felicis*, they should win with a higher, but unknown probability ($\mathcal{M}_J : \theta > 0.5$). Johnson then sets out to determine whether this small amount of data ($k = 4$ wins in $N = 4$ games) contains enough evidence to warrant strong suspicions.

The discriminating evidence is given by the Bayes factor, $BF_{J0} = P(k|\mathcal{M}_J)/P(k|\mathcal{M}_0)$, where the marginal likelihoods (with capital $P(\cdot)$ since number of wins are discrete) can be calculated one model at a time. Since the outcomes of the four coin tosses are assumed independent given θ , the probability of k successes in any sequence of length N is given by the binomial distribution: $\binom{N}{k} \theta^k (1-\theta)^{N-k}$, where the binomial coefficient $\binom{N}{k}$ is the number of ways N items can arrange themselves in groups of size k (e.g., 4 items can be arranged into a group of 4 exactly 1 way). Thus, for \mathcal{M}_0 ,

$$\begin{aligned} P(k|\mathcal{M}_0) &= \binom{4}{4} 0.5^4 \times 0.5^0 \\ &= \frac{1}{2^4} = \frac{1}{16}. \end{aligned}$$

For \mathcal{M}_J , Johnson needs to express her prior knowledge of the parameter θ . Since she knows very little about the potion *felix felicis*, she takes all values between 0.5 and 1.0 to be equally plausible, so that $P(\theta|\mathcal{M}_J) = 2I(0.5 < \theta < 1.0)$. The shape of this prior density is depicted

in the left half of Figure 1.7. Hence,

$$\begin{aligned}
 P(k|\mathcal{M}_J) &= \int_{\Theta} p(\theta|\mathcal{M}_J) \times P(k|\theta, \mathcal{M}_J) d\theta \\
 &= \int_{\Theta} 2I(0.5 < \theta < 1.0) \times \binom{4}{4} \theta^4 (1-\theta)^0 d\theta \\
 &= 2 \int_{0.5}^{1.0} \theta^4 d\theta \\
 &= 2 \left[\frac{\theta^5}{5} \right]_{0.5}^{1.0} = \frac{2}{5} (1^5 - 0.5^5) = \frac{31}{80}
 \end{aligned}$$

Thus, the data are implied $(31/80) / (1/16) = 6.2$ times more strongly by \mathcal{M}_J than by \mathcal{M}_0 (i.e., $BF_{J0} = 6.2$). Johnson concludes that these data afford only a modest amount of evidence—certainly not enough evidence to support a controversial and consequential recommendation—and decides to return to tallying quidditch-related nose fractures instead.

Example 7b: “Luck of the Irish — Part 2”

As might be expected, the Irish quidditch controversy did not fail to pique interest throughout the wizarding world. Independently of the Ministry statistician, Barnabas Cuffe, Editor-in-Chief of the *Daily Prophet*—England’s premier magical newspaper—had noticed the same peculiar luck in the Irish team’s pregame coin tosses. In the editor’s case, however, attention to the coin tosses was not a coincidence – in fact, “liquid luck” had helped him win a few career-saving coin tosses in a mildly embarrassing part of his journalistic past.

Cuffe’s experience with *felix felicis* is straightforward: on eleven different occasions did he sip the potion just before a toin coss would decide which of two journalistic leads he would pursue that day – his colleague would pursue the other. He recalls clearly that on each of the eleven occasions, his leads carried him in the thick of dramatic, newsworthy events while his colleague’s leads turned out dead ends. Cuffe was promoted; his colleague dismissed.

As it happens, Cuffe is an accomplished statistician, and he reasons in exactly the same way as Angelina Johnson (the junior statistician at the Ministry). If there is no cheating the winning probability should be 50% each time ($\mathcal{M}_0 : \theta = 0.5$). If there *is* cheating, the winning probability should be higher. In contrast to Johnson, however, Cuffe has a good idea how much higher the winning probability θ will be with *felix felicis*: before evaluating the Irish captain’s luck he can estimate θ from additional information y that only he possesses.

Cuffe starts by writing down Equation 1.10 and filling in the quantities on the right hand side. Among these is the prior density $p(\theta)$, which gives the density at each possible value of θ *before considering his own eleven winning coin tosses y*. A reasonable place to start (as before) is that all values between 0.5 and 1.0 are equally plausible: $p(\theta) = 2I(0.5 < \theta < 1.0) = 2I_\theta$ (where we introduce I_θ as a shorthand for $I(0.5 < \theta < 1.0)$, the appropriate

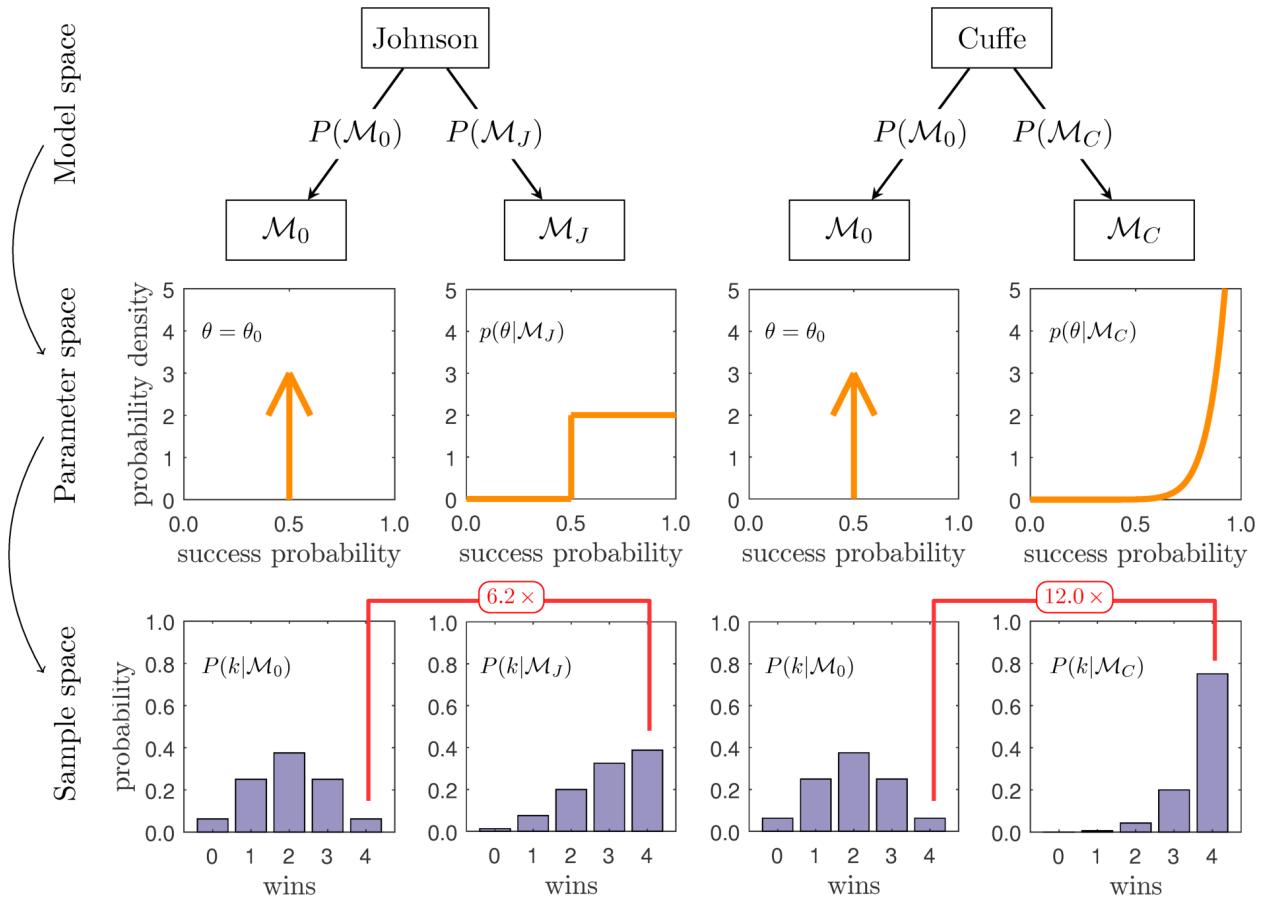


Figure 1.7: The structure of Johnson and Cuffe's models, which can be viewed as more complex (rotated) versions of earlier path diagrams. **Top:** The model space shows the contending models. In this case, both Johnson and Cuffe are comparing two models. The prior probabilities for the models are left unspecified. **Middle:** The parameter space shows what each model predicts about the true value of θ (i.e., each model's *conditional prior distribution*). Johnson and Cuffe both use a point null model, which packs all of its mass into a single point (shown as the arrow spiking at $\theta = .5$). However, they have different background knowledge about *felix felicis*, so their prior distributions for θ under their respective alternative model differ. Note that $p(\theta|\mathcal{M}_C)$ is obtained from updating $p(\theta|\mathcal{M}_J)$ with 11 additional *felix felicis* successes. **Bottom:** The sample space shows what each model predicts about the data to be observed (i.e., each model's prior predictive distribution). The Bayes factor is formed by taking the ratio of the probability each model attached to the observed data, which was four wins in four coin tosses. Since the predictions from the null model are identical for Cuffe and Johnson, the difference in their Bayes factors is due to the higher marginal likelihood Cuffe's alternative model placed on the Irish captain winning all four coin tosses.

indicator function). He also uses the same binomial likelihood function as Johnson, hence,

$$\begin{aligned}
 p(\theta|y) &= \frac{p(\theta) \times p(y|\theta)}{\int_{\Theta} p(\theta) \times p(y|\theta) d\theta} \\
 &= \frac{2I_{\theta} \times \binom{11}{11}\theta^{11}(1-\theta)^0}{\int_{\Theta} 2I_{\theta} \times \binom{11}{11}\theta^{11}(1-\theta)^0 d\theta} = \frac{2I_{\theta} \times \theta^{11}}{2 \int_{0.5}^{1.0} \theta^{11} d\theta} \\
 &= \frac{I_{\theta} \times \theta^{11}}{\left[\frac{\theta^{12}}{12}\right]_{0.5}^{1.0}} = \frac{I_{\theta} \times \theta^{11}}{\frac{1}{12}(1.0^{12} - 0.5^{12})} \approx 12\theta^{11}I_{\theta}
 \end{aligned}$$

This calculation¹⁶ yields Cuffe's *posterior density* of the winning probability θ , which captures his knowledge and uncertainty of the value of θ under luck doping. The shape of this density function is depicted in the right half of Figure 1.7. Crucially, Cuffe can use this knowledge to perform the same analysis as the Ministry statistician with only one difference: **yesterday's posterior $p(\theta|y)$ is today's prior $p(\theta|\mathcal{M}_C)$** . The fact that the latter notation of the prior does not include mention of y serves to illustrate that densities and probabilities are often implicitly conditional on (sometimes informal) background knowledge. Note, for instance, that the entire calculation above assumes that *felix felicis* was taken, but this is not made explicit in the mathematical notation.

Unknowingly repeating Johnson's calculation, Cuffe finds that the probability of the Irish team captain's $k = 4$ winning coin tosses assuming no luck doping is again $p(k|\mathcal{M}_0) = 1/16$. His calculation for the probability of the $k = 4$ wins assuming luck doping is

$$\begin{aligned}
 P(k|\mathcal{M}_C) &= \int_{\Theta} p(\theta|\mathcal{M}_C) \times p(k|\theta, \mathcal{M}_C) d\theta \\
 &\approx \int_{0.5}^{1.0} 12\theta^{11}I_{\theta} \times \binom{4}{4}\theta^4(1-\theta)^0 d\theta \\
 &= 12 \left[\frac{\theta^{16}}{16} \right]_{0.5}^{1.0} = \frac{12}{16} (1^{16} - 0.5^{16}) \approx \frac{12}{16}
 \end{aligned}$$

To complete his analysis, Cuffe takes the ratio of marginal likelihoods,

$$BF_{C0} = P(k|\mathcal{M}_C)/P(k|\mathcal{M}_0) \approx 12,$$

which is strong—but not very strong—evidence in favor of Cuffe's luck doping model.

Inspired partly by the evidence and partly by the recklessness that follows from years of *felix felicis* abuse, editor Cuffe decides to publish an elaborate exposé condemning both the Irish quidditch team for cheating and the Ministry of Magic for failing to act on strong evidence of misconduct.

¹⁶Note that here and below, we make use of a convenient approximation: $0.5^k \approx 0$ for large values of k . Making the calculation exact is not difficult but requires a rather unpleasant amount of space. Also note that the indicator function from the prior density carries over to the posterior density.

Discussion This final, two-part example served mostly to illustrate the effects of prior knowledge on inference. This is somewhat in contrast to Example 6, where the prior information was overwhelmed by the data. In the two scenarios here, the Ministry junior statistician and the *Prophet* editor are both evaluating evidence that discriminates between two models. Both consider a “nil model” in which all parameters are known (the fairness of a coin implies that the parameter θ must be 0.5), but they critically differ in their definition of the alternative model. The Ministry statistician, having no particular knowledge of the luck doping potion, considers all better-than-chance values equally plausible, whereas the *Prophet* editor can quantify and insert relevant prior information that specifies the expected effects of the drug in question to greater precision.

As illustrated in the bottom row of Figure 1.7, these three models (the chance model \mathcal{M}_0 , the Ministry model \mathcal{M}_J , and the *Prophet* model \mathcal{M}_C) make distinct predictions: \mathcal{M}_0 predicts a distribution of Irish coin toss wins that is symmetric about $k = 2$; \mathcal{M}_J predicts a right-leaning distribution with a greater probability of four Irish wins; and \mathcal{M}_C predicts an even greater such probability. More specifically, the marginal likelihoods are $P(k|\mathcal{M}_0) = 5/80$, $P(k|\mathcal{M}_J) = 31/80$, and $P(k|\mathcal{M}_C) \approx 60/80$, and the Bayes factor between any two of these models is given by forming the appropriate ratio.

This example illustrates a general property in Bayesian model comparison: A model that makes precise predictions can be confirmed to a much stronger extent than a model that makes vague predictions, while at the same time the precision of its predictions makes it easier to disconfirm. The reason Cuffe was able to obtain a higher Bayes factor than Johnson is because his alternative model made much more precise predictions; \mathcal{M}_C packed three-quarters of its prior predictive distribution into $k = 4$, whereas \mathcal{M}_J spread its probability more broadly among the potential outcomes. Since Cuffe’s precise prediction was correct, he was rewarded with a larger Bayes factor. However, Cuffe’s prediction was risky: if the Irish captain had won any fewer than all four coin tosses, \mathcal{M}_0 would have been supported over \mathcal{M}_C . In contrast, the Bayes factor would still favor \mathcal{M}_J when $k = 3$ because Johnson’s model is more conservative in its predictions. In sum, the ability to incorporate meaningful theoretical information in the form of a prior distribution allows for more informed predictions and hence more efficient inferences (Lee & Vanpaemel, this volume).

Broader appeal and advantages of Bayesian inference

The Bayesian approach is a common sense approach. It is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data.

W. Edwards et al. (1963)

In our opinion, the greatest theoretical advantage of Bayesian inference is that it unifies all statistical practices within the consistent formal system of probability theory. Indeed, the unifying framework of Bayesian inference is so uniquely well-suited for scientific inference that these authors see the two as synonymous. Inference is the process of combining multiple

sources of information into one, and the rules for formally combining information derive from two simple rules of probability. Inference can be as straightforward as determining the event of interest (in our notation, usually \mathcal{M} or θ) and the relevant data and then exploring what the sum and product rules tell us about their relationship.

As we have illustrated, common statistical applications such as parameter estimation and hypothesis testing naturally emerge from the sum and product rules. However, these rules allow us to do much more, such as make precise quantitative predictions about future data. This intuitive way of making predictions can be particularly informative in discussions about what one should expect in future studies – it is perhaps especially useful for predicting and evaluating the outcome of a replication attempt, since we can derive a set of new predictions after accounting for the results of the original study (e.g., Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, & Ly, 2016).

The practical advantages of using probability theory as the basis of scientific and statistical inference are legion. One of the most appealing in our opinion is it allows us to make probabilistic statements about the quantities of actual interest, such as “There is a 90% probability the participants are guessing,” or “The probability is .5 that the population mean is negative.” It also allows us to construct hierarchical models that more accurately capture the structure of our data, which often includes modeling theoretically-meaningful variability at the participant, task, item, or stimulus level (Gelman & Hill, 2007; Lee & Wagenmakers, 2013; Rouder, Morey, & Pratte, In Press).

Bayesian inference also gracefully handles so-called *nuisance parameters*. In most of our present examples there has been only a single quantity of interest – in order to help keep the examples simple and easy to follow. In real applications, however, there are typically many parameters in a statistical model, some of which we care about and some of which we do not. The latter are called nuisance parameters because we have little interest in them: we only estimate them out of necessity. For example, if we were estimating the mean of a normal distribution (as in Example 4) and did not know the population standard deviation, then we would have to assign it a prior density, such that the overall prior density would be of the form $p(\mu, \sigma)$; after collecting data X , the posterior density would be of the form $p(\mu, \sigma|X)$. Since we are generally only interested in the parameter μ , estimating σ out of necessity, σ is considered a nuisance parameter. To make inferences about μ we merely integrate out σ from the posterior density using the sum rule: $p(\mu|X) = \int_{\Sigma} p(\mu, \sigma|X)d\sigma$, from which we can do inference about μ . Similarly, in Examples 7 and 7b, the exact win rate from a luck-doped coin toss is not of primary interest, only whether the coin tossed in the four games was plausibly fair or not. Here, the bias parameter of the coin can be seen as a nuisance parameter. Dealing with nuisance parameters in a principled way is a unique advantage of the Bayesian framework: except for certain special cases, frequentist inference can become paralyzed by nuisance parameters.

The ability of Bayesian inference to deal with nuisance parameters also allows it to flexibly handle one of the biggest statistical challenges for data analysts: situations in which the assumptions of the statistical model regarding the data are badly violated. For example, one of the most common assumptions violated is that of normality (e.g., due to the presence of many outliers). In technical terms, this means that we may not think the normal likelihood function adequately characterizes the data-generating mechanism for the inference problem

at hand. In Bayesian inference the choice of likelihood is important because, as we have seen in the estimation examples above, with even moderate samples sizes the likelihood quickly begins to dominate the prior densities. To resolve this issue a Bayesian can construct two models: one that uses a normal likelihood function (model \mathcal{M}_N), and one that uses a likelihood function with wider tails (model \mathcal{M}_W), such as a t distribution with few degrees of freedom. After collecting data we then have a posterior distribution for the parameters of interest for each model, $p(\theta|X, \mathcal{M}_N)$ and $p(\theta|X, \mathcal{M}_W)$. If we assign prior probabilities to these two models (we emphasize that a “model” consists of both a prior distribution for the parameters and a likelihood function for the data), $P(\mathcal{M}_N)$ and $P(\mathcal{M}_W)$, we can calculate their posterior probabilities $P(\mathcal{M}_N|X)$ and $P(\mathcal{M}_W|X)$. We are then in a position to use the sum rule to marginalize over the different models (as Dr. Bones did with the various prior densities in Example 6), allowing us to find the *model-averaged* posterior density for θ ,

$$p(\theta|X) = P(\mathcal{M}_N|X)p(\theta|X, \mathcal{M}_N) + P(\mathcal{M}_W|X)p(\theta|X, \mathcal{M}_W).$$

Note that model averaging is in a sense the flip-side of model selection: In model selection, the identity of the model is central while the *model parameters* are sometimes seen as nuisance variables to be integrated away. By contrast, in the previous equation the *model identities* are treated as nuisance variables while the shared model parameters remain central (see Roberts, 1965; Etz & Wagenmakers, in press). The flexibility to perform model averaging across any variable we care to name (e.g. Hoeting, Madigan, Raftery, & Volinsky, 1999; Link & Barker, 2009) is a unique advantage of Bayesian inference.

Finally, Bayesian analysis allows for immense freedom in data collection because it respects the *likelihood principle* (Berger & Wolpert, 1988). The likelihood principle states that the likelihood function of the data contains all of the information relevant to the evaluation of statistical evidence. What this implies is that other properties of the data or experiment that do not factor into the likelihood function are *irrelevant* to the statistical inference based on the data (Lindley, 1993; Royall, 1997). Adherence to the likelihood principle means that one is free to do analyses without needing to adhere to rigid sampling plans, or even have any plan at all (Rouder, 2014). Note that we did not consider the sampling plan in any of our examples above, and none of the inferences we made would have changed if we had. Within a Bayesian analysis, “It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience” (Edwards et al., 1963, p. 193).

Conclusion

[W]e believe that Bayes’ theorem is not only useful, but in fact leads to the *only* correct formulas for solving a large number of our cryptanalytic problems.

F. T. Leahy (1960) [emphasis original]

The goal of this introduction has been to familiarize the reader with the fundamental principles of Bayesian inference. Other contributions in this special issue (Dienes & McLatchie,

this volume; Kruschke & Liddell, this volume) focus on why and how Bayesian methods are preferable to the methods proposed in the *New Statistics* (Cumming, 2014). The Bayesian approach to all inferential problems follows from two simple formal laws: the sum and product rules of probability. Taken together and in their various forms, these two rules make up the entirety of Bayesian inference – from testing simple hypotheses and estimating parameters, to comparing complex models and producing quantitative predictions.

The Bayesian method is unmatched in its flexibility, is rooted in relatively straightforward calculus, and uniquely allows researchers to make statements about the relative probability of theories and parameters – and to update those statements with more data. That is, the laws of probability show us how our scientific opinions can evolve to cohere with the results of our empirical investigations. For these reasons, we recommend that social scientists adopt Bayesian methods rather than the *New Statistics*, and we hope that the present introduction will contribute to deterring the field from taking an evolutionary step in the wrong direction.

Acknowledgments

The authors would like to thank J. P. de Ruiter, Stephan Franke, and Zita Oravecz for helpful comments, Brian Clayton for the Illustration, and J. K. Rowling for the Harry Potter universe. The authors were supported by NSF grants #1230118 and #1534472 from the Methods, Measurements, and Statistics panel and by John Templeton Foundation grant #48192. AE was further supported by the National Science Foundation Graduate Research Fellowship Program (#DGE-1321846).

References

- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward (CA): Institute of Mathematical Statistics.
- Consonni, G., & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23, 332–353.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
- de Finetti, B. (1974). *Theory of probability*, vol. 1. New York: John Wiley & Sons.
- Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42, 204–223.
- Dienes, Z., & McLatchie, N. (this volume). Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review*.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Efron, B., & Morris, C. (1977). Stein’s paradox in statistics. *Scientific American*, 236, 119–127.

- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project. *PLOS ONE*, 11, e0149794.
- Etz, A., & Wagenmakers, E.-J. (in press). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*.
- Evans, M. (2014). Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle". *Statistical Science*, 29(2), 242–246.
- Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470), 680–701.
- Gelman, A. (2010). Bayesian statistics then and now. *Statistical Science*, 25(2), 162–165.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton (FL): Chapman & Hall/CRC.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 55–61.
- Hill, B. M. (1974). On coherence, inadmissibility and inference about many parameters in the theory of least squares. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics* (pp. 555–584). North-Holland Amsterdam.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.
- Jaynes, E. T. (1984). The intuitive inadequacy of classical statistics. *Epistemologia*, 7(43), 43–74.
- Jaynes, E. T. (1986). Bayesian methods: General background. In J. H. Justice (Ed.), *Maximum entropy and bayesian methods in applied statistics* (pp. 1–25). Cambridge University Press.
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophy Society*, 31, 203–222.
- Jeffreys, H. (1939). *Theory of probability* (1 ed.). Oxford, UK: Oxford University Press.
- Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
- Jeffreys, H. (1973). *Scientific inference* (3 ed.). Cambridge, UK: Cambridge University Press.
- Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review*, 121(2), 206.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kruschke, J., & Liddell, T. (this volume). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and planning from a Bayesian perspective. *Psychonomic Bulletin and Review*.
- Leahy, F. (1960). Bayes marches on. *National Security Agency Technical Journal*, 5(1), 49–61.
- Lee, M. D., & Vanpaemel, W. (this volume). Determining informative priors for cognitive models. *Psychonomic Bulletin and Review*.

- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lindley, D. V. (1985). *Making decisions* (2 ed.). London: Wiley.
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15, 22–25.
- Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49, 293–337.
- Link, W. A., & Barker, R. J. (2009). Bayes factors and multimodel inference. In *Modeling demographic processes in marked populations* (pp. 595–615). Springer.
- 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.
- Marin, J.-M., & Robert, C. P. (2010). On resolving the Savage–Dickey paradox. *Electronic Journal of Statistics*, 4, 643–654.
- Marsman, M., & Wagenmakers, E.-J. (2016). Three insights from a Bayesian interpretation of the one-sided P value. *Educational and Psychological Measurement*.
- Matzke, D., Boehm, U., & Vandekerckhove, J. (this volume). Bayesian inference for psychology, part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin and Review*.
- 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(1), 103–123.
- Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6–18.
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, 36, 97–131.
- Open Science Collaboration, T. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria.
- Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.
- Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72, 33–37.
- Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, 60(309), 50–62.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301–308.
- Rouder, J. N., & Morey, R. D. (2011). A Bayes–factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18, 682–689.
- Rouder, J. N., Morey, R. D., & Pratte, M. S. (In Press). Bayesian hierarchical models. In *New handbook of mathematical psychology, volume. 1: Measurement and methodology*. Cambridge University Press.
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, 8, 520-547.

- Rouder, J. N., & Vandekerckhove, J. (this volume). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin and Review*.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
- Scamander, N. A. F. (2001). *Fantastic beasts and where to find them*. London, UK: Obscurus Books.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64, 583–639.
- The Mathworks, Inc. (2015). *MATLAB version R2015a*. Natick, MA.
- van Ravenzwaaij, D., Cassey, P., & Brown, S. (this volume). A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin and Review*.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.
- Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6, 142–228.
- Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143, 1457–1475.
- 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, J., et al. (this volume). Bayesian inference for psychology, part II: Example applications with JASP. *Psychonomic Bulletin and Review*.
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (this volume). Bayesian inference for psychology, part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*.
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (in press). Bayesian benefits for the pragmatic researcher. *Perspectives on Psychological Science*.
- Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48, 413–426.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92–107.
- Winkler, R. L. (1972). *An introduction to Bayesian inference and decision*. Holt, Rinehart and Winston New York.
- Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, 38, 715–731.
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42, 369–390.

Appendix: Computer code for “The measure of an elf”

<pre> 1 % MATLAB/Octave code for Example 6 2 3 % Define the three models 4 e = 5; 5 s = 15; 6 7 h0 = @(x) (x<e & x>-e) / (2*e); 8 hn = @(x) normpdf(x,-e,s)*2.*(x<-e); 9 hp = @(x) normpdf(x,e,s)*2.*(x> e); 10 11 % Define the data and likelihood 12 d = -2; 13 n = 100; 14 sem = sqrt((s^2 + s^2) / (2*n)); 15 16 likelihood = @(x) normpdf(d,x,sem); 17 18 % Define the integrands and integrate 19 fn = @(x)likelihood(x).*hn(x); 20 mn = quadgk(fn,-inf,-e,'waypoints',[-e,e]); 21 22 f0 = @(x)likelihood(x).*h0(x); 23 m0 = quadgk(f0,-e,e,'waypoints',[-e,e]); 24 25 fp = @(x)likelihood(x).*hp(x); 26 mp = quadgk(fp,e,inf,'waypoints',[-e,e]); 27 28 ev = [mn,m0,mp]; 29 30 % Apply Bayes' rule 31 eq19 = @(p,m) p.*m ./ sum(p.*m); 32 33 marchbanks = [.25,.50,.25]; 34 granger = [.15,.70,.15]; 35 runcorn = [.45,.10,.45]; 36 37 eq19(marchbanks,ev) 38 % ans = 0.0061 0.9939 0.0000 39 eq19(granger,ev) 40 % ans = 0.0026 0.9974 0.0000 41 eq19(runcorn,ev) 42 % ans = 0.0122 0.9878 0.0000 </pre>	<pre> # R code for Example 6 # # Define the three models e <- 5 s <- 15 h0 <- function(x) (x<e & x>-e) / (2*e) hn <- function(x) dnorm(x,-e,s)*2*(x<-e) hp <- function(x) dnorm(x, e,s)*2*(x> e) # Define the data and likelihood d <- -2 n <- 100 sem <- sqrt((s^2 + s^2) / (2*n)) like <- function(x) dnorm(d,x,sem) # Define the integrands and integrate fn <- function(x) like(x)*hn(x) mn <- integrate(fn,-Inf,-e)\$value f0 <- function(x) like(x)*h0(x) m0 <- integrate(f0,-e,e)\$value fp <- function(x) like(x)*hp(x) mp <- integrate(fp,e,Inf)\$value ev <- c(mn,m0,mp) # Apply Bayes' rule eq19 <- function(p,m) p*m / sum(p*m) marchbanks <- c(.25,.50,.25) granger <- c(.15,.70,.15) runcorn <- c(.10,.10,.80) eq19(marchbanks,ev) # [1] 6.1457e-03 9.9385e-01 4.1385e-07 eq19(granger,ev) # [1] 2.6432e-03 9.9736e-01 1.7799e-07 eq19(runcorn,ev) # [1] 1.2216e-02 9.8778e-01 6.5810e-06 </pre>
---	--

MATLAB/Octave users who do not have access to the Statistics Toolbox can add on line 6:

```
normpdf = @(x,m,s) exp(-((x-m)./s).^2/2)./sqrt(2.*s.^2.*pi);
```


Theoretical advantages and practical ramifications

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly,
Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F. Gronau, Martin
Šmíra, Sacha Epskamp, Dora Matzke, Jeffrey N. Rouder, and Richard D.
Morey

Theoretical satisfaction and practical implementation are the twin ideals of coherent statistics.

D. V. Lindley, 1980

The psychology literature is rife with p values. In almost every published research article in psychology, substantive claims are supported by p values, preferably ones smaller than .05. For instance, the December 2014 issue of *Psychonomic Bulletin & Review* featured 24 empirical brief reports, all of which reported p values. The dominance of the p value statistical framework is so complete that its presence feels almost prescriptive (“every empirical article in psychology shall feature at least one p value.”). In Part I of this two-part series we aim to demonstrate that there exists a valid and feasible alternative –Bayesian inference– whose adoption brings considerable benefits, both in theory and in practice.

Based on a superficial assessment, the continued popularity of p values over Bayesian methods may be difficult to understand. The concept of p value null hypothesis statistical testing (NHST) has been repeatedly critiqued on a number of important points (e.g., W. Edwards, Lindman, & Savage, 1963; Morrison & Henkel, 1970; Mulaik & Steiger, 1997; Wagenmakers, 2007), and few methodologists have sought to defend the practice. One of the critiques is that p values are often misinterpreted as Bayesian posterior probabilities, such that it is all too easy to believe that $p < .05$ warrants the rejection of the null hypothesis \mathcal{H}_0 , and consequently supports the acceptance of the alternative hypothesis \mathcal{H}_1 . This interpretation of p values is tempting but incorrect (Gigerenzer, Krauss, & Vitouch, 2004). A p value is the probability of obtaining results at least as extreme as those observed given that

the null hypothesis is true. The transition from this concept to the decision, “I accept the alternative hypothesis”, is a leap that is logically invalid. The p value does not take into account the prior plausibility of \mathcal{H}_0 , and neither does it recognize the fact that data unusual under \mathcal{H}_0 can also be unusual under \mathcal{H}_1 (Wagenmakers et al., in press). Other pressing problems with p values will be discussed shortly.

From a psychological perspective, however, a number of arguments may help explain the continued popularity of p values over Bayesian methods.¹ First, researchers practice and preach the methodology that they were once taught themselves; interrupting this self-perpetuating educational cycle requires that researchers invest serious effort to learn new methods. Second, by breaking away from the dominant group of p value practitioners, researchers choose to move away from the in-group and expose themselves to the associated risks of academic exclusion. Third, just like fish form schools to escape predation, researchers may believe that there is security in repeating procedures that are popular; surely, they may feel, “if the procedure I use is standard in the field, then any detractors must be overstating their case”. Fourth, many psychologists are primarily interested in addressing substantive research questions, not in the finer details of statistical methodology; such methodological disinterest feeds the desire for simple procedures that work well enough to convince the reviewers. In this sense the current p value fixation is similar to a statistical ritual (i.e., the “null ritual”, Gigerenzer, 2004). Fifth, the p value framework, when misinterpreted, offers a simple solution to deal with the uncertainty inherent in noisy data: when $p < .05$, reject \mathcal{H}_0 and accept \mathcal{H}_1 ; when $p > .10$, retain \mathcal{H}_0 . When misapplied in this way, p values appear to make it easy for researcher to draw strong conclusions even when the empirical results are noisy and uninformative. Sixth, researchers may feel that by using non-standard methods (i.e., anything other than the p value) they reduce their chances of getting their work published or having it understood by their colleagues. Seventh, researchers interested in methodology have often internalized their statistical education to such an extent that they have difficulty accepting that the method they have used all their life may have serious limitations; when new information conflicts with old habits, the resulting cognitive dissonance can be reduced by discounting or ignoring the new information. Finally, it is possible that researchers may agree with the p value critiques, yet are unable to adopt alternative (Bayesian) inferential procedures. The reason for this inability is straightforward: virtually all statistical software packages produce p values easily, whereas Bayesian methods cannot count on the same level of support. Many of these arguments hold for statistical innovations in general, not just for p value NHST (Sharpe, 2013).

In general, then, powerful psychological and societal forces are at play, making it nigh impossible to challenge the dominant methodology. Nonetheless, the edifice of NHST appears to show subtle signs of decay. This is arguably due to the recent trials and tribulations collectively known as the “crisis of confidence” in psychological research, and indeed, in empirical research more generally (e.g., Begley & Ellis, 2012; Button et al., 2013; Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012; Pashler & Wagenmakers, 2012; Simmons, Nelson, & Simonsohn, 2011). This crisis of

¹These arguments are speculative to the degree that they are based entirely on our personal experience and common-sense; in other words, our arguments have not been subjected to rigorous empirical tests.

confidence has stimulated a methodological reorientation away from the current practice of p value NHST. A series of recent articles have stressed the limitations of p values and proposed alternative methods of analysis (e.g., Cumming, 2008, 2014; Halsey, Curran-Everett, Vowler, & Drummond, 2015; Johnson, 2013; Kruschke, 2010a, 2011; Nuzzo, 2014; Simonsohn, 2015b). In response, flagship journals such as *Psychological Science* have issued editorials warning against the uncritical and exclusive use of p values (Lindsay, 2015); similar warnings have been presented in the *Psychonomic Bulletin & Review* Statistical Guidelines for authors; finally, the journal *Basic And Applied Social Psychology* has banned p values altogether (Trafimow & Marks, 2015).

In order to reduce psychologists' dependence on p values it is essential to present alternatives that are concrete and practical. One such alternative is inference from confidence intervals (i.e., the "new statistics", Cumming, 2014; Grant, 1962). We see two main limitations for the new statistics. The first limitation is that confidence intervals are not Bayesian, which means that they forego the benefits that come with the Bayesian approach (a list of such benefits is provided below); moreover, confidence intervals share the fate of p values in the sense that they are prone to fallacies and misinterpretations (Greenland et al., in press; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). The second limitation is that confidence intervals presume that the effect under consideration exists; in other words, their use implies that every problem of inference is a problem of parameter estimation rather than hypothesis testing. Although we believe that effect size estimation is important and should receive attention, the question of size ("how big is the effect?") comes into play only after the question of presence ("is there an effect?") has been convincingly addressed (Morey, Rouder, Verhagen, & Wagenmakers, 2014). In his monograph "Theory of Probability", Bayesian pioneer Harold Jeffreys makes a sharp distinction between estimation and testing, discussing each in separate chapters: "In the problems of the last two chapters we were concerned with the estimation of the parameters in a law, the form of the law itself being given. We are now concerned with the more difficult question: in what circumstances do observations support a change of the form of the law itself? *This question is really logically prior to the estimation of the parameters, since the estimation problem presupposes that the parameters are relevant.*" (Jeffreys, 1961, p. 245; italics ours). The same sentiment was recently expressed by Simonsohn (2015b, p. 559): "Only once we are past asking whether a phenomenon exists at all and we come to accept it as qualitatively correct may we become concerned with estimating its magnitude more precisely. Before lines of inquiry arrive at the privileged position of having identified a phenomenon that is generally accepted as qualitatively correct, researchers require tools to help them distinguish between those that are and are not likely to get there." We believe it is a mistake to mandate either an estimation or a testing approach across the board; instead, the most productive mode of inference depends on the substantive questions that researchers wish to have answered. As illustrated below, the problems with p values are not a reason to abandon hypothesis testing – they are a reason to abandon p values.

As a concrete and practical alternative to hypothesis testing using p values, we propose to conduct hypothesis testing using Bayes factors (e.g., J. O. Berger, 2006; Jeffreys, 1935, 1961; Kass & Raftery, 1995). The Bayes factor hypothesis test compares the predictive adequacy of two competing statistical models, thereby grading the evidence provided by the data on

a continuous scale, and quantifying the change in belief that the data bring about for the two models under consideration. Bayes factors have many practical advantages; for instance, they allow researchers to quantify evidence, and they allow this evidence to be monitored continually, as data accumulate, and without needing to know the intention with which the data were collected (Rouder, 2014; Wagenmakers, 2007).

In order to profit from the practical advantages that Bayesian parameter estimation and Bayes factor hypothesis tests have to offer it is vital that the procedures of interest can be executed in accessible, user-friendly software package. In part II of this series (Wagenmakers et al., this volume) we introduce JASP (jasp-stats.org; JASP Team, 2016), a free and open-source program with a graphical user interface familiar to users of SPSS. With JASP, users are able to conduct classical analyses as well as Bayesian analyses, without having to engage in computer programming or mathematical derivation.

The overarching goal of Part I this series is to present Bayesian inference as an attractive alternative to p value NHST. To this end, a concrete example is used to highlight ten practical advantages of Bayesian parameter estimation and Bayesian hypothesis testing over their classical counterparts. Next we briefly address a series of ten objections against the Bayes factor hypothesis test. Our hope is that by raising awareness about Bayesian benefits (and by simultaneously providing a user-friendly software program, see Wagenmakers et al., this volume) we can help accelerate the adoption of Bayesian statistics in psychology and other disciplines.

Bayesian Inference and its Benefits

To facilitate the exposition below we focus on a concrete example: the height advantage of candidates for the US presidency (Stulp, Buunk, Verhulst, & Pollet, 2013). The data from the first 46 US presidential elections can be analyzed in multiple ways, but here we are concerned with the Pearson correlation ρ between the proportion of the popular vote and the height ratio (i.e., height of the president divided by the height of his closest competitor). Figure 2.1 shows that taller candidates tend to attract more votes; the sample correlation r equals .39 and is significantly different from zero ($p = .007$, two-sided test). A classical confidence interval for ρ ranges from .12 to .61. We now turn to a Bayesian analysis of these data, first discussing estimation, then discussing hypothesis testing of the correlation ρ . Our exposition is necessarily brief and selective; a complete treatment of Bayesian inference requires a monograph (e.g., Bernardo & Smith, 1994; Jeffreys, 1961; Jaynes, 2003; Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012; O'Hagan & Forster, 2004). In addition, we have made an effort to communicate the concepts and ideas without recourse to equations and derivations. Readers interested in the mathematical underpinnings of Bayesian inference are advised to turn to other sources (e.g., Ly, Verhagen, & Wagenmakers, 2016b; Marin & Robert, 2007; O'Hagan & Forster, 2004; Pratt, Raiffa, & Schlaifer, 1995; Rouder, Morey, Speckman, & Province, 2012; an overview and a reading list are provided in this issue, Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2016).

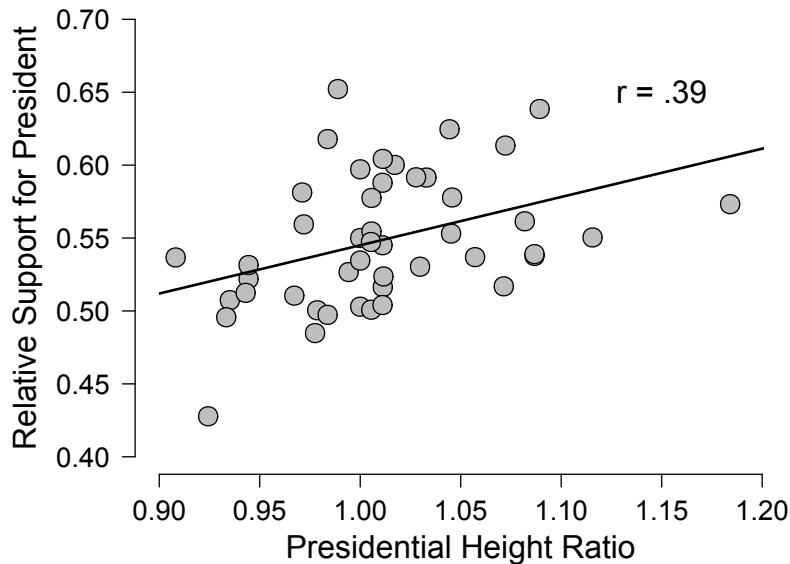


Figure 2.1: The proportion of the popular vote versus the height ratio between a US president and his closest competitor for the first 46 elections. Data obtained from Stulp et al. (2013). Figure based on JASP.

Bayesian Parameter Estimation

A Bayesian analysis may proceed as follows. The model under consideration assumes that the data are bivariate Normal, and interest centers on the unknown correlation coefficient ρ . In Bayesian statistics, the uncertainty about ρ before seeing the data is quantified by a probability distribution known as the prior. Here we specify a default prior distribution, one that stipulates that every value of ρ is equally plausible a priori (Jeffreys, 1961); this yields a uniform distribution ranging from -1 to 1 , shown in Figure 2.2 by the dotted line.² It is possible to specify different models by changing the prior distribution. For instance, later we will incorporate the knowledge that ρ is expected to be positive, which can be accomplished by using a uniform prior distribution that ranges only from 0 to 1 . For the moment, we refrain from doing so here because the classical NHST analysis is also two-sided.

Next the prior distribution is combined with the information from the data (i.e., the likelihood; A. W. F. Edwards, 1992; Myung, 2003; Royall, 1997) and the result is a posterior distribution. This posterior distribution quantifies the uncertainty about ρ after having seen the data. Figure 2.2 shows that compared to the prior distribution, the posterior

²The prior distributions for the other parameters from the bivariate Normal are inconsequential for inference about ρ and can be assigned vague prior distributions (Ly et al., 2016b). A slightly different and less transparent Bayesian model for the Pearson correlation coefficient is presented in Wetzels and Wagenmakers (2012).

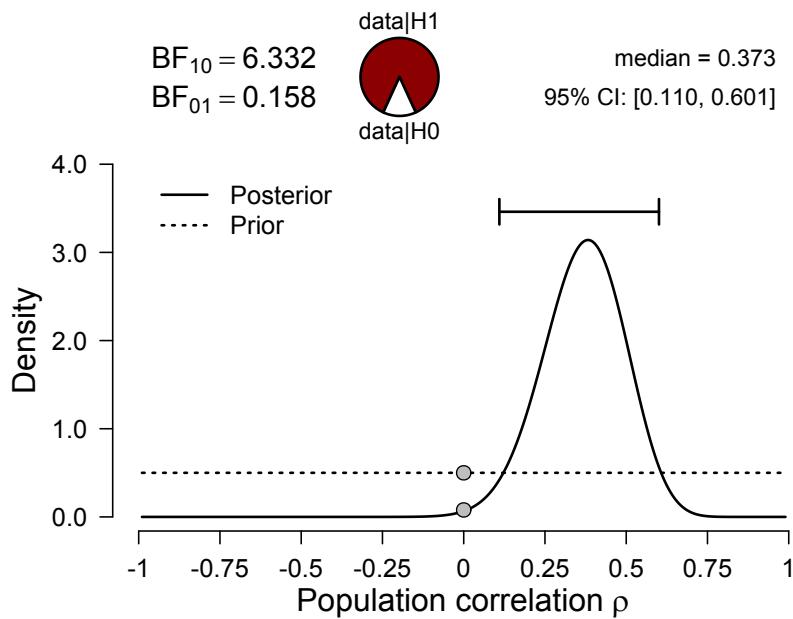


Figure 2.2: Prior and posterior distribution for the correlation between the proportion of the popular vote and the height ratio between a US president and his closest competitor. The default two-sided Bayes factor is visualized by the ratio between the prior and posterior ordinate at $\rho = 0$ and equals 6.33 in favor of the alternative hypothesis over the null hypothesis. Figure from JASP.

distribution assigns relatively little mass to values lower than 0 and higher than .70. A 95% credible interval ranges from .11 to .60, which means that one can be 95% confident that the true value of ρ lies between .11 and .60. When the posterior distribution is relatively peaked compared to the prior, this means that the data were informative and much has been learned. Note that the area under the prior and the posterior distribution has to equal 1; consequently, if some values of ρ are less likely under the posterior then they were under the prior, the reverse pattern needs to hold for at least some other values of ρ .

Benefits of Bayesian Parameter Estimation

In psychology, Bayesian parameter estimation techniques have recently been promoted by Jeff Rouder and colleagues (e.g., Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Lu, et al., 2007; Rouder, Lu, Morey, Sun, & Speckman, 2008), by Michael Lee and colleagues (e.g., Lee, 2008, 2011; Lee, Fuss, & Navarro, 2006), and by John Kruschke (e.g., Kruschke, 2010b, 2010a, 2011). Because the results of classical parameter estimation techniques (i.e., confidence intervals) are sometimes numerically similar to those obtained using Bayesian methods (i.e., credible intervals), it is tempting to conclude that the difference is not of practical interest. This is, however, a misconception. Below we indicate several arguments in favor of Bayesian parameter estimation using posterior distributions over classical parameter

estimation using confidence intervals. For more details and examples see Morey et al. (2016). Before proceeding, it is important to recall the definition of a classical confidence interval: An $X\%$ confidence interval for a parameter θ is an interval generated by a procedure that in repeated sampling has an $X\%$ probability of containing the true value of θ (Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Neyman, 1937). Thus, the confidence in the classical confidence interval resides in its performance in repeated use, across hypothetical replications. In contrast, the confidence in the Bayesian credible interval refers directly to the situation at hand (see benefit 3 below and see Wagenmakers, Morey, & Lee, 2016). Table 2.1 lists five benefits of Bayesian estimation over classical estimation. We will discuss each in turn.

Benefit 1. Bayesian estimation can incorporate prior knowledge

The posterior distribution is a compromise between the prior (i.e., what was known before the data arrived), and the likelihood (i.e., the extent to which the data update the prior). By selecting an appropriate prior distribution, researchers are able to insert substantive knowledge and add useful constraint (Vanpaemel, 2010; Vanpaemel & Lee, 2012). This is not a frivolous exercise that can be misused to obtain arbitrary results (Lindley, 2004). For instance, consider the estimation of IQ. Based on existing knowledge, it is advisable to use a Gaussian prior distribution with mean 100 and standard deviation 15. Another example concerns the estimation of a participant's latent ability to discriminate signal from noise in a psychophysical present-absent task. In the absence of ability, the participant still has a 50% probability of guessing the correct answer. Hence, the latent rate θ of correct judgements is bounded from below by 0.5 (Morey, Rouder, & Speckman, 2008; Rouder, Morey, Speckman, & Pratte, 2007). Any statistical paradigm that cannot incorporate such knowledge seems overly restrictive and incomplete. The founding fathers of classical inference –including “Student” and Fisher– mentioned explicitly that their methods apply only in the absence of any prior knowledge (Jeffreys, 1961, pp. 380-382).

To see how easy it is to add meaningful constraints to the prior distribution, consider again the example on the US presidents (see also Lee & Wagenmakers, 2013; Wagenmakers, Verhagen, & Ly, 2016). Assume that, before the data were examined, the correlation was believed to be positive; that is, it was thought that taller candidates attract more votes, not less. This restriction can be incorporated by assigning ρ a uniform distribution from 0 to 1 (Hoijtink, Klugkist, & Boelen, 2008; Hoijtink, 2011; Klugkist, Laudy, & Hoijtink, 2005). The results are shown in Figure 2.3. Note that the area under the one-sided prior distribution needs to equal 1, which explains why it is twice as high as the two-sided prior distribution shown in Figure 2.2.

A comparison between Figure 2.2 and Figure 2.3 also reveals that the restriction did not meaningfully alter the posterior distribution. This occurs because most of the posterior mass was already consistent with the restriction, and hence the one-sided restriction necessitated only a minor adjustment to the posterior obtained from the two-sided prior. In contrast, the classical one-sided 95% confidence interval ranges from .16 to 1, containing all values that would not be rejected by a one-sided $\alpha = .05$ significance test. This one-sided interval is very different from the two-sided interval that ranged from .12 to .61. In light of the data, and in light of the posterior distribution, the one-sided confidence interval does not appear

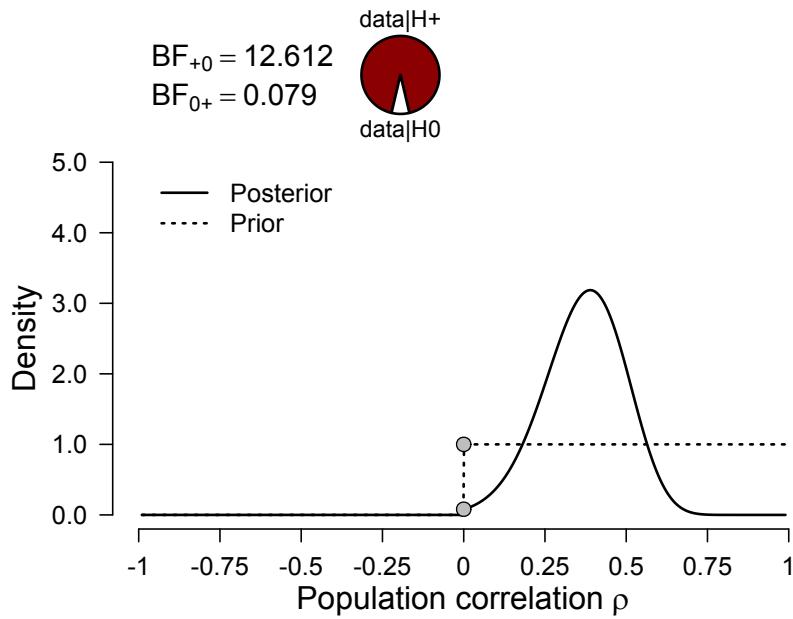


Figure 2.3: One-sided prior and posterior distribution for the correlation between the proportion of the popular vote and the height ratio between a US president and his closest competitor. The default one-sided Bayes factor is visualized by the ratio between the prior and posterior ordinate at $\rho = 0$ and equals 12.61 in favor of the alternative hypothesis over the null hypothesis. Figure from JASP.

to provide an intuitive or desirable summary of the uncertainty in estimating ρ .³ To further stress the difference between the Bayesian and classical one-sided intervals, note that for the present data the one-sided classical interval that presumes the opposite restriction (i.e., taller candidates are assumed to attract fewer votes) yields an interval that ranges from -1 to 0.58 , that is, covering all of the negative range and most of the positive range. In sharp contrast, the restriction to negative correlations yields a Bayesian one-sided credible interval with negative bounds that are very close to zero, as one would expect.

In sum, Bayesian estimation methods allow researchers to add substantive prior knowledge. The classical framework is incapable of doing so except for the simplest case of an order-restriction, where it yields intervals that do not provide useful information about the precision with which parameters were estimated.

³The rationale behind the one-sided classical confidence interval is difficult to teach. One statistics teacher remarked “one-sided classical confidence intervals really blow students’ minds, and not in a good way.” Another statistics teacher said that she simply refuses to cover the concept at all, in order to prevent student riots.

Benefit 2. Bayesian estimation can quantify confidence that θ lies in a specific interval

The posterior distribution for a parameter θ provides a complete summary of what we know about this parameter. Using this posterior distribution, we can answer questions such as “how much more likely is the value $\theta = .6$ versus the value $\theta = .4?$ ” – this equals the ratio of the heights of the posterior distribution at those values. Also, we can use the posterior distribution to quantify how likely it is that θ falls in a specific interval, say, between .2 and .4 – this equals the posterior mass in that interval (Wagenmakers, Morey, & Lee, 2016).

In contrast, the classical confidence interval procedure can do no more than provide X% confidence intervals. It is not possible within the classical framework to specify the interval bounds and then ask for the probability or confidence that the true value is within these bounds. This is a serious limitation. For instance, one criterion for the diagnosis of an intellectual disability is an IQ below 70. Hence it may be important to know the probability that a person’s IQ is in the interval from 0 to 70, given a series of test scores. With classical statistics, this question cannot be addressed. Pratt et al. (1995, p. 258) formulate this concern as follows:

“A feature of confidence regions which is particularly disturbing is the fact that the confidence level must be selected in advance and the region we then look at is imposed by chance and may not be at all one we are interested in. Imagine the plight of a manager who exclaims, ‘I understand [does he?] the meaning that the demand for XYZ will lie in the interval 973 to 1374 with confidence .90. However, I am particularly interested in the interval 1300 to 1500. What confidence can I place on that interval?’ Unfortunately, this question *cannot* be answered. Of course, however, it is possible to give a posterior probability to that particular interval—or any other—based on the sample data and on a codification of the manager’s prior judgments.”

Cox (1958, p. 363) expresses a similar concern (see also Lindley, 1965, p. 23):

“(...) the method of confidence intervals, as usually formulated, gives only one interval at some preselected level of probability. (...) For when we write down the confidence interval (...) for a completely unknown normal mean, there is certainly a sense in which the unknown mean θ is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if θ does lie outside the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts.”

Benefit 3. Bayesian estimation conditions on what is known (i.e., the data)

The Bayesian credible interval (and Bayesian inference in general) conditions on all that is known. This means that inference is based on the specific data set under consideration, and that performance of the methodology for other hypothetical data sets is irrelevant. In contrast, the classical confidence interval is based on average performance across hypothetical data sets. To appreciate the difference, consider a scale that works perfectly in 95% of the

cases, but returns a value of “1 kilo” in the remaining 5%. Suppose you weigh yourself on this scale and the result is “70 kg”. Classically, your confidence in this value should be 95%, because the scale is accurate in 95% of all cases. However, the data tell you that the scale has not malfunctioned, and hence you can be 100% confident in the result. Similarly, suppose the scale returns “1 kilo”. Classically, you can have 95% confidence in this result. Logically, however, the value of “1 kilo” tells you that the scale has malfunctioned, and you have learned nothing at all about your weight (J. O. Berger & Wolpert, 1988).

Another example is the 50% confidence interval for a binomial rate parameter θ (i.e., θ is allowed to take on values between 0 and 1). A classically valid 50% interval can be constructed by ignoring the data and randomly reporting either the interval $(0 - 0.5)$ or $(0.5 - 1)$. This random interval procedure will cover the true value in 50% of the cases. Of course, when the data are composed of 10 successes out of 10 trials the interval $(0 - 0.5)$ is nonsensical; however, the confidence of the classical procedure is based on average performance, and the average performance of the random interval is 50%.

Thus, one of the crucial differences between classical and Bayesian procedures is that classical procedures are generally “pre-data”, whereas Bayesian procedures are “post-data” (Jaynes, 2003).⁴ One final example, taken from by J. O. Berger and Wolpert (1988), should suffice to make the distinction clear. The situation is visualized in Figure 2.4: two balls are dropped, one by one, in the central tube located at θ . Each ball travels down the central tube until it arrives at the T-junction, where it takes either the left or the right tube with equal probability, where the final outcome is registered as $\theta - 1$ and $\theta + 1$, respectively.

Consider that the first ball registers as “12”. Now there are two scenarios, both equally likely a priori, that provide radically different information. In the first scenario, the second ball lands in the other tube. For instance, the second ball can register as a “14”. In this case, we know with 100% certainty that θ is 13 – the middle value. In the second scenario, the second ball lands in the same tube as the first one, registering another “12”. This datum is wholly uninformative, as we still do not know whether θ equals 13 (when “12” is the left tube) or 11 (when “12” is the right tube). Hence we simply guess that the balls have traveled down the left tube and state that θ equals 13. The first scenario always yields 100% accuracy and the second scenario yields 50% accuracy. Both scenarios are equally likely to occur and hence the overall probability that the above procedure correctly infers the true value of θ is 75%. This indicates how well the procedure performs in repeated use, averaged across the sample space (i.e., all possible data sets).

However, consider that two balls have been observed and you are asked what you have learned about θ . Even classical statisticians agree that in cases such as these, one should not report an unconditional confidence of 75%; instead, one should take into account that the first scenario is different from the second, and draw different conclusions depending on the data at hand. As a technical side note, the negative consequences of averaging across hypothetical data sets that are fundamentally different is known as the problem of “recognizable/relevant subsets”. Ultimately, the problem can only be overcome by conditioning on the data that were observed, but doing so removes the conceptual basis of classical inference. In Bayesian

⁴This difference was already clear to Laplace, who argued that the post-data viewpoint is “obviously” the one that should be employed (Gillispie, 1997, p. 82).

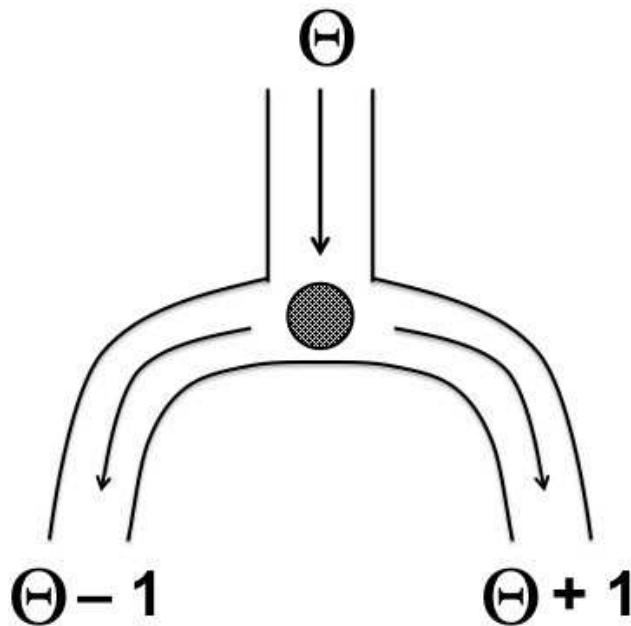


Figure 2.4: Two balls are dropped consecutively in a tube at location θ ; each ball lands randomly at tube location $\theta - 1$ or $\theta + 1$. When the two balls land in different locations, θ is known with 100% certainty; when the two balls land in the same location, θ is known with 50% certainty. The pre-data average of 75% confidence is meaningless after the data have been observed. The example is taken from J. O. Berger and Wolpert (1988).

inference, the problem of relevant subsets does not occur (for a more detailed discussion see e.g., Brown, 1967; Cornfield, 1969; Gleser, 2002; Morey et al., 2016; Pierce, 1973; Pratt, 1961). Relevant subsets are easy to detect in somewhat contrived examples such as the above; however, they also exist in standard inference situations such as the comparison of two means (Buehler & Fedderson, 1963).

The conceptual and practical difference between classical and Bayesian intervals is eloquently summarized by Jaynes (1976, pp. 200-201):

“Our job is not to follow blindly a rule which would prove correct 90% of the time in the long run; there are an infinite number of radically different rules, all with this property. Our job is to draw the conclusions that are most likely to be right in the specific case at hand (...) To put it differently, the sampling distribution of an estimator is not a measure of its reliability in the individual case, because considerations about samples that have not been observed, are simply not relevant to the problem of how we should reason from the one that has been observed. A doctor trying to diagnose the cause of Mr. Smith’s stomachache would not be helped by statistics about the number of patients who complain instead of a sore arm or stiff neck. This does not mean that there are no connections at all between individual case and long-run performance; for if we have found the procedure which is ‘best’ in each individual case, it is hard to see how

it could fail to be ‘best’ also in the long run (...) The point is that the converse does not hold; having found a rule whose long-run performance is proved to be as good as can be obtained, it does not follow that this rule is necessarily the best in any particular individual case. One can trade off increased reliability for one class of samples against decreased reliability or another, in a way that has no effect on long-run performance; but has a very large effect on performance in the individual case.”

Benefit 4. Bayesian estimation is coherent (i.e., not internally inconsistent)

One of the defining characteristics of Bayesian inference is that it is coherent, meaning that all inferential statements must be mutually consistent; in other words, Bayesian inference does not depend on the way a problem is framed (de Finetti, 1974; Lindley, 1985, 2006; Ramsey, 1926). In Bayesian statistics, coherence is guaranteed by the laws of probability theory: “Coherence acts like geometry in the measurement of distance; it forces several measurements to obey the system.” (Lindley, 2000, p. 306). For instance, when we know that for a posterior distribution, $p(0 < \rho < 0.3) = a$ and $p(0.3 < \rho < 0.4) = b$, then it has to follow that $p(0 < \rho < 0.4) = a + b$. Any other conclusion violates the laws of probability theory and is termed incoherent or absurd (Lindley, 1985). A famous example of incoherence is provided by Tversky and Kahneman (1983, p. 297), who gave participants the following background story:

“Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.”

After reading the story, participants were asked to provide the probability of several statements, including the following two:

1. “Linda is a bank teller. (T)”
2. “Linda is a bank teller and is active in the feminist movement. (T&F)”

The results showed that the great majority of participants judged the conjunction statement T&F to be more probable than the constituent statement T. This conjunction error violates the laws of probability theory, according to which the probability of T&F can never be higher than the probability of either of its constituents (see also Nilsson, Winman, Juslin, & Hansson, 2009). Within the restrictions of the normative Bayesian framework, violations of logic and common sense can never occur.

Coherence is about fitting together different pieces of information in a way that is internally consistent, and this can be done in only one way: by obeying the laws of probability theory. Consider the following example. A bent coin is tossed twice: the first toss comes up heads, and the second toss comes up tails. Assume that, conditional on the angle of the bent coin, the tosses are independent. Then the final inference about the angle should not depend on the order with the data were observed (i.e., heads-tails or tails-heads). Similarly, the final inference should not depend on whether the data were analyzed sequentially, one

at a time, or as a single batch. This sequential form of coherence can only be obtained by continual updating of the prior distribution, such that the posterior distribution after datum i becomes the prior distribution for the analysis of datum $i + 1$; without a prior distribution, coherence is impossible and inferential statements are said to be absurd. Coherence also ensures that Bayesian inference is equally valid for all sample sizes – there is no need for “rules of thumb” to identify sample sizes below which inference cannot be trusted.

Coherence has been argued to be the core element of Bayesian inference; for instance, Ramsey (1926) argued that “the most generally accepted parts of logic, namely, formal logic, mathematics and the calculus of probabilities, are all concerned simply to ensure that our beliefs are not self-contradictory” (see Eagle (Ed.), 2011, p. 65); Jeffreys (1961, p. ix) starts the preface to the Bayesian classic “Theory of Probability” by stating that “The chief object of this work is to provide a method of drawing inferences from observational data that will be self-consistent and can also be used in practice”. Moreover, Lindley (1985) used the term “coherent statistics” instead of “Bayesian statistics”, and Joyce (1998) highlighted the importance of coherence by proving that “any system of degrees of belief that violates the axioms of probability can be replaced by an alternative system that obeys the axioms and yet is more accurate *in every possible world*” (see Eagle (Ed.), 2011, p. 89).

In contrast to Bayesian inference, the concept of coherence plays no role in the classical framework. The resulting problems become manifest when different sources of information need to be combined. In the classical framework, the usual remedy against incoherence is to focus on one source of information only. Even though this hides the problem from view, it does not eliminate it, because almost any data set can be divided into arbitrary batches, and the final inference should not depend on the order or method of division.

Benefit 5. Bayesian estimation extends naturally to complicated models

The principles of Bayesian estimation hold for simple models just as they do for complicated models (e.g., Gelman & Hill, 2007; Gelman et al., 2014). Regardless of model complexity, Bayesian inference features only one estimator: the posterior distribution. When this posterior distribution cannot be obtained analytically, it is usually possible to draw samples from it using numerical algorithms such as Markov chain Monte Carlo (MCMC; Gelfand & Smith, 1990; Gilks, Richardson, & Spiegelhalter, 1996; van Ravenzwaaij, Cassey, & Brown, in press). By increasing the number of MCMC samples, the posterior distribution can be approximated to arbitrary precision. With the help of MCMC sampling, Bayesian inference proceeds almost mechanically, allowing for straightforward inference even in relatively complex models (e.g., Lunn et al., 2012).

Consider the use of hierarchical nonlinear process models in cognitive psychology. Most models in cognitive psychology are *nonlinear* in that they are more than the sum of effects plus noise. An example of a nonlinear model is Yonelinas’ dual process model, in which memory performance is a mixture of recollection, modeled as a discrete all-or-none process, and familiarity, modeled as a continuous signal-detection process (e.g., Yonelinas, 2002). In realistic settings each of several people observe each of several items, but each person-item combination is unique. It is reasonable to assume variation across people and items, and once the model is expanded to include people and item effects, it is not only nonlinear, but quite

numerous in parameters. One approach is to aggregate data across people, items, or both. The drawback is that the fit to aggregated data will be substantially distorted and perhaps reflect the psychological processing of nobody (Estes, 1956; Heathcote, Brown, & Mewhort, 2000; Rouder et al., 2005). A superior approach is to construct hierarchical nonlinear process models that simultaneously account for psychological process and nuisance variation from people and items. Pratte and Rouder (2012), for example, fit an expanded, hierarchical dual process model with about 2000 parameters. It is not obvious to us how to fit such models in a classical framework.⁵ Fortunately, the analysis is tractable and relatively straightforward using Bayesian inference with MCMC sampling.

Thus, Bayesian estimation is ideally suited for models that respect the complexity inherent in psychological data; such realistic models can be hierarchical, involve mixtures, contain nonlinearities, or be based on detailed considerations of the underlying psychological process (Lee & Wagenmakers, 2013; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Despite their surface differences, all such models obey the same conceptual principles, and parameter estimation is merely a matter of “turning the Bayesian handle”:

“What is the principal distinction between Bayesian and classical statistics?

It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria.

I have heard people who should know better use this dullness as an argument against Bayesianism. One might as well complain that Newton’s dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy’s epicyclic system.” (Dawid, 2000, p. 326)

Bayesian Hypothesis Testing

In Bayesian parameter estimation, the inferential end-goal is the posterior distribution. In the earlier example featuring election outcomes, the posterior distribution for ρ allowed an answer to the question “What do we know about the correlation between height and popularity in the US elections, assuming from the outset that such a correlation exists?” From this formulation, it is clear that we cannot use the posterior distribution alone for the purpose of hypothesis testing: the prior formulation $\rho \sim \text{Uniform}[-1, 1]$ presupposes that ρ is relevant, that is, it presupposes that ρ is unequal to zero.⁶ To test an invariance or a general law, this law needs to be assigned a separate prior probability (Etz & Wagenmakers, 2016; Haldane, 1932; Jeffreys, 1961, 1973, 1980; Ly et al., 2016b; Wrinch & Jeffreys, 1921, 1923): to test $\mathcal{H}_0 : \rho = 0$, this hypothesis needs to be taken serious a priori. In the election example, this means that we should explicitly consider the hypothesis that taller candidates do not attract a larger or smaller proportion of the popular vote. This is something that the estimation framework fails to do. Consequently, as stated by J. O. Berger (2006, p.

⁵Using maximum likelihood estimation, general-purpose gradient decent algorithms in Matlab, R, and Excel often fail in nonlinear contexts with more than just a few dozen parameters.

⁶Under a continuous prior probability distribution, the probability assigned to any single point (i.e., $\rho = 0$) is zero.

	Bayesian Inference	Classical Inference	References
Desiderata for Parameter Estimation			
To incorporate prior knowledge	✓	✗	1,2
To quantify confidence that θ lies in a specific interval	✓	✗	3
To condition on what is known (i.e., the data)	✓	✗	4,5
To be coherent (i.e., not internally inconsistent)	✓	✗	6,7
To extend naturally to complicated models	✓	✗	8,9
Desiderata for Hypothesis Testing			
To quantify evidence that the data provide for \mathcal{H}_0 vs. \mathcal{H}_1	✓	✗	10,11
To quantify evidence in favor of \mathcal{H}_0	✓	✗	12,13
To allow evidence to be monitored as data accumulate	✓	✗	14,15
To not depend on unknown or absent sampling plans	✓	✗	16,17
To not be “violently biased” against \mathcal{H}_0	✓	✗	18,19,20

Table 2.1: Select overview of advantages of Bayesian inference over classical inference. See text for details. References: 1 = Dienes (2011); 2 = Vanpaemel (2010); 3 = Pratt et al. (1995, p. 258); 4 = J. O. Berger and Wolpert (1988); 5 = Jaynes (2003); 6 = Lindley (1985); 7 = Lindley (2000); 8 = Pratte and Rouder (2012); 9 = Lunn et al. (2012); 10 = Jeffreys (1935); 11 = Jeffreys (1961); 12 = Rouder et al. (2009); 13 = Wagenmakers (2007); 14 = W. Edwards et al. (1963); 15 = Rouder (2014); 16 = J. O. Berger and Berry (1988); 17 = Lindley (1993); 18 = W. Edwards (1965); 19 = J. O. Berger and Delampady (1987); 20 = Sellke et al. (2001).

383): “[...] 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).”

Hence, when the goal is hypothesis testing, Bayesians need to go beyond the posterior distribution. To answer the question “To what extent do the data support the presence of a correlation?” one needs to compare two models: a null hypothesis that states the absence of the effect (i.e., $\mathcal{H}_0 : \rho = 0$) and an alternative hypothesis that states its presence. In Bayesian statistics, this alternative hypothesis needs to be specified exactly. In our election scenario, the alternative hypothesis we discuss first is specified as $\mathcal{H}_1 : \rho \sim \text{Uniform}(-1, 1)$, that is, every value of ρ is judged to be equally likely a priori (Jeffreys, 1961; Ly et al., 2016b).⁷

With the competing hypotheses \mathcal{H}_0 and \mathcal{H}_1 fully specified, the process of updating their

⁷Specification of prior distributions is an important component for Bayes factor hypothesis testing, as the prior distributions define a model’s complexity and hence exert a lasting effect on the test outcome. We will return to this issue later.

relative plausibilities is described by a simplification of Bayes' rule:

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

In this equation, the prior model odds $p(\mathcal{H}_1)/p(\mathcal{H}_0)$ indicate the relative plausibility of the two models before seeing the data. After observing the data, the relative plausibility is quantified by the posterior model odds, that is, $p(\mathcal{H}_1 \mid \text{data})/p(\mathcal{H}_0 \mid \text{data})$. The change from prior to posterior odds brought about by the data is referred to as the Bayes factor, that is, $p(\text{data} \mid \mathcal{H}_1)/p(\text{data} \mid \mathcal{H}_0)$. Because of the subjective nature of the prior model odds, the emphasis of Bayesian hypothesis testing is on the amount by which the data shift one's beliefs, that is, on the Bayes factor. When the Bayes factor BF_{10} equals 6.33, the data are 6.33 times more likely under \mathcal{H}_1 than under \mathcal{H}_0 . When the Bayes factor equals $\text{BF}_{10} = 0.2$, the data are 5 times more likely under \mathcal{H}_0 than under \mathcal{H}_1 . Note that the subscripts "10" in BF_{10} indicate that \mathcal{H}_1 is in the numerator of Equation 2.1 and \mathcal{H}_0 is in the denominator, whereas the subscripts "01" indicate the reverse. Hence, $\text{BF}_{10} = 1/\text{BF}_{01}$.

An alternative interpretation of the Bayes factor is in terms of the models' relative predictive performance (Wagenmakers, Grünwald, & Steyvers, 2006; Wagenmakers, Morey, & Lee, 2016). Consider two models, \mathcal{H}_0 and \mathcal{H}_1 , and two observations, $y = (y_1, y_2)$. The Bayes factor $\text{BF}_{10}(y)$ is given by $p(y_1, y_2 \mid \mathcal{H}_1)/p(y_1, y_2 \mid \mathcal{H}_0)$, that is, the ratio of the advance probability that the competing models assign to the data. Thus, both models make a probabilistic prediction about the data, and the model with the best prediction is preferred. This predictive interpretation can also be given a sequential slant. To see this, recall that according to the definition of conditional probability, $p(y_1, y_2) = p(y_1)p(y_2 \mid y_1)$. In the current example, both \mathcal{H}_0 and \mathcal{H}_1 make a prediction about the first data point, yielding $\text{BF}_{10}(y_1) = p(y_1 \mid \mathcal{H}_1)/p(y_1 \mid \mathcal{H}_0)$ – the relative predictive performance for the first data point. Next, both models incorporate the knowledge gained from the first data point and make a prediction for the second observation, yielding $\text{BF}_{10}(y_2 \mid y_1) = p(y_2 \mid y_1, \mathcal{H}_1)/p(y_2 \mid y_1, \mathcal{H}_0)$ – the relative predictive performance for the second data point, given the knowledge obtained from the first. These one-step-ahead sequential forecasts can be combined –using the law of conditional probability– to produce a model's overall predictive performance (cf. Dawid's prequential principle; e.g., Dawid, 1984): $\text{BF}_{10}(y) = \text{BF}_{10}(y_1) \times \text{BF}_{10}(y_2 \mid y_1)$. The accumulation of one-step-ahead sequential forecasts provides a fair assessment of a model's predictive adequacy, penalizing undue model complexity and thereby implementing a form of Occam's razor⁸ (i.e., the principle of parsimony, Jeffreys & Berger, 1992; Lee & Wagenmakers, 2013; Myung & Pitt, 1997; Myung, Forster, & Browne, 2000; Vandekerckhove, Matzke, & Wagenmakers, 2015; Wagenmakers & Waldorp, 2006). The predictive interpretation of the Bayes factor is conceptually relevant because it means that inference can be meaningful even without either of the models being true in some absolute sense (Morey, Romeijn, & Rouder, 2013; but see van Erven, Grünwald, & de Rooij, 2012).

⁸An overly complex model mistakes noise for signal, tailoring its parameters to data patterns that are idiosyncratic and nonrepeatable. This predilection to "overfit" is exposed when the model is forced to make out-of-sample predictions, because such predictions will be based partly on noise.

From the Bayesian perspective, evidence is an inherently relative concept. Therefore it makes little sense to try and evaluate evidence for a specific hypothesis without having specified exactly what the alternative hypothesis predicts. In the words of Peirce (1878a), “When we adopt a certain hypothesis, it is not alone because it will explain the observed facts, but also because the contrary hypothesis would probably lead to results contrary to those observed.” (as quoted in Hartshorne & Weiss, 1932, p. 377). As outlined below, this is one of the main differences with classical hypothesis testing, where the p value quantifies the unusualness of the data under the null hypothesis (i.e., the probability of obtaining data at least as extreme as those observed, given that the null hypothesis is true), leaving open the possibility that the data are even more likely under a well-specified and plausible alternative hypothesis.

In sum, Bayes factors compare the predictive adequacy of two competing statistical models. By doing so, they grade the evidence provided by the data on a continuous scale, and quantify the change in belief that the data bring about for the two models under consideration. Its long history and direct link to Bayes’ rule make the Bayes factor “the standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis & Raftery, 1997, p. 648) and “the primary tool used in Bayesian inference for hypothesis testing and model selection” (J. O. Berger, 2006, p. 378). We consider the Bayes factor (or its logarithm) a *thermometer for the intensity of the evidence* (Peirce, 1878b). In our opinion, such a thermometer is exactly what researchers desire when they wish to measure the extent to which their observed data support \mathcal{H}_1 or \mathcal{H}_0 .

Benefits of Bayesian Hypothesis Testing

In psychology, several researchers have recently proposed, developed, and promoted Bayes factor hypothesis testing (e.g., Dienes, 2008, 2011, 2014; Hoijtink, 2011; Klugkist et al., 2005; Masson, 2011; Morey & Rouder, 2011; Mulder et al., 2009; Rouder et al., 2009, 2012; Vanpaemel, 2010; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Table 2.1 provides a non-exhaustive list of five specific benefits of Bayesian hypothesis testing over classical p value hypothesis testing (see also Kass & Raftery, 1995, p. 773). We now briefly discuss each of these benefits in turn. Other benefits of Bayesian hypothesis testing include those already mentioned for Bayesian parameter estimation above.

Benefit 1. The Bayes factor quantifies evidence that the data provide for \mathcal{H}_0 vs. \mathcal{H}_1

As mentioned above, the Bayes factor is inherently comparative: it weighs the support for one model against that of another. This contrasts with the p value, which is calculated conditional on the null hypothesis \mathcal{H}_0 being true; the alternative hypothesis \mathcal{H}_1 is left unspecified and hence its predictions are irrelevant as far as the calculation of the p value is concerned. Consequently, data that are unlikely under \mathcal{H}_0 may lead to its rejection, even though these data are just as unlikely under \mathcal{H}_1 – and are therefore perfectly uninformative (Wagenmakers et al., in press). Figure 2.5 provides a cartoon highlighting that p value NHST considers one side of the coin.



Figure 2.5: A boxing analogy of the p value (Wagenmakers et al., in press). The referee uses null hypothesis significance testing and therefore considers only the deplorable state of boxer \mathcal{H}_0 (i.e., the null hypothesis). His decision to reject \mathcal{H}_0 puzzles the public. Figure available at <http://www.flickr.com/photos/23868780@N00/12559689854/>, courtesy of Dirk-Jan Hoek, under CC license <https://creativecommons.org/licenses/by/2.0/>.

The practical relevance of this concern was underscored by the infamous court case of Sally Clark (Dawid, 2005; Hill, 2005; Nobles & Schiff, 2005). Both of Sally Clark's children had died at an early age, presumably from cot death or SIDS (sudden infant death syndrome). The probability of a mother having to face such a double tragedy was estimated to be 1 in 73 million. Such a small probability may have influenced judge and jury, who in November 1999 decided to sentence Sally Clark to jail for murdering her two children. In an open letter published in 2002, the president of the Royal Statistical Society Peter Green explained why the probability of 1 in 73 million is meaningless: "The jury needs to weigh up two competing explanations for the babies' deaths: SIDS or murder. The fact that two deaths by SIDS is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation." (Nobles & Schiff, 2005, p. 19). This point of critique is not just relevant for the case of Sally Clark, but applies to all inferences based on the p value.

Bayes factors compare two competing models or hypotheses: \mathcal{H}_0 and \mathcal{H}_1 . Moreover, Bayes factors do so by fully conditioning on the observed data y . In contrast, the p value is a tail-area integral that depends on hypothetical outcomes more extreme than the one observed in the sample at hand. Such a practice violates the likelihood principle and results in paradoxical conclusions (for examples see J. O. Berger & Wolpert, 1988; Wagenmakers, 2007). Indeed, our personal experience suggests that this is one of the most widespread misconceptions that practitioners have about p values: interpreting a p value as the "probability of obtaining these results given that the null hypothesis is true". However, as mentioned

above, the p value equals the probability of obtaining results *at least as extreme* as those observed given that the null hypothesis is true. As remarked by Jeffreys (1980, p. 453): “I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.” Towards the end of his life, this critique was acknowledged by one of the main protagonists of the p value, Ronald Fisher himself.⁹ In discussing inference for a binomial rate parameter based on observing 3 successes out of 14 trials, Fisher argued for the use of likelihood, implicitly acknowledging Jeffreys’ concern:

“Objection has sometimes been made that the method of calculating Confidence Limits by setting an assigned value such as 1% on the frequency of observing 3 or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly the same manner as the value 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation.” (Fisher, 1959, p. 68).

Benefit 2. The Bayes factor can quantify evidence in favor of \mathcal{H}_0

It is evident from Equation 2.1 that the Bayes factor is able to quantify evidence in favor of \mathcal{H}_0 . In the Bayesian framework, no special status is attached to either of the hypotheses under test; after the models have been specified exactly, the Bayes factor mechanically assesses each model’s one-step-ahead predictive performance, and expresses a preference for the model that was able to make the most accurate series of sequential forecasts (Wagenmakers et al., 2006). When the null hypothesis \mathcal{H}_0 predicts the observed data better than the alternative hypothesis \mathcal{H}_1 , this signifies that the additional complexity of \mathcal{H}_1 is not warranted by the data.

The fact that the Bayes factor can quantify evidence in favor of the null hypothesis can be of considerable substantive importance (e.g., Gallistel, 2009; Rouder et al., 2009). For instance, the hypothesis of interest may predict an invariance, that is, the absence of an effect across a varying set of conditions. The ability to quantify evidence in favor of the null hypothesis is also important for replication research, and should be of interest to any researcher who wishes to learn whether the observed data provide evidence of absence or absence of evidence (Dienes, 2014). Specifically, the possible outcomes of the Bayes factor can be assigned to three discrete categories: (1) evidence in favor of \mathcal{H}_1 (i.e., evidence in favor of the presence of an effect); (2) evidence in favor of \mathcal{H}_0 (i.e., evidence in favor of the absence of an effect); (3) evidence that favors neither \mathcal{H}_1 nor \mathcal{H}_0 . An example of evidence for absence is $BF_{01} = 15$, where the observed data are 15 times more likely to occur under \mathcal{H}_0 than under \mathcal{H}_1 . An example of absence of evidence is $BF_{01} = 1.5$, where the observed data are only 1.5 times more likely to occur under \mathcal{H}_0 than under \mathcal{H}_1 . Evidentially these scenarios are very different, and it is clearly useful and informative to discriminate between

⁹The first p value was calculated by Pierre-Simon Laplace in the 1770s; the concept was formally introduced by Karl Pearson in 1900 as a central component to his Chi-squared test (<http://en.wikipedia.org/wiki/P-value#History>).

the two. However, the p value is not able to make the distinction, and in either of the above scenarios one may obtain $p = .20$. In general, the standard p value NHST is unable to provide a measure of evidence in favor of the null hypothesis.

Benefit 3. The Bayes factor allows evidence to be monitored as data accumulate

The Bayes factor can be thought of as a thermometer for the intensity of the evidence. This thermometer can be read out, interpreted, and acted on at any point during data collection (cf. the stopping rule principle; J. O. Berger & Wolpert, 1988). Using Bayes factors, researchers are free to monitor the evidence as the data come in, and terminate data collection whenever they like, such as when the evidence is deemed sufficiently compelling, or when the researcher has run out of resources (e.g., J. O. Berger, 1985, Chapter 7; W. Edwards et al., 1963; Rouder, 2014; Wagenmakers, 2007). This freedom has substantial practical ramifications, and allows experiments to be conducted in a manner that is both efficient and ethical (e.g., Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, in press).

Consider the hypothetical case where a memory researcher, professor Bumbledorf, has planned to test 40 children with severe epilepsy using intracranial EEG. In scenario 1, Bumbledorf tests 20 children and finds that the data are so compelling that the conclusion hits her straight between the eyes (i.e., Berkson's interocular traumatic test, W. Edwards et al., 1963, p. 217). Should Bumbledorf feel forced to test 20 children more, inconveniencing the patients and wasting resources that could be put to better use? In scenario 2, Bumbledorf tests all 40 children and feels that, although the data show a promising trend, the results are not statistically significant ($p = .11$). Should Bumbledorf be disallowed from testing additional children, thereby possibly preventing the patients' earlier efforts from advancing science by contributing to data that yield an unambiguous conclusion? With Bayes factors, there are no such conundrums (J. O. Berger & Mortera, 1999); in scenario 1, Bumbledorf can stop after 20 patients and report the Bayes factor; in scenario 2, Bumbledorf is free to continue testing until the results are sufficiently compelling. This freedom stands in sharp contrast to the standard practice of p value NHST, where adherence to the sampling plan is critical; this means that according to standard p value NHST dogma, Bumbledorf is forced to test the remaining 20 patients in scenario 1 ("why did you even look at the data after 20 patients?"), and Bumbledorf is prohibited from testing addition patients in scenario 2 ("maybe you should have planned for more power").

It should be acknowledged that the standard framework of p value NHST can be adjusted so that it can accommodate sequential testing, either in a continual fashion, with an undetermined number of tests (e.g., Botella, Ximénez, Revuelta, & Suero, 2006; Fitts, 2010; Frick, 1998; Wald & Wolfowitz, 1948) or in an interrupted fashion, with a predetermined number of tests (e.g., Lakens & Evers, 2014). From a Bayesian perspective, however, corrections for sequential monitoring are an anathema. Anscombe (1963, p. 381) summarized the conceptual point of contention:

“ ‘Sequential analysis’ is a hoax(...) So long as all observations are fairly reported, the sequential stopping rule that may or may not have been followed is irrelevant. The experimenter should feel entirely uninhibited about continuing or

discontinuing his trial, changing his mind about the stopping rule in the middle, etc., because the interpretation of the observations will be based on what was observed, and not on what might have been observed but wasn't."

Benefit 4. The Bayes factor does not depend on unknown or absent sampling plans

The Bayes factor is not affected by the sampling plan, that is, the intention with which the data were collected. This sampling-plan-irrelevance follows from the likelihood principle (J. O. Berger & Wolpert, 1988), and it means that Bayes factors may be computed and interpreted even when the intention with which the data are collected is ambiguous, unknown, or absent. This is particularly relevant when the data at hand are obtained from a natural process, and the concepts of "sampling plan" and "experiment" do not apply.

As a concrete demonstration of the practical problems of p values when the sampling plan is undefined, consider again the election example and the data shown in Figure 2.1. We reported that for this correlation, $p = .007$. However, this p value was computed under a fixed sample size scenario; that is, the p value was computed under the assumption that an experimenter set out to run 46 elections and then stop. This sampling plan is absurd and by extension, so is the p value. But what is the correct sampling plan? It could be something like "US elections will continue every four years until democracy is replaced with a different system of government or the US ceases to exist". But even this sampling plan is vague – we only learn that we can expect quite a few elections more.

In order to compute a p value, one could settle for the fixed sample size scenario and simply not worry about the details of the sampling plan. However, consider the fact that new elections will continue be added to the set. How should such future data be analyzed? One can pretend, after every new election, that the sample size was fixed. However, this myopic perspective induces a multiple comparison problem – every new test has an additional non-zero probability of falsely rejecting the null hypothesis, and the myopic perspective therefore fails to control the overall Type I error rate.¹⁰

In contrast to p value NHST, the Bayes factor can be meaningfully interpreted even when the data at hand have been generated by real-world processes outside of experimental control. Figure 2.6 shows how the data from the US elections can be analyzed as they come in over time, an updating process that can be extended continually and indefinitely, as long as the US electoral process exists. This example also emphasizes the intimate connection between the benefit of monitoring the evidence as it unfolds over time, and the benefit of being able to compute the evidence from data outside of experimental control: both benefits occur because the Bayes factor does not depend on the intention with which the data are collected (i.e., hypothetical data sets that are not observed).

¹⁰For sequential tests the multiple comparisons are not independent; this reduces but does not eliminate the rate with which the Type I error increases.

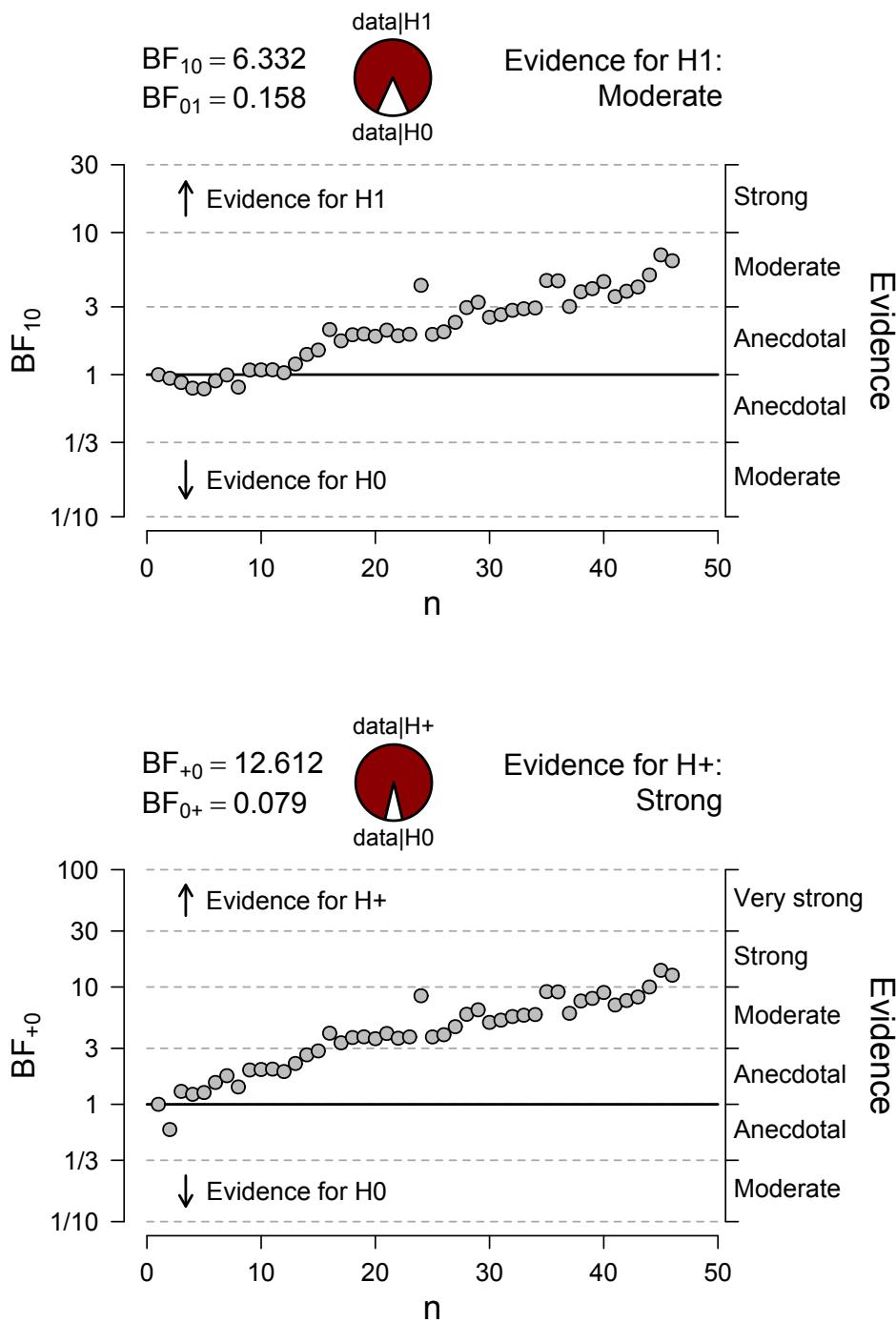


Figure 2.6: Forty-six election-long evidential flow for the presence of a correlation between the relative height of the US president and his proportion of the popular vote. Top panel: two-sided analysis; bottom panel: one-sided analysis. Figure based on JASP.

Benefit 5. The Bayes factor is not “violently biased” against \mathcal{H}_0

Given a complete specification of the models under test, the Bayes factor provides a precise assessment of their relative predictive adequacy. Poor predictive adequacy of \mathcal{H}_0 alone is not a sufficient reason to prefer \mathcal{H}_1 ; it is the balance between predictions from \mathcal{H}_0 and \mathcal{H}_1 that is relevant for the assessment of the evidence. As discussed under benefit 1 above, this contrasts with the NHST p value, which only considers the unusualness of the data under \mathcal{H}_0 . Consequently, statisticians have repeatedly pointed out that “Classical significance tests are violently biased against the null hypothesis.” (W. Edwards, 1965, p. 400; see also Johnson, 2013; Sellke et al., 2001). Based on a comparison between p values and Bayes factors, (J. O. Berger & Delampady, 1987, p. 330) argued that “First and foremost, when testing precise hypotheses, formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against \mathcal{H}_0 .” In a landmark article, W. Edwards et al. (1963, p. 228) concluded that “Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest.” Finally, Lindley suggested, somewhat cynically perhaps, that this bias is precisely the reason for the continued popularity of p values: “There is therefore a serious and systematic difference between the Bayesian and Fisherian calculations, in the sense that a Fisherian approach much more easily casts doubt on the null value than does Bayes. Perhaps this is why significance tests are so popular with scientists: they make effects appear so easily.” (Lindley, 1986, p. 502).

The p value bias against \mathcal{H}_0 is also evident from the election example, where a correlation of .39, displayed in Figure 2.1, yields $p = .007$ and $BF_{10} = 6.33$. Even though in this particular case both numbers roughly support the same conclusion (i.e., “reject \mathcal{H}_0 ” versus “evidence for \mathcal{H}_1 ”), the p value may suggest that the evidence is compelling, whereas the Bayes factor leaves considerable room for doubt. An extensive empirical comparison between p values and Bayes factors can be found in Wetzels et al. (2011). For a Bayesian interpretation of the classical p value see Marsman and Wagenmakers (in press).

In sum, the Bayes factor conditions on the observed data to grade the degree of evidence that the data provide for \mathcal{H}_0 versus \mathcal{H}_1 . As a thermometer for the intensity of the evidence –either for \mathcal{H}_0 or for \mathcal{H}_1 – the Bayes factor allows researchers to monitor the evidential flow as the data accumulate, and stop whenever they feel the evidence is compelling or the resources have been depleted. Bayes factors can be computed and interpreted even when the intention with which the data have been collected is unknown or entirely absent, such as when the data are provided by a natural process without an experimenter. Moreover, its predictive nature ensures that the Bayes factor does not require either model to be true.

Ten Objections to the Bayes Factor Hypothesis Test

Up to this point we have provided a perspective on Bayesian estimation and Bayesian hypothesis testing that may be perceived as overly optimistic. Bayesian inference does not solve all of the problems that confront the social sciences today. Other important problems include the lack of data sharing and the blurred distinction between exploratory and confir-

matory work (e.g., Chambers, 2013; De Groot, 1956/2014; Nosek et al., 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), not to mention the institutional incentive structure to “publish or perish” (Nosek et al., 2012). Nevertheless, as far as statistical inference is concerned, we believe that the adoption of Bayesian procedures is a definite step in the right direction.

In addition, our enthusiasm for Bayes factor hypothesis testing is shared by only a subset of modern-day Bayesian statisticians (e.g., Albert, 2007; J. O. Berger & Pericchi, 2001; Bové & Held, 2011; Liang, Paulo, Molina, Clyde, & Berger, 2008; Maruyama & George, 2011; Ntzoufras, Dellaportas, & Forster, 2003; Ntzoufras, 2009; O’Hagan, 1995; Overstall & Forster, 2010; Raftery, 1999; for an alternative perspective see e.g., Robert, 2016). In fact, the topic of Bayes factors is contentious to the extent that it provides a dividing line between different schools of Bayesians. In recognition of this fact, and in order to provide a more balanced presentation, we now discuss a list of ten objections against the approach we have outlined so far. A warning to the uninitiated reader: some of the objections and counterarguments may be difficult to understand from a superficial reading alone; trained statisticians and philosophers have debated these issues for many decades, without much resolution in sight.

Objection 1: Estimation is Always Superior to Testing

As mentioned in the introduction, it is sometimes argued that researchers should abandon hypothesis tests in favor of parameter estimation (e.g., Cumming, 2014). We agree that parameter estimation is an important and unduly neglected part of the inductive process in current-day experimental psychology, but we believe that ultimately both hypothesis testing and parameter estimation have their place, and a complete report features results from both approaches (J. O. Berger, 2006).

Parameter estimation is most appropriate when the null hypothesis is not of any substantive research interest. For instance, in political science one may be interested in polls that measure the relative popularity of various electoral candidates; the hypothesis that all candidates are equally popular is uninteresting and irrelevant. Parameter estimation is also appropriate when earlier work has conclusively ruled out the null hypothesis as a reasonable explanation of the phenomenon under consideration. For instance, a study of the Stroop effect need not assign prior mass to the hypothesis that the effect is absent. In sum, whenever prior knowledge or practical considerations rule out the null hypothesis as a plausible or interesting explanation then a parameter estimation approach is entirely defensible and appropriate.

Other research scenarios, however, present legitimate testing problems. An extreme example concerns precognition: the question at hand is not “Assuming that people can look into the future, how strong is the effect?” – rather, the pertinent question is “Can people look into the future?”. The same holds for medical clinical trials, where the question at hand is not “Assuming the new treatment works, how strong is the effect?” but instead is “Does the new treatment work?”. Note that in these examples, the parameter estimation question presupposes that the effect exists, whereas the hypothesis testing question addresses whether that supposition is warranted in the first place.

The relation between estimation and testing is discussed in detail in Jeffreys's book "Theory of Probability". For instance, Jeffreys provides a concrete example of the difference between estimation and testing:

"The distinction between problems of estimation and significance arises in biological applications, though I have naturally tended to speak mainly of physical ones. Suppose that a Mendelian finds in a breeding experiment 459 members of one type, 137 of the other. The expectations on the basis of a 3 : 1 ratio would be 447 and 149. The difference would be declared not significant by any test. But the attitude that refuses to attach any meaning to the statement that the simple rule is right must apparently say that if any predictions are to be made from the observations the best that can be done is to make them on the basis of the ratio 459/137, with allowance for the uncertainty of sampling. I say that the best is to use the 3/1 rule, considering no uncertainty beyond the sampling errors of the new experiments. In fact the latter is what a geneticist would do. The observed result would be recorded and might possibly be reconsidered at a later stage if there was some question of differences of viability after many more observations had accumulated; but meanwhile it would be regarded as confirmation of the theoretical value. This is a problem of what I call significance.

But what are called significance tests in agricultural experiments seem to me to be very largely problems of pure estimation. When a set of varieties of a plant are tested for productiveness, or when various treatments are tested, it does not appear to me that the question of presence or absence of differences comes into consideration at all. It is already known that varieties habitually differ and that treatments have different effects, and the problem is to decide which is the best; that is, to put the various members, as far as possible, in their correct order." (Jeffreys, 1961, p. 389).¹¹

Moreover, Jeffreys argues that a sole reliance on estimation results in inferential chaos:

"These are all problems of pure estimation. But their use as significance tests covers a looseness of statement of what question is being asked. They give the correct answer if the question is: If there is nothing to require consideration of some special values of the parameter, what is the probability distribution of that parameter given the observations? But the question that concerns us in significance tests is: If some special value has to be excluded before we can assert any other value, what is the best rule, on the data available, for deciding whether to retain it or adopt a new one? The former is what I call a problem of estimation, the latter of significance. Some feeling of discomfort seems to attach itself to the assertion of the special value as *right* since it may be slightly wrong but not sufficiently to be revealed by a test on the data available; but no significance test

¹¹Jeffreys's statement that treatment effects are the domain of estimation may appear inconsistent with our claim that medical clinical trials are the domain of testing. However, the difference is that Jeffreys's treatment effects are random, whereas the treatment in a clinical trial is targeted (see also footnote 1 in Bayarri, Benjamin, Berger, & Sellke, 2016).

asserts it as certainly right. We are aiming at the best way of progress, not at the unattainable ideal of immediate certainty. What happens if the null hypothesis is retained after a significance test is that the maximum likelihood solution or a solution given by some other method of estimation is rejected. The question is, When we do this, do we expect thereby to get more or less correct inferences than if we followed the rule of keeping the estimation solution regardless of any question of significance? I maintain that the only possible answer is that we expect to get more. The difference as estimated is interpreted as random error and irrelevant to future observations. In the last resort, if this interpretation is rejected, there is no escape from the admission that a new parameter may be needed for every observation, and then all combination of observations is meaningless, and the only valid presentation of data is a mere catalogue without any summaries at all.” (Jeffreys, 1961, pp. 387-388)

In light of these and other remarks, Jeffreys’s maxim may be stated as follows: “Do not try to estimate something until you are sure there is something to be estimated.”¹²

Finally, in some applications the question of estimation never arises. Examples include cryptography (Turing, 1941/2012; Zabell, 2012), the construction of phylogenetic trees (Huelsenbeck & Ronquist, 2001), and the comparison of structurally different models (e.g., in the field of response time analysis: the diffusion model versus the linear ballistic accumulator model; in the field of categorization: prototype versus exemplar models; in the field of visual working memory: discrete slot models versus continuous resource models; in the field of long-term memory: multinomial processing tree models versus models based on signal detection theory).

In sum, hypothesis testing and parameter estimation are both important. In the early stages of a research paradigm, the focus of interest may be on whether the effect is present or absent; in the later stages, if the presence of the effect has been firmly established, the focus may shift towards an estimation approach.

Objection 2: Bayesian Hypothesis Tests Can Indicate Evidence for Small Effects That Are Practically Meaningless

An objection that is often raised against NHST may also be raised against Bayes factor hypothesis testing: with large sample sizes, even small and practically meaningless effects will be deemed “significant” or “strongly supported by the data”. This is true. However, what is practically relevant is context-dependent – in some contexts, small effects can have large consequences. For example, N. J. Goldstein, Cialdini, and Griskevicius (2008) reported

¹²This is inspired by what is known as Hyman’s maxim for ESP, namely “Do not try to explain something until you are sure there is something to be explained.” (Alcock, 1994, p. 189, see also <http://www.skeptic.com/insight/history-and-hymans-maxim-part-one/>). For a similar perspective see Paul Alper’s comment on what Harriet Hall termed “Tooth fairy science” https://www.causeweb.org/wiki/chance/index.php/Chance_News_104#Tooth_fairy_science: “Yes, you have learned something. But you haven’t learned what you think you’ve learned, because you haven’t bothered to establish whether the Tooth Fairy really exists”.

that messages to promote hotel towel reuse are more effective when they also attend guests to descriptive norms (e.g., “the majority of guests reuse their towels”). Based on a total of seven published experiments, a Bayesian meta-analysis suggests that this effect is present ($BF_{10} \approx 37$) but relatively small, around 6% (Scheibehenne, Jamil, & Wagenmakers, in press). The practical relevance of this result depends on whether or not it changes hotel policy; the decision to change the messages or leave them intact requires hotels to weigh the costs of changing the messages against the expected gains from having to wash fewer towels; for a large hotel, a 6% gain may result in considerable savings.

Thus, from a Bayesian perspective, context-dependence is recognized and incorporated through an analysis that computes expected utilities for a set of possible actions (Lindley, 1985). The best action is the one with the highest expected utility. In other words, the practicality of the effects can be taken into account, if needed, by adding an additional layer of considerations concerning utility. Another method to address this objection is to specify the null hypothesis not as a point but as a practically relevant interval around zero (Morey & Rouder, 2011).¹³

Objection 3: Bayesian Hypothesis Tests Promote Binary Decisions

It is true that Jeffreys and other statisticians have suggested rough descriptive guidelines for the Bayes factor (for a more detailed discussion see Wagenmakers et al., this volume). These guidelines facilitate a discrete verbal summary of a quantity that is inherently continuous. More importantly, regardless of whether it is presented in continuous numerical or discrete verbal form, the Bayes factor grades the evidence that the data provide for \mathcal{H}_0 versus \mathcal{H}_1 – thus, the Bayes factor relates to evidence, not decisions (Ly, Verhagen, & Wagenmakers, 2016a). As pointed out above, decisions require a consideration of actions and utilities of outcomes (Lindley, 1985). In other words, the Bayes factor measure the change in beliefs brought about by the data, or –alternatively– the relative predictive adequacy of two competing models; in contrast, decisions involve the additional consideration of actions and their consequences.

Objection 4: Bayesian Hypothesis Tests Are Meaningless Under Mis-specification

The Bayes factor is a measure of relative rather than absolute performance. When the Bayes factor indicates overwhelming support in favor of \mathcal{H}_1 over \mathcal{H}_0 , for instance, this does not imply that \mathcal{H}_1 provides an acceptable account of the data. Instead, the Bayes factor indicates only that the predictive performance of \mathcal{H}_1 is superior to that of \mathcal{H}_0 ; the absolute performance of \mathcal{H}_1 may well be abysmal.

A simple example illustrates the point. Consider a test for a binomial proportion parameter θ . Assume that the null hypothesis specifies a value of interest θ_0 , and assume that the alternative hypothesis postulates that θ is lower than θ_0 , with each value of θ judged equally likely a priori. Hence, the Bayes factor compares $\mathcal{H}_0 : \theta = \theta_0$ against $\mathcal{H}_1 : \theta \sim \text{Uniform}(0, \theta_0)$

¹³We plan to include this functionality in a future version of JASP.

(e.g., Haldane, 1932; Etz & Wagenmakers, 2016). Now assume that the data consist of a sequence of length n that features only successes (e.g., items answered correctly, coin tosses landing tails, patients being cured). In this case the predictions of \mathcal{H}_0 are superior to those of \mathcal{H}_1 . A straightforward derivation¹⁴ shows that the Bayes factor in favor of \mathcal{H}_0 against \mathcal{H}_1 equals $n + 1$, *regardless* of θ_0 .¹⁵ Thus, when n is large the Bayes factor will indicate decisive relative support in favor of \mathcal{H}_0 over \mathcal{H}_1 ; at the same time, however, the absolute predictive performance of \mathcal{H}_0 depends crucially on θ_0 , and becomes abysmal when θ_0 is low.

The critique that the Bayes factor does not quantify absolute fit is therefore entirely correct, but it pertains to statistical modeling across the board. Before drawing strong inferential conclusions, it is always wise to plot the data, inspect residuals, and generally confirm that the model under consideration is not misspecified in a major way. The canonical example of this is Anscombe's quartet, displayed here in Figure 2.7 (see also Andraszewicz et al., 2015; Anscombe, 1973; Heathcote, Brown, & Wagenmakers, 2015; Lindsay, 2015). Each panel of the quartet displays two variables with the same mean and variance. Moreover, for the data in each panel the Pearson correlation coefficient equals $r = 0.816$. An automatic analysis of the data from each panel yields the same four p values, the same four confidence intervals, the same four Bayes factors, and the same four credible intervals. Yet a mere glance at Figure 2.7 suggests that these inferential conclusions are meaningful only for the data from the top left panel.

Objection 5: Vague Priors are Preferable over Informed Priors

Bayes factors cannot be used with extremely vague or “uninformative” prior distributions for the parameters under test. For instance, a t -test on effect size δ cannot specify $\mathcal{H}_1 : \delta \sim \text{Uniform}(-\infty, \infty)$, as this leaves the Bayes factor undefined. The use of an almost uninformative prior does not solve the problem; the specification $\mathcal{H}_1 : \delta \sim \text{Uniform}(-10^{100}, 10^{100})$ means that for all sets of reasonable data, the null hypothesis will be strongly preferred. The reason for this behavior is that with such a vague prior, \mathcal{H}_1 predicts that effect size is virtually certain to be enormous; these predictions are absurd, and \mathcal{H}_1 is punished accordingly (Rouder & Morey, 2012).

Consequently, a reasonable comparison between \mathcal{H}_0 and \mathcal{H}_1 requires that both models are specified in a reasonable way (e.g., Dienes, 2011; Vanpaemel, 2010; Vanpaemel & Lee, 2012). Vague priors for effect size are not reasonable. In parameter estimation such unreasonableness usually does not have negative consequences, but this is different for Bayes factor hypothesis testing. Thus, the core problem is not with Bayes factors – the core problem is with unreasonable prior distributions.

Objection 6: Default Priors are not Sufficiently Subjective

Jeffreys (1961) and other “objective” Bayesians have proposed default priors that are intended to be used regardless of the area of substantive application. These default priors

¹⁴See supplemental materials available at the Open Science Framework, <https://osf.io/m6bi8/>.

¹⁵This surprising result holds as long as $\theta_0 > 0$.

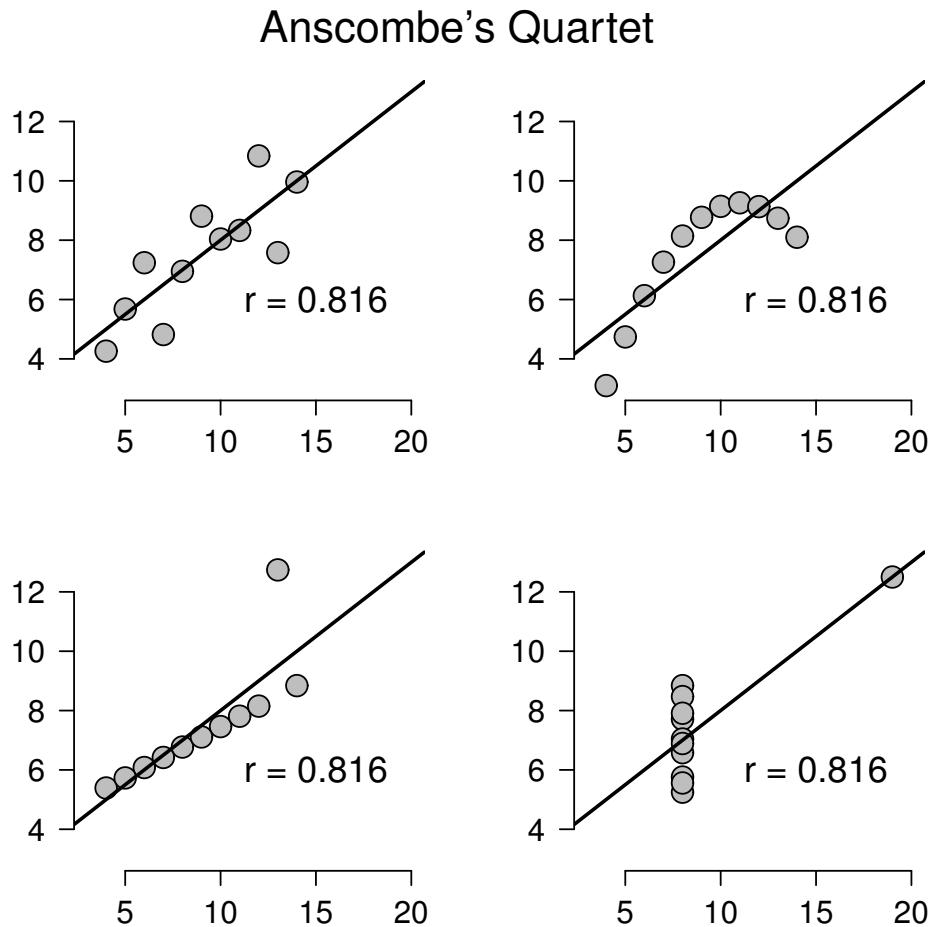


Figure 2.7: “Anscombe’s quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, $r = 0.816$. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. Despite the equivalence of the four data patterns in terms of popular summary measures, the graphical displays reveal that the patterns are very different from one another, and that the Pearson correlation (a linear measure of association) is only valid for the data set from the top left panel.”(Heathcote et al., 2015, p. 34). Figure available at <http://tinyurl.com/zv2shlx> under CC license <https://creativecommons.org/licenses/by/2.0/>.

provide a reference result that can be refined by including subjective knowledge. However, “subjective” Bayesians may argue that this needs to be done always, and the subjectivity in the specification of priors for Bayes factor hypothesis testing does not go far enough. For instance, the t -test involves the specification $\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, r)$. But is it reasonable for the Cauchy distribution to be centered on zero, such that the most likely value for effect size under \mathcal{H}_1 equals zero? Perhaps not (e.g., Johnson, 2013). In addition, the Cauchy form itself may be questioned. Perhaps each analysis attempt should be preceded by a detailed

prior elicitation process, such that \mathcal{H}_1 can be specified in a manner that incorporates all prior knowledge that can be brought to bear on the problem at hand.

The philosophical position of the subjective Bayesian is unassailable, and if the stakes are high enough then every researcher would do well to turn into a subjective Bayesian. However, the objective or consensus Bayesian methodology affords substantial practical advantages: it requires less effort, less knowledge, and it facilitates communication (e.g., J. Berger, 2004; but see M. Goldstein, 2006). For more complicated models, it is difficult to see how a subjective specification can be achieved in finite time. Moreover, the results of an objective analysis may be more compelling to other researchers than those of a subjective analysis (Morey, Wagenmakers, & Rouder, *in press*). Finally, in our experience, the default priors usually yield results that are broadly consistent with those that would be obtained with a more subjective analysis (see also Jeffreys, 1963). Nevertheless, the exploration of more subjective specifications requires more attention (e.g., Dienes, 2014; Verhagen & Wagenmakers, 2014).

Objection 7: Subjective Priors are not Sufficiently Objective

This is an often-heard objection to Bayesian inference in general: the priors are subjective, and in scientific communication one needs to avoid subjectivity at all cost. Of course, this objection ignores the fact that the specification of statistical models is also subjective – the choice between probit regression, logistic regression, and hierarchical zero-inflated Poisson regression is motivated subjectively, by a mix of prior knowledge and experience with the statistical model under consideration. The same holds for power analyses that are conducted using a particular effect size, the choice of which is based on a subjective combination of previous experimental outcomes and prior knowledge. Moreover, the scientific choices of what hypothesis to test, and how to design a good experiment are all subjective. Despite their subjectivity, the research community has been able, by and large, to assess the reasonableness of the choices made by individual researchers.

When the choice is between a method that is objective but unreasonable versus a method that is subjective but reasonable, most researchers would prefer the latter. The default priors for the Bayes factor hypothesis tests are a compromise solution: they attempt to be reasonable without requiring a complete subjective specification.

Objection 8: Default Priors are Prejudiced Against Small Effects

On his influential blog, Simonsohn has recently argued that default Bayes factor hypothesis tests are prejudiced against small effects.¹⁶ This claim raises the question “Prejudiced compared to what?” Small effects certainly receive more support from a classical analysis, but, as discussed above, this occurs mainly because the classical paradigm is biased against the null as the predictions made by \mathcal{H}_1 are ignored (cf. Figure 2.5). Furthermore, note that for large sample sizes, Bayes factors are guaranteed to strongly support a true \mathcal{H}_1 , even for very small true effect sizes. Moreover, the default nested prior specification of \mathcal{H}_1 makes it

¹⁶<http://datacolada.org/2015/04/09/35-the-default-bayesian-test-is-prejudiced-against-small-effects/>

difficult to collect compelling evidence for \mathcal{H}_0 , so the most prominent advantage is generally with \mathcal{H}_1 , not with \mathcal{H}_0 .

These considerations mean that a Bayes factor analysis may be misleading only under the following combination of factors: a small sample size, a small true effect size, and a prior distribution that represents the expectation that effect size is large. Even under this unfortunate combination of circumstances, the extent to which the evidence is misleading will be modest, at least for reasonable prior distributions and reasonable true effect sizes. The relevant comparison is not between the default Bayes factor and some unattainable Platonic ideal; the relevant comparison is between default Bayes factors and p values. Here we believe that practical experience will show that Bayes factors are more informative and have higher predictive success than that provided by p values.

Objection 9: Increasing Sample Size Solves All Statistical Problems

An increase in sample size will generally reduce the need for statistical inference: with large samples, the signal-to-noise ratio often becomes so high that the data pass Berkson's interocular traumatic test. However, "The interocular traumatic test is simple, commands general agreement, and is often applicable; well-conducted experiments often come out that way. But the enthusiast's interocular trauma may be the skeptic's random error. A little arithmetic to verify the extent of the trauma can yield great peace of mind for little cost." (W. Edwards et al., 1963, p. 217).

Moreover, even high-powered experiments can yield completely uninformative results (Wagenmakers, Verhagen, & Ly, 2016). Consider Study 6 from Donnellan, Lucas, and Cesario (2015), one of nine replication attempts on the reported phenomenon that lonely people take hotter showers (in order to replace the lack of social warmth with physical warmth; Bargh & Shalev, 2012). Although the overall results provided compelling evidence in favor of the null hypothesis (Wagenmakers, Verhagen, & Ly, 2016), three of the nine studies by Donnellan et al. (2015) produced only weak evidence for \mathcal{H}_0 , despite relatively large sample sizes. For instance, Study 6 featured $n = 553$ with $r = .08$, yielding a one-sided $p = 0.03$. However, the default one-sided Bayes factor equals an almost perfectly uninformative $BF_{0+} = 1.61$. This example demonstrates that a high-powered experiment does not need to provide diagnostic information; power is a pre-experimental concept that is obtained by considering all the hypothetical data sets that can be observed. In contrast, evidence is a post-experimental concept, taking into account only the data set that was actually obtained (Wagenmakers et al., 2015).

Objection 10: Bayesian Procedures Can be Hacked Too

In an unpublished paper, Simonsohn has argued that Bayes factors are not immune to the biasing effects of selective reporting, ad-hoc use of transformations and outlier removal, etc. (Simonsohn, 2015a).¹⁷ In other words, Bayes factors can be "hacked" too, just like p values. This observation is of course entirely correct. Any reasonable statistical method

¹⁷The paper is available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2374040.

should be sensitive to selective reporting, for else it does not draw the correct conclusions in case the data were obtained without it. Bayes factors are elegant and often informative, but they cannot work miracles and the value of a Bayes factor rests on the reliability and representativeness of the data at hand.

The following example illustrates a more subtle case of “B-hacking” that is able to skew statistical conclusions obtained from a series of experiments. In 2011, Bem published an article in the *Journal of Personality and Social Psychology* in which he argued that eight of nine experiments provided statistical evidence for precognition (Bem, 2011), that is, the ability of people to anticipate a completely random event (e.g., on which side of the computer screen a picture is going to appear). A default Bayes factor analysis by Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) showed that the evidence was not compelling and in many cases even supported \mathcal{H}_0 . In response, Bem, Utts, and Johnson (2011) critiqued the default prior distribution and re-analyzed the data using their own subjective “precognition prior”. Based on this prior distribution, Bem et al. (2011) reported a combined Bayes factor of 13,669 in favor of \mathcal{H}_1 . The results seems to contrast starkly with those of Wagenmakers et al. (2011); can the subjective specification of the prior distribution exert such a huge effect?

The conflict between Bem et al. (2011) and Wagenmakers et al. (2011) is more apparent than real. For each experiment separately, the Bayes factors from Bem et al. (2011) and Wagenmakers et al. (2011) are relatively similar, a result anticipated by the sensitivity analysis reported in the online supplement to Wagenmakers et al. (2011). The impressive Bayes factor of 13,669 in favor of the precognition hypothesis was obtained by multiplying the Bayes factors for the individual experiments. However, this changes the focus of inference from individual studies to the entire collection of studies as a whole. Moreover, as explained above, multiplying Bayes factors without updating the prior distribution is a statistical mistake (Jeffreys, 1961; Rouder & Morey, 2011; Wagenmakers, Verhagen, & Ly, 2016).

In sum, the Bayes factor conclusions from Bem et al. (2011) and Wagenmakers et al. (2011) are in qualitative agreement about the relatively low evidential impact of the individual studies reported in Bem (2011). The impression of a conflict is caused by a change in inferential focus coupled with a statistical mistake. Bayesian inference is coherent and optimal, but it is not a magic potion that protects against malice or statistical misunderstanding.

Concluding Comments

Substantial practical rewards await the pragmatic researcher who decides to adopt Bayesian methods of parameter estimation and hypothesis testing. Bayesian methods can incorporate prior information, they do not depend on the intention with which the data were collected, and they can be used to quantify and monitor evidence, both in favor of \mathcal{H}_0 and \mathcal{H}_1 . In depressing contrast, classical procedures apply only in the complete absence of knowledge about the topic at hand, they require knowledge of the intention with which the data were collected, they are biased against the null hypothesis, and they can yield conclusions that, although valid on average, may be absurd for the case at hand.

Despite the epistemological richness and practical benefits of Bayesian parameter estimation and Bayesian hypothesis testing, the practice of reporting p values continues its dominant reign. As outlined in the introduction, the reasons for resisting statistical innovation are manyfold (Sharpe, 2013). In recent years our work has focused on overcoming one reason for resistance: the real or perceived difficulty of obtaining default Bayesian answers for run-of-the-mill statistical scenarios involving correlations, the t -test, ANOVA and others. To this aim we have developed JASP, a software program that allows the user to conduct both classical and Bayesian analyses.¹⁸ An in-depth discussion of JASP is provided in Part II of this series (Wagenmakers et al., this volume).

References

- Albert, J. (2007). *Bayesian computation with R*. New York: Springer.
- Alcock, J. (1994). Afterword: An analysis of psychic sleuths' claims. In J. Nickell (Ed.), *Psychic sleuths: ESP and sensational cases* (pp. 172–190). Buffalo, NY: Prometheus Books.
- Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, 41, 521–543.
- Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, 58, 365–383.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17–21.
- Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion*, 12, 154–162.
- Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.
- Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531–533.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101, 716–719.
- Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis*, 1, 1–17.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.
- 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.

¹⁸The development of JASP was made possible by the ERC grant “Bayes or bust: Sensible hypothesis tests for social scientists”.

- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2, 317–352.
- Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94, 542–554.
- Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In P. Lahiri (Ed.), *Model selection* (pp. 135–207). Beachwood, OH: Institute of Mathematical Statistics Lecture Notes—Monograph Series, volume 38.
- Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.
- Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods*, 38, 65–76.
- Bové, D. S., & Held, L. (2011). Hyper- g priors for generalized linear models. *Bayesian Analysis*, 6, 387–410.
- Brown, L. (1967). The conditional level of Student's t test. *The Annals of Mathematical Statistics*, 38, 1068–1071.
- Buehler, R. J., & Fedderson, A. P. (1963). Note on a conditional property of Student's t . *The Annals of Mathematical Statistics*, 34, 1098–1100.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 1–12.
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49, 609–610.
- Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics*, 25, 617–657.
- Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29, 357–372.
- Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286–300.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
- Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147, 278–292.
- Dawid, A. P. (2000). Comment on “The philosophy of statistics” by D. V. Lindley. *The Statistician*, 49, 325–326.
- Dawid, A. P. (2005). Statistics on trial. *Significance*, 2, 6–8.
- de Finetti, B. (1974). *Theory of probability, vol. 1 and 2*. New York: John Wiley & Sons.
- De Groot, A. D. (1956/2014). The meaning of “significance” for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, 148, 188–194.

- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave MacMillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781.
- Donnellan, M. B., Lucas, R. E., & Cesario, J. (2015). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1. *Emotion*, 15, 109–119.
- Eagle (Ed.), A. (2011). *Philosophy of probability: Contemporary readings*. New York: Routledge.
- Edwards, A. W. F. (1992). *Likelihood*. Baltimore, MD: The Johns Hopkins University Press.
- Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, 63, 400–402.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, 53, 134–140.
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2016). How to become a Bayesian in eight easy steps: An annotated reading list. *Manuscript submitted for publication*.
- Etz, A., & Wagenmakers, E.-J. (2016). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Manuscript submitted for publication and uploaded to ArXiv*.
- Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). New York: Hafner.
- Fitts, D. A. (2010). Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behavior Research Methods*, 42, 3–22.
- Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, 30, 690–697.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439–453.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton (FL): Chapman & Hall/CRC.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.
- Gillispie, C. C. (1997). *Pierre-Simon Laplace 1749–1827: A life in exact science*. Princeton, NJ: Princeton University Press.
- Gleser, L. J. (2002). Setting confidence intervals for bounded parameters: Comment. *Statistical Science*, 17, 161–163.
- Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1, 403–420.
- Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35, 472–482.
- Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 69, 54–61.
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (in press). Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician*.
- Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 55–61.
- Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods*, 12, 179–185.
- Hartshorne, C., & Weiss, P. (Eds.). (1932). *Collected papers of Charles Sanders Peirce: Volume II: Elements of logic*. Cambridge: Harvard University Press.
- Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7, 185–207.
- Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York: Springer.
- Hill, R. (2005). Reflections on the cot death cases. *Significance*, 2, 13–15.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157–1164.
- Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: Chapman & Hall/CRC.
- Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.
- Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 754–755.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 696–701.
- JASP Team. (2016). *JASP (Version 0.8)*[Computer software].
- Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science, Vol. II* (pp. 175–257). Dordrecht, Holland: D. Reidel Publishing Company.

- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
- Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80, 64–72.
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31, 203–222.
- Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
- Jeffreys, H. (1963). Review of “the foundations of statistical inference”. *Technometrics*, 3, 407–410.
- Jeffreys, H. (1973). *Scientific inference* (3 ed.). Cambridge, UK: Cambridge University Press.
- Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 451–453). Amsterdam, The Netherlands: North-Holland Publishing Company.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524–532.
- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317.
- Joyce, J. M. (1998). A non-pragmatic vindication of probabilism. *Philosophy of Science*, 65, 575–603.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10, 477–493.
- Kruschke, J. K. (2010a). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Burlington, MA: Academic Press.
- Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14, 293–300.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6, 299–312.
- Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9, 278–292.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1–15.
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55, 1–7.
- Lee, M. D., Fuss, I., & Navarro, D. (2006). A Bayesian approach to diffusion models of decision-making and response time. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 809–815). Cambridge, MA: MIT Press.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with

- the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, 92, 648–655.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.
- Lindley, D. V. (1965). *Introduction to probability & statistics from a Bayesian viewpoint. Part 2. Inference*. Cambridge: Cambridge University Press.
- Lindley, D. V. (1980). Jeffreys's contribution to modern statistical thought. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 35–39). Amsterdam, The Netherlands: North-Holland Publishing Company.
- Lindley, D. V. (1985). *Making decisions* (2 ed.). London: Wiley.
- Lindley, D. V. (1986). Comment on “tests of significance in theory and practice” by D. J. Johnstone. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 35, 502–504.
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15, 22–25.
- Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49, 293–337.
- Lindley, D. V. (2004). That wretched prior. *Significance*, 1, 85–87.
- Lindley, D. V. (2006). *Understanding uncertainty*. Hoboken: Wiley.
- Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26, 1827–1832.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton (FL): Chapman & Hall/CRC.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Marin, J.-M., & Robert, C. P. (2007). *Bayesian core: A practical approach to computational Bayesian statistics*. New York: Springer.
- Marsman, M., & Wagenmakers, E.-J. (in press). Three insights from a Bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*.
- Maruyama, Y., & George, E. I. (2011). Fully Bayes factors with a generalized g -prior. *The Annals of Statistics*, 39, 2740–2765.
- Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43, 679–690.
- 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., Romeijn, J., & Rouder, J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, 66, 68–75.

- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16, 406–419.
- 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.
- Morey, R. D., Rouder, J. N., Verhagen, A. J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, 25, 1289–1290.
- Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (in press). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. New Brunswick (N.J.): Transaction Publishers.
- Mulaik, S., & Steiger, J. (1997). *What if there were no significance tests*. Mahwah, New Jersey: Erlbaum.
- Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, 53, 530–546.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90–100.
- Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, 44(1–2).
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79–95.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 236, 333–380.
- Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, 138, 517–534.
- Nobles, R., & Schiff, D. (2005). Misleading statistics within criminal trials: The Sally Clark case. *Significance*, 2, 17–19.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science*, 348, 1422–1425.
- Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23, 217–243.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615–631.
- Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.
- Ntzoufras, I., Dellaportas, P., & Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111, 165–180.
- Nuzzo, R. (2014). Statistical errors. *Nature*, 506, 150–152.

- O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, 57, 99–138.
- O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.
- Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54, 3269–3288.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530.
- Peirce, C. S. (1878a). Deduction, induction, and hypothesis. *Popular Science Monthly*, 13, 470–482.
- Peirce, C. S. (1878b). The probability of induction. *Popular Science Monthly*, 12, 705–718.
- Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *The Annals of Statistics*, 1, 241–250.
- Pratt, J. W. (1961). Review of Lehmann, E. L., testing statistical hypotheses. *Journal of the American Statistical Association*, 56, 163–167.
- Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge, MA: MIT Press.
- Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1591–1607.
- Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, 27, 411–427.
- Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan Paul.
- Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72, 33–37.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301–308.
- Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, 137, 370–389.
- Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12, 195–223.
- Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, 72, 621–642.
- Rouder, J. N., & Morey, R. D. (2011). A Bayes-factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18, 682–689.
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877–903.
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. P. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, 14, 597–605.

- 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.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
- Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (in press). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*.
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (in press). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.
- Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, 55, 62–71.
- Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, 18, 572–582.
- 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.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- Simonsohn, U. (2015a). Posterior-hacking: Selective reporting invalidates Bayesian results also. *Unpublished manuscript*.
- Simonsohn, U. (2015b). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26, 559–569.
- Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, 24, 159–171.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic And Applied Social Psychology*, 37, 1–2.
- Turing, A. M. (1941/2012). The applications of probability to cryptography. *UK National Archives, HW 25/37*.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- van Erven, T., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society B*, 74, 361–417.
- van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (in press). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.

- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.
- Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, 19, 1047–1056.
- Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143, 1457–1475.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, 14, 779–804.
- Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50, 149–166.
- 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., et al. (this volume). Bayesian statistical inference for psychological science. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*.
- 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. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48, 413–426.
- Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., et al. (2015). A power fallacy. *Behavior Research Methods*, 47, 913–917.
- Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (in press). 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*. John Wiley and Sons.
- Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, 50(2).
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100, 426–432.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 627–633.
- Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19, 326–339.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6, 291–298.
- Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 19, 1057–1064.
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42, 369–390.

- Wrinch, D., & Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 45, 368–374.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46, 441–517.
- Zabell, S. (2012). Commentary on Alan M. Turing: The applications of probability to cryptography. *Cryptologia*, 36, 191–214.

Example applications with JASP

**Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil,
Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F. Gronau, Damian
Dropmann, Bruno Boutin, Frans Meerhoff, Patrick Knight, Akash Raj,
Erik-Jan van Kesteren, Johnny van Doorn, Martin Šmíra, Sacha Epskamp,
Alexander Etz, Dora Matzke, Tim de Jong, Don van den Bergh,
Alexandra Sarafoglou, Helen Steingrover, Koen Derkx, Jeffrey N.
Rouder, and Richard D. Morey**

As demonstrated in part I of this series, Bayesian inference unlocks a series of advantages that remain unavailable to researchers who continue to rely solely on classical inference (Wagenmakers et al., *in press*). For example, Bayesian inference allows researchers to update knowledge, to draw conclusions about the specific case under consideration, to quantify evidence for the null hypothesis, and to monitor evidence until the result is sufficiently compelling or the available resources have been depleted. Generally, Bayesian inference yields intuitive and rational conclusions within a flexible framework of information updating. As a method for drawing scientific conclusions from data, we believe that Bayesian inference is more appropriate than classical inference.

Pragmatic researchers may have a preference that is less pronounced. These researchers may feel it is safest to adopt an inclusive statistical approach, one in which classical and Bayesian results are reported together; if both results point in the same direction this increases one's confidence that the overall conclusion is robust. Nevertheless, both pragmatic researchers and hardcore Bayesian advocates have to overcome the same hurdle, namely, the difficulty in transitioning from Bayesian theory to Bayesian practice. Unfortunately, for many researchers it is difficult to obtain Bayesian answers to statistical questions for standard scenarios involving correlations, the *t*-test, analysis of variance (ANOVA), and others. Until recently, these tests had not been implemented in any software, let alone user-friendly software. And in the absence of software, few researchers feel enticed to learn about Bayesian

inference and few teachers feel enticed to teach it to their students.

To narrow the gap between Bayesian theory and Bayesian practice we developed JASP (JASP Team, 2017), an open-source statistical software program with an attractive graphical user interface (GUI). The JASP software package is cross-platform and can be downloaded free of charge from jasp-stats.org. Originally conceptualized to offer only Bayesian analyses, the current program allows its users to conduct both classical and Bayesian analyses.¹ Using JASP, researchers can conduct Bayesian inference by dragging and dropping the variables of interest into analysis panels, whereupon the associated output becomes available for inspection. JASP comes with default priors on the parameters that can be changed whenever this is deemed desirable.

This article summarizes the general philosophy behind the JASP program and then presents five concrete examples that illustrate the most popular Bayesian tests implemented in JASP. For each example we discuss the correct interpretation of the Bayesian output. Throughout, we stress the insights and additional possibilities that a Bayesian analysis affords, referring the reader to background literature for statistical details. The article concludes with a brief discussion of future developments for Bayesian analyses with JASP.

The JASP Philosophy

The JASP philosophy is based on several interrelated design principles. First, JASP is free and open-source, reflecting our belief that transparency is an essential element of scientific practice. Second, JASP is inferentially inclusive, featuring classical and Bayesian methods for parameter estimation and hypothesis testing. Third, JASP focuses on the statistical methods that researchers and students use most often; to retain simplicity, add-on modules are used to implement more sophisticated and specialized statistical procedures. Fourth, JASP has a graphical user interface that was designed to optimize the user's experience. For instance, output is dynamically updated as the user selects input options, and tables are in APA format for convenient copy-pasting in text editors such as LibreOffice and Microsoft Word. JASP also uses progressive disclosure, which means that initial output is minimalist and expanded only when the user makes specific requests (e.g., by ticking check boxes). In addition, JASP output retains its state, meaning that the input options are not lost – clicking on the output brings the input options back up, allowing for convenient review, discussion, and adjustment of earlier analyses. Finally, JASP is designed to facilitate open science; from JASP 0.7 onward, users are able to save and distribute data, input options, and output results together as a .jasp file. Moreover, by storing the .jasp file on a public repository such as the Open Science Framework (OSF), reviewers and readers can have easy access to the data and annotated analyses that form the basis of a substantive claim. As illustrated in Figure 3.1, the OSF has a JASP previewer that presents the output from a .jasp file regardless of whether the user has JASP installed. In addition, users with an OSF account can upload, download, edit, and sync files stored in their OSF repositories from within JASP. The examples discussed in this article each come with an annotated .jasp file available on

¹Bayesian advocates may consider the classical analyses a Bayesian Trojan horse.

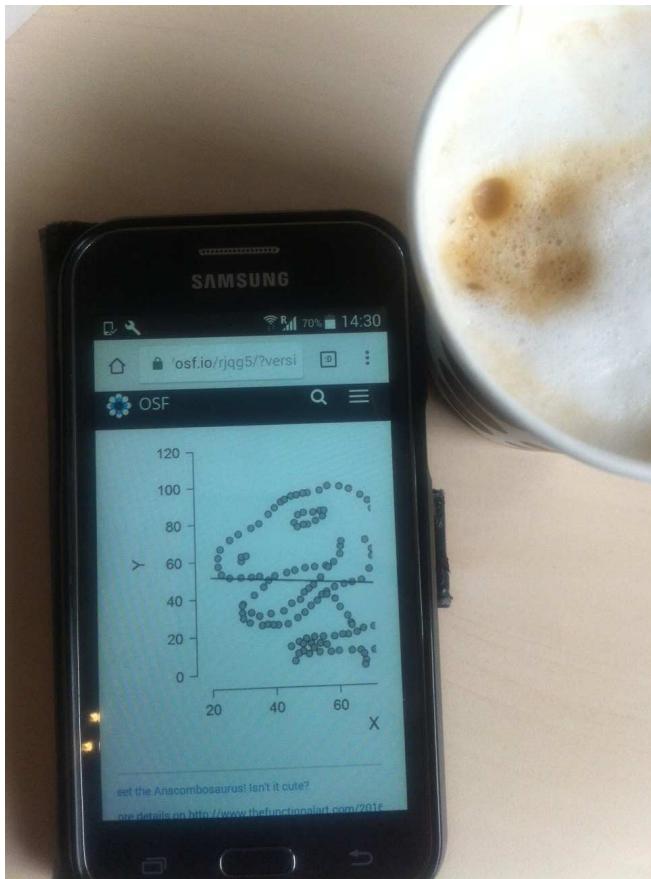


Figure 3.1: The JASP previewer allows users to inspect the annotated output of a .jasp file on the OSF, even without JASP installed and without an OSF account. The graph shown on the cell phone displays the Anscombosaurus (see <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>). Figure available at <https://osf.io/m6bi8/> under a CC-BY license.

the OSF at <https://osf.io/m6bi8/>. Several analyses are illustrated with videos on the JASP YouTube channel.

The JASP GUI is familiar to users of SPSS and has been programmed in C++, html, and javascript. The inferential engine is based on R (R Development Core Team, 2004) and –for the Bayesian analyses– much use is made of the **BayesFactor** package developed by Morey and Rouder (2015) and the **conting** package developed by Overstall and King (2014a). The latest version of JASP uses the functionality of more than 110 different R packages; a list is available on the JASP website at <https://jasp-stats.org/r-package-list/>. The JASP installer does not require that R is installed separately.

Our long-term goals for JASP are two-fold: the primary goal is to make Bayesian benefits more widely available than they are now, and the secondary goal is to reduce the field's dependence on expensive statistical software programs such as SPSS.

Example 1: “A Bayesian Correlation Test for the Height Advantage of US Presidents”

For our first example we return to the running example from Part I. This example concerned the height advantage of candidates for the US presidency (Stulp, Buunk, Verhulst, & Pollet, 2013). Specifically, we were concerned with the Pearson correlation ρ between the proportion of the popular vote and the height ratio (i.e., height of the president divided by the height of his closest competitor). In other words, we wished to assess the evidence that the data provide for the hypothesis that taller presidential candidates attract more votes. The scatter plot was shown in Figure 1 of Part I. Recall that the sample correlation r equaled .39 and was significantly different from zero ($p = .007$, two-sided test, 95% CI [.116, .613]); under a default uniform prior, the Bayes factor equaled 6.33 for a two-sided test and 12.61 for a one-sided test (Wagenmakers et al., in press).

Here we detail how the analysis is conducted in JASP. The left panel of Figure 3.2 shows a spreadsheet view of the data that the user has just loaded from a .csv file using the file tab.² Each column header contains a small icon denoting the variable’s measurement level: continuous, ordinal, or nominal (Stevens, 1946). For this example, the ruler icon signifies that the measurement level is continuous. When loading a data set, JASP uses a “best guess” to determine the measurement level. The user can click the icon, and change the variable type if this guess is incorrect.

After loading the data, the user can select one of several analyses. Presently the functionality of JASP (version 0.8.1) encompasses the following procedures and tests:

- Descriptives (with the option to display a matrix plot for selected variables).
- Reliability analysis (e.g., Cronbach’s α , Gutmann’s λ_6 , and McDonald’s ω).
- Independent samples t -test, paired samples t -test, and one sample t -test. Key references for the Bayesian implementation include Jeffreys (1961); Ly, Verhagen, and Wagenmakers (2016b, 2016a); Rouder, Speckman, Sun, Morey, and Iverson (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009).
- ANOVA, repeated measures ANOVA, and ANCOVA. Key references for the Bayesian implementation include Rouder, Morey, Speckman, and Province (2012), Rouder, Morey, Verhagen, Swagman, and Wagenmakers (in press), and Rouder, Engelhardt, McCabe, and Morey (in press).
- Correlation. Key references for the Bayesian implementation include Jeffreys (1961), Ly et al. (2016b), and Ly, Marsman, and Wagenmakers (in press) for Pearson’s ρ , and van Doorn, Ly, Marsman, and Wagenmakers (in press) for Kendall’s tau.
- Linear regression. Key references for the Bayesian implementation include Liang, Paulo, Molina, Clyde, and Berger (2008); Rouder and Morey (2012), and Zellner and Siow (1980).

²JASP currently reads the following file formats: .jasp, .txt, .csv (i.e., a plain text file with fields separated by commas), .ods (i.e., OpenDocument Spreadsheet, a file format used by OpenOffice), and .sav (i.e., the SPSS file format).

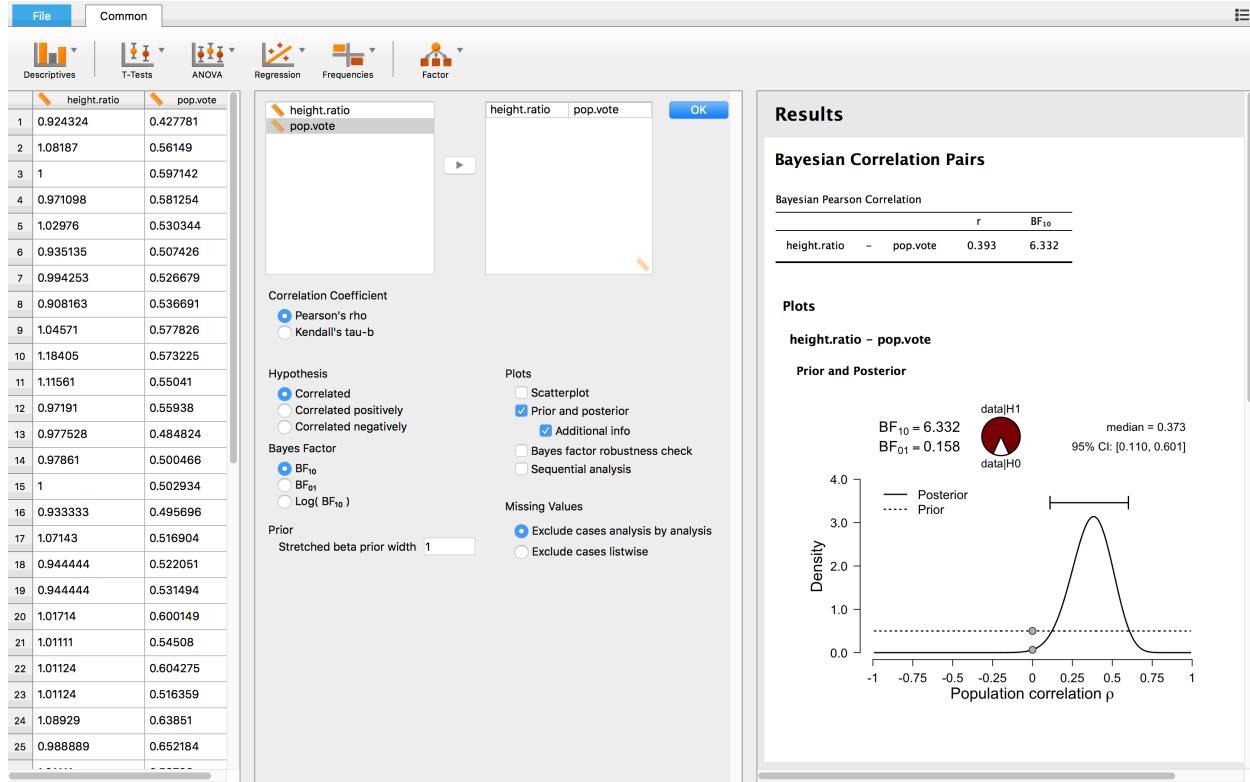


Figure 3.2: JASP screenshot for the two-sided test for the presence of a correlation between the relative height of the US president and his proportion of the popular vote. The left panel shows the data in spreadsheet format; the middle panel shows the analysis input options; the right panel shows the analysis output.

- Binomial test. Key references for the Bayesian implementation include Jeffreys (1961) and O'Hagan and Forster (2004).
- Contingency tables. Key references for the Bayesian implementation include Gunel and Dickey (1974) and Jamil, Ly, et al. (in press).
- Log-linear regression. Key references for the Bayesian implementation include Overstall and King (2014a) and Overstall and King (2014b).
- Principal component analysis and exploratory factor analysis.

Except for reliability analysis and factor analysis, the above procedures are available both in their classical and Bayesian form. Future JASP releases will expand this core functionality and add logistic regression, multinomial tests, and a series of nonparametric techniques. More specialized statistical procedures will be provided through add-on packages so that the main JASP interface retains its simplicity.

The middle panel of Figure 3.2 shows that the user selected a Bayesian Pearson correlation analysis. The two variables to be correlated were selected through dragging and

dropping. The middle panel also shows that the user has not specified the sign of the expected correlation under \mathcal{H}_1 – hence, JASP will conduct a two-sided test. The right panel of Figure 3.2 shows the JASP output; in this case, the user requested and received:

1. The Bayes factor expressed as BF_{10} (and its inverse $\text{BF}_{01} = 1/\text{BF}_{10}$), grading the intensity of the evidence that the data provide for \mathcal{H}_1 versus \mathcal{H}_0 (for details see Part I).
2. A proportion wheel that provides a visual representation of the Bayes factor.
3. The posterior median and a 95% credible interval, summarizing what has been learned about the size of the correlation coefficient ρ assuming that \mathcal{H}_1 holds true.
4. A figure showing (a) the prior distribution for ρ under \mathcal{H}_1 (i.e., the uniform distribution, which is the default prior proposed by Jeffreys, 1961 for this analysis; the user can adjust this default specification if desired), (b) the posterior distribution for ρ under \mathcal{H}_1 , (c) the 95% posterior credible interval for ρ under \mathcal{H}_1 , and (d) a visual representation of the Savage-Dickey density ratio, that is, grey dots that indicate the height of the prior and the posterior distribution at $\rho = 0$ under \mathcal{H}_1 ; the ratio of these heights equals the Bayes factor for \mathcal{H}_1 versus \mathcal{H}_0 (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

Thus, in its current state JASP provides a relatively comprehensive overview of Bayesian inference for ρ , featuring both estimation and hypothesis testing methods.

Before proceeding we wish to clarify the meaning of the proportion wheel or “pizza plot”. The wheel was added to assist researchers who are unfamiliar with the odds formulation of evidence – the wheel provides a visual impression of the continuous strength of evidence that a given Bayes factor provides. In the presidents example $\text{BF}_{10} = 6.33$, such that the observed data are 6.33 times more likely under \mathcal{H}_1 than under \mathcal{H}_0 . To visualize this ratio, we transform it to the 0-1 interval and plot the resulting magnitude as the proportion of a circle (e.g., Tversky, 1969, Figure 1; Lipkus & Hollands, 1999). For instance, the presidents example has a ratio of $\text{BF}_{10} = 6.33$ and a corresponding proportion of $6.33/7.33 \approx 0.86$;³ consequently, the red area (representing the support in favor of \mathcal{H}_1) covers 86% of the circle and the white area (representing the support in favor of \mathcal{H}_0) covers the remaining 14%.

Figure 3.3 gives three further examples of proportion wheels. In each panel, the red area represents the support that the data y provide for \mathcal{H}_1 , and the white area represents the complementary support for \mathcal{H}_0 . Figure 3.3 shows that when $\text{BF}_{10} = 3$, the null hypothesis still occupies a non-negligible 25% of the circle’s area. The wheel can be used to intuit the strength of evidence even more concretely, as follows. Imagine the wheel is a dart board. You put on a blindfold and the board is attached to a wall in a random orientation. You then throw a series of darts until the first one hits the board. You remove the blindfold and observe that the dart has landed in the smaller area. *How surprised are you?* We propose that this measure of imagined surprise provides a good intuition for degree of evidence that a particular Bayes factor conveys (Jamil, Marsman, Ly, Morey, & Wagenmakers, in press).

³With unit prior odds, a ratio of x corresponds to a proportion of $x/(x + 1)$.

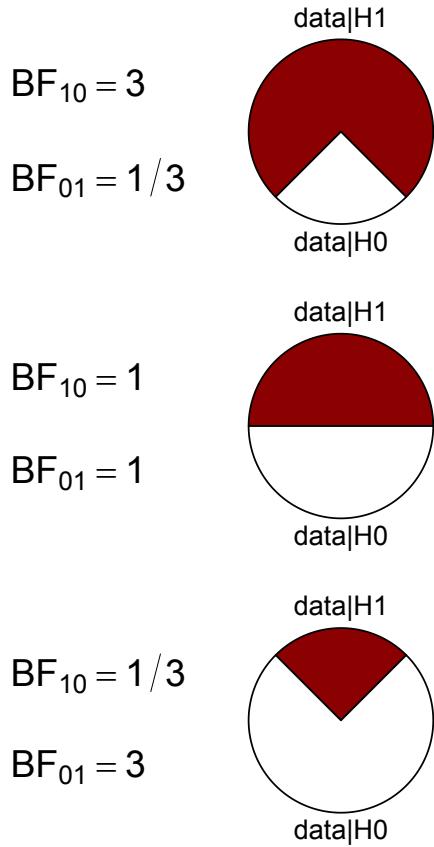


Figure 3.3: Proportion wheels visualize the strength of evidence that a Bayes factor provides. Ratios are transformed to a magnitude between 0 and 1 and plotted as the proportion of a circular area. Imagine the wheel is a dartboard; you put on a blindfold, the wheel is attached to the wall in random orientation, and you throw darts until you hit the board. You then remove the blindfold and find that the dart has hit the smaller area. How surprised are you? The level of imagined surprise provides an intuition for the strength of a Bayes factor. The analogy is visualized in the appendix.

The top panel of Figure 3.3, for instance, represents $\text{BF}_{10} = 3$. Having the imaginary dart land in the white area would be somewhat surprising, but in most scenarios not sufficiently surprising to warrant a strong claim such as the one that usually accompanies a published article. Yet many p -values near the .05 boundary (“reject the null hypothesis”) yield evidence that is weaker than $\text{BF}_{10} = 3$ (e.g., Berger & Delampady, 1987; Edwards, Lindman, & Savage, 1963; Johnson, 2013; Wagenmakers et al., in press; Wetzels et al., 2011). The dart board analogy is elaborated upon in the appendix.

The proportion wheel underscores the fact that the Bayes factor provides a graded, continuous measure of evidence. Nevertheless, for historical reasons it may happen that a discrete judgment is desired (i.e., an all-or-none preference for \mathcal{H}_0 or \mathcal{H}_1). When the competing models are equally likely a priori, then the probability of making an error equals the size

of the smaller area. Note that this kind of “error control” differs from that which is sought by classical statistics. In the Bayesian formulation the probability of making an error refers to the individual case, whereas in classical procedures it is obtained as an average across all possible data sets that could have been observed. Note that the long-run average need not reflect the probability of making an error for a particular case (Wagenmakers et al., in press).

JASP offers several ways in which the present analysis may be refined. In Part I we already showed the results of a one-sided analysis in which the alternative hypothesis \mathcal{H}_+ stipulated the correlation to be positive; this one-sided analysis can be obtained by ticking the check box “correlated positively” in the input panel. In addition, the two-sided alternative hypothesis has a default prior distribution which is uniform from -1 to 1 ; a user-defined prior distribution can be set through the input field “Stretched beta prior width”. For instance, by setting this input field to 0.5 the user creates a prior distribution with smaller width, that is, a distribution which assigns more mass to values of ρ near zero.⁴ Additional check boxes create sequential analyses and robustness checks, topics that will be discussed in the next example.

Example 2: “A Bayesian T-test for a Kitchen Roll Rotation Replication Experiment”

Across a series of four experiments, the data reported in Topolinski and Sparenberg (2012) provided support for the hypothesis that clockwise movements induce psychological states of temporal progression and an orientation toward the future and novelty. Concretely, in their Experiment 2, one group of participants rotated kitchen rolls clockwise, whereas the other group rotated them counterclockwise. While rotating the rolls, participants completed a questionnaire assessing openness to experience. The data from Topolinski and Sparenberg (2012) showed that, in line with their main hypothesis, participants who rotated the kitchen rolls clockwise reported more openness to experience than participants who rotated them counterclockwise (but see Francis, 2013).

We recently attempted to replicate the kitchen roll experiment from Topolinski and Sparenberg (2012), using a preregistered analysis plan and a series of Bayesian analyses (Wagenmakers et al., 2015, <https://osf.io/uszvx/>). Thanks to the assistance of the original authors, we were able to closely mimic the setup of the original study. The apparatus and setup for the replication experiment are shown in Figure 3.4.

Before turning to a JASP analysis of the data, it is informative to recall the stopping rule procedure specified in the online preregistration form (<https://osf.io/p3isc/>):

“We will collect a minimum of 20 participants in each between-subject condition (i.e., the clockwise and counterclockwise condition, for a minimum of 40 participants in total). We will then monitor the Bayes factor and stop the experiment whenever the critical hypothesis test (detailed below) reach a Bayes

⁴Statistical detail: the stretched beta prior is a $\text{beta}(a, a)$ distribution transformed to cover the interval from -1 to 1 . The prior width is defined as $1/a$. For instance, setting the stretched beta prior width equal to 0.5 is conceptually the same as using a $\text{beta}(2, 2)$ distribution on the 0 - 1 interval and then transforming it to cover the interval from -1 to 1 , such that it is then symmetric around $\rho = 0$.

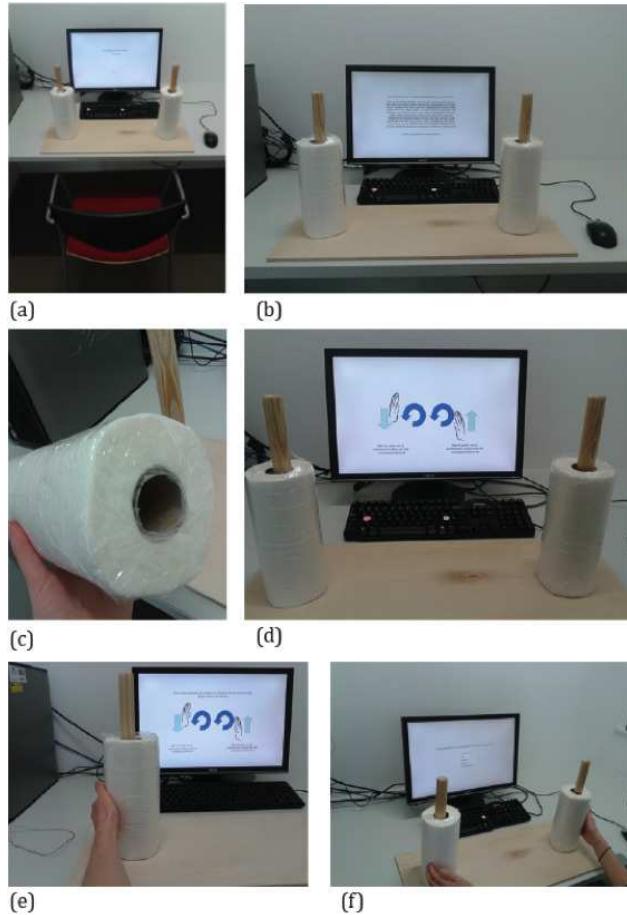


Figure 3.4: The experimental setting from Wagenmakers et al. (2015): (a) the set-up; (b) the instructions; (c) a close-up of one of the sealed paper towels; (d) the schematic instructions; Photos (e) and (f) give an idea of how a participant performs the experiment. Figure available at <https://www.flickr.com/photos/130759277@N05/>, under CC license <https://creativecommons.org/licenses/by/2.0/>.

factor that can be considered “strong” evidence (Jeffreys, 1961); this means that the Bayes factor is either 10 in favor of the null hypothesis, or 10 in favor of the alternative hypothesis. The experiment will also stop whenever we reach the maximum number of participants, which we set to 50 participants per condition (i.e., a maximum of 100 participants in total). Finally, the experiment will also stop on October 1st, 2013. From a Bayesian perspective the specification of this sampling plan is needlessly precise; we nevertheless felt the urge to be as complete as possible.”

In addition, the preregistration form indicated that the Bayes factor of interest is the default one-sided t -test as specified in Rouder et al. (2009) and Wetzels et al. (2009). The two-sided version of this test was originally proposed by Jeffreys (1961), and it involves a comparison of two hypothesis for effect size δ : the null hypothesis \mathcal{H}_0 postulates that

effect size is absent (i.e., $\delta = 0$), whereas the alternative hypothesis \mathcal{H}_1 assigns δ a Cauchy prior centered on 0 with interquartile range $r = 1$ (i.e., $\delta \sim \text{Cauchy}(0, 1)$). The Cauchy distribution is similar to the normal distribution but has fatter tails; it is a t -distribution with a single degree of freedom. Jeffreys chose the Cauchy because it makes the test “information consistent”: with two observations measured without noise (i.e., $y_1 = y_2$) the Bayes factor in favor of \mathcal{H}_1 is infinitely large. The one-sided version of Jeffreys’s test uses a folded Cauchy with positive effect size only, that is, $\mathcal{H}_+ : \delta \sim \text{Cauchy}^+(0, 1)$.

The specification $\mathcal{H}_+ : \delta \sim \text{Cauchy}^+(0, 1)$ is open to critique. Some people feel that this distribution is unrealistic because it assigns too much mass to large effect sizes (i.e., 50% of the posterior mass is on values for effect size larger than 1); in contrast, others feel that this distribution is unrealistic because it assigns most mass to values near zero (i.e., $\delta = 0$ is the most likely value). It is possible to reduce the value of r , and, indeed, the `BayesFactor` package uses a default value of $r = \frac{1}{2}\sqrt{2} \approx 0.707$, a value that JASP has adopted as well. Nevertheless, the use of a very small value of r implies that \mathcal{H}_1 and \mathcal{H}_0 closely resemble one another in the sense that both models make similar predictions about to-be-observed data; this setting therefore makes it difficult to obtain compelling evidence, especially in favor of a true \mathcal{H}_0 (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, in press). In general, we feel that reducing the value of r is recommended if the location of the prior distribution is also shifted away from $\delta = 0$. Currently JASP fixes the prior distribution under \mathcal{H}_1 to the location $\delta = 0$, and consequently we recommend that users deviate from the default setting only when they realize the consequences of their choice.⁵ Note that Gronau, Ly, and Wagenmakers (2017) recently extended the Bayesian t -test to include prior distributions on effect size that are centered away from zero. We plan to add these “informed t -tests” to JASP in May 2017.

We are now ready to analyze the data in JASP. Readers who wish to confirm our results can open JASP, go to the File tab, Select “Open”, go to “Examples”, and select the “Kitchen Rolls” data set that is available at <https://osf.io/m6bi8/>. As shown in the left panel of Figure 3.5, the data feature one row for each participant. Each column corresponds to a variable; the dependent variable of interest here is in the column “mean NEO”, which contains the mean scores of each participant on the shortened 12-item version of the openness to experience subscale of the Neuroticism–Extraversion–Openness Personality Inventory (NEO PI-R; Costa & McCrae, 1992; Hoekstra, Ormel, & de Fruyt, 1996). The column “Rotation” includes the crucial information about group membership, with entries either “counter” or “clock”.

In order to conduct the analysis, selecting the “T-test” tab reveals the option “Bayesian Independent Samples T-test”, the dialog of which is displayed in the middle panel of Figure 3.5. We have selected “mean NEO” as the dependent variable, and “Rotation” as the grouping variable. After ticking the box “Descriptives”, the output displayed in the right panel of Figure 3.5 indicates that the mean openness-to-experience is slightly larger in the counterclockwise group (i.e., $N = 54$; $M = .71$) than in the clockwise group (i.e., $N = 48$;

⁵For an indication of how Bayes factors can be computed under any proper prior distribution see <http://jeffrouder.blogspot.nl/2016/01/what-priors-should-i-use-part-i.html>, also available as a pdf file at the OSF project page <https://osf.io/m6bi8/>.

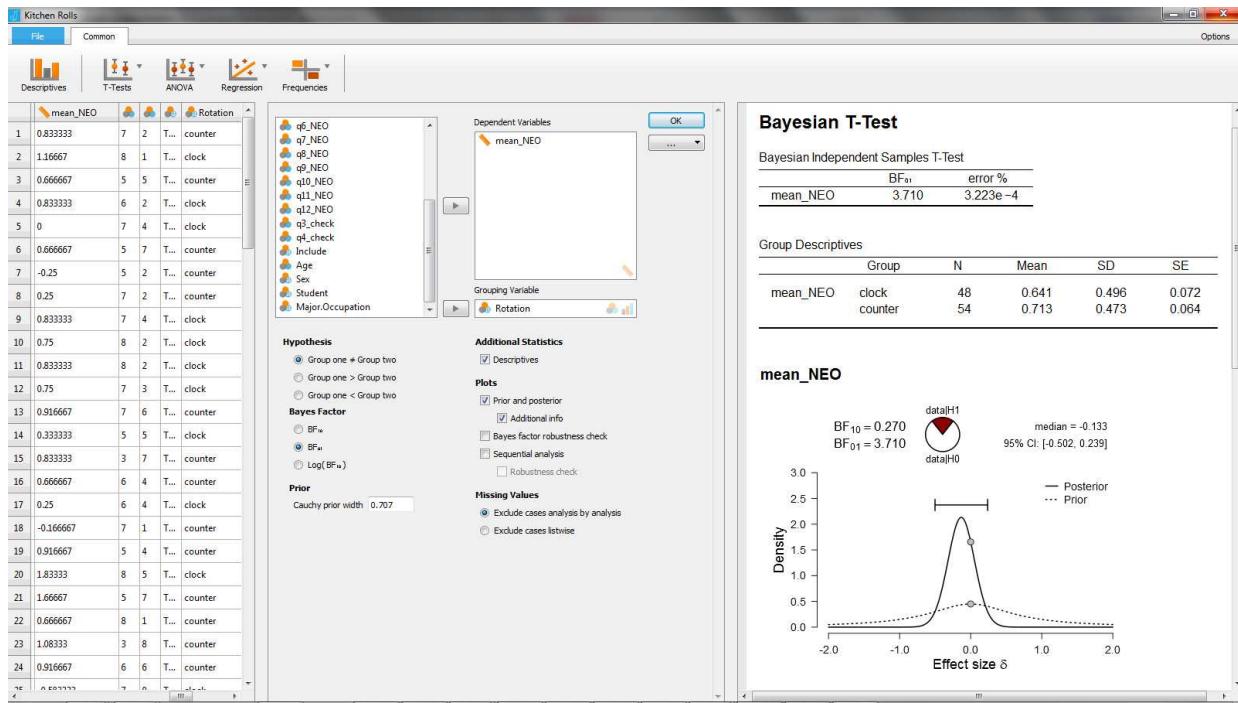


Figure 3.5: JASP screenshot for the two-sided test of the kitchen roll replication experiment (Wagenmakers et al., 2015). The left panel shows the data in spreadsheet format; the middle panel shows the analysis input options; the right panel shows the analysis output. NB. The “error %” indicates the size of the error in the integration routine relative to the Bayes factor, similar to a coefficient of variation.

$M = .64$) – note that the effect goes in the direction opposite to that hypothesized by Topolinski and Sparenberg (2012).

For demonstration purposes, at first we refrain from specifying the direction of the test. To contrast our results with those reported by Wagenmakers et al. (2015), we have set the Cauchy prior width to its JASP default $r = 0.707$ instead of Jeffreys’s value $r = 1$. We have also ticked the plotting options “Prior and posterior” and “Additional info”. This produces the plot shown in the right panel of Figure 3.5. It is evident that most of the posterior mass is negative. The posterior median is -0.13 , and a 95% credible interval ranges from -0.50 to 0.23 . The Bayes factor is 3.71 in favor of \mathcal{H}_0 over the two-sided \mathcal{H}_1 . This indicates that the observed data are 3.71 times more likely under \mathcal{H}_0 than under \mathcal{H}_1 . Because the Bayes factor favors \mathcal{H}_0 , in the input panel we have selected “ BF_{01} ” under “Bayes Factor” – it is easier to interpret $BF_{01} = 3.71$ than it is to interpret the mathematically equivalent statement $BF_{10} = 0.27$.

After this initial investigation we now turn to an analysis of the preregistered order-restricted test (with the exception of using $r = 0.707$ instead of the preregistered $r = 1$). The output of the “Descriptives” option has revealed that “clock” is group 1 (because it is on top), and “counter” is group 2. Hence, we can incorporate the order restriction in our inference by ticking the “Group one > Group two” box under “Hypothesis” in the input

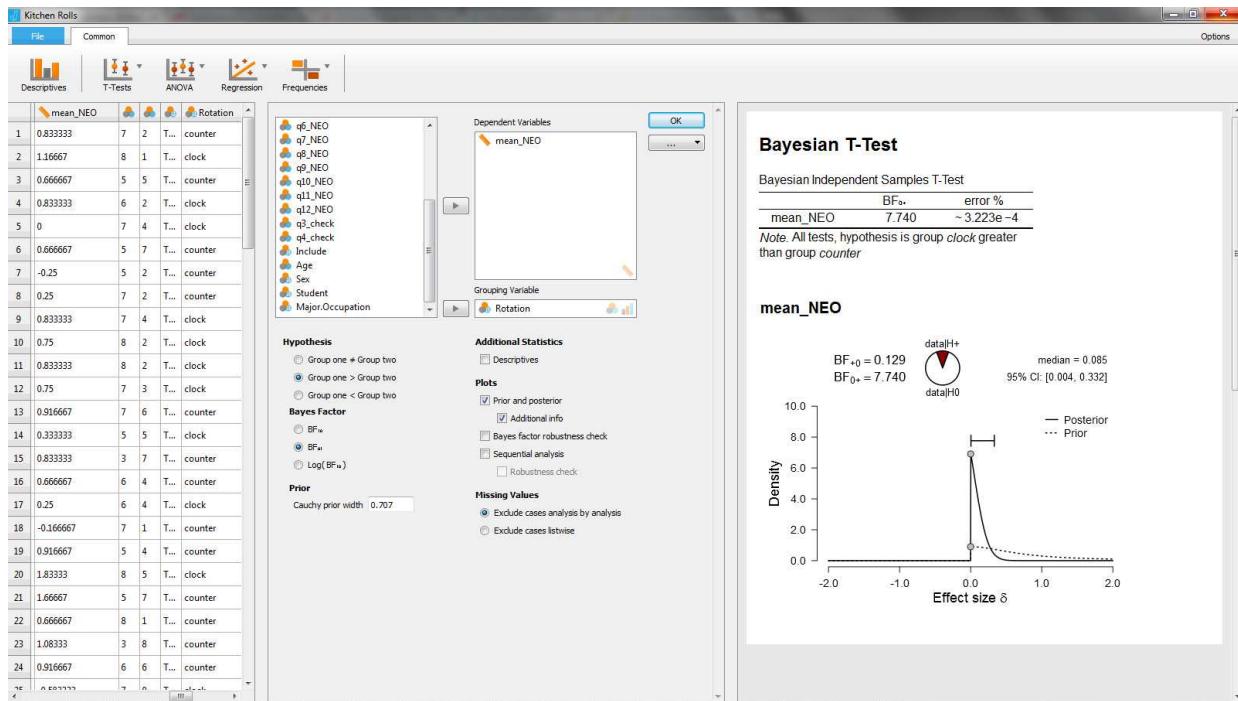


Figure 3.6: JASP screenshot for the one-sided test of the kitchen roll replication experiment (Wagenmakers et al., 2015). The left panel shows the data in spreadsheet format; the middle panel shows the analysis input options; the right panel shows the analysis output.

panel, as is shown in the middle panel of Figure 3.6.

The output for the order-restricted test is shown in the right panel of Figure 3.6. As expected, incorporating the knowledge that the observed effect is in the direction opposite to the one that was hypothesized increases the relative evidence in favor of \mathcal{H}_0 (see also Matzke et al., 2015). Specifically, the Bayes factor has risen from 3.71 to 7.74, meaning that the observed data are 7.74 times more likely under \mathcal{H}_0 than under \mathcal{H}_+ .

As an aside, note that under \mathcal{H}_+ the posterior distribution is concentrated near zero but does not have mass on negative values, in accordance with the order-restriction imposed by \mathcal{H}_+ . In contrast, the classical one-sided confidence interval ranges from $-.23$ to ∞ . This classical interval contrasts sharply with its Bayesian counterpart, and, even though the classical interval is mathematically well-defined (i.e., it contains all values that would not be rejected by a one-sided $\alpha = .05$ significance test, see also Wagenmakers et al., in press), we submit that most researchers will find the classical result neither intuitive nor informative.

Next we turn to a robustness analysis and quantify the evidential impact of the width r of the Cauchy prior distribution. The middle panel of Figure 3.7 shows that the option “Bayes factor robustness check” is ticked, and this produces the upper plot in the right panel of Figure 3.7. When the Cauchy prior with r equals zero, \mathcal{H}_1 is identical to \mathcal{H}_+ , and the Bayes factor equals 1. As the width r increases and \mathcal{H}_+ starts to predict that the effect is positive, the evidence in favor of \mathcal{H}_0 increases; for the JASP default value $r = .707$, the Bayes factor $BF_{0+} = 7.73$; for Jeffreys’s default $r = 1$, the Bayes factor $BF_{0+} = 10.75$; and

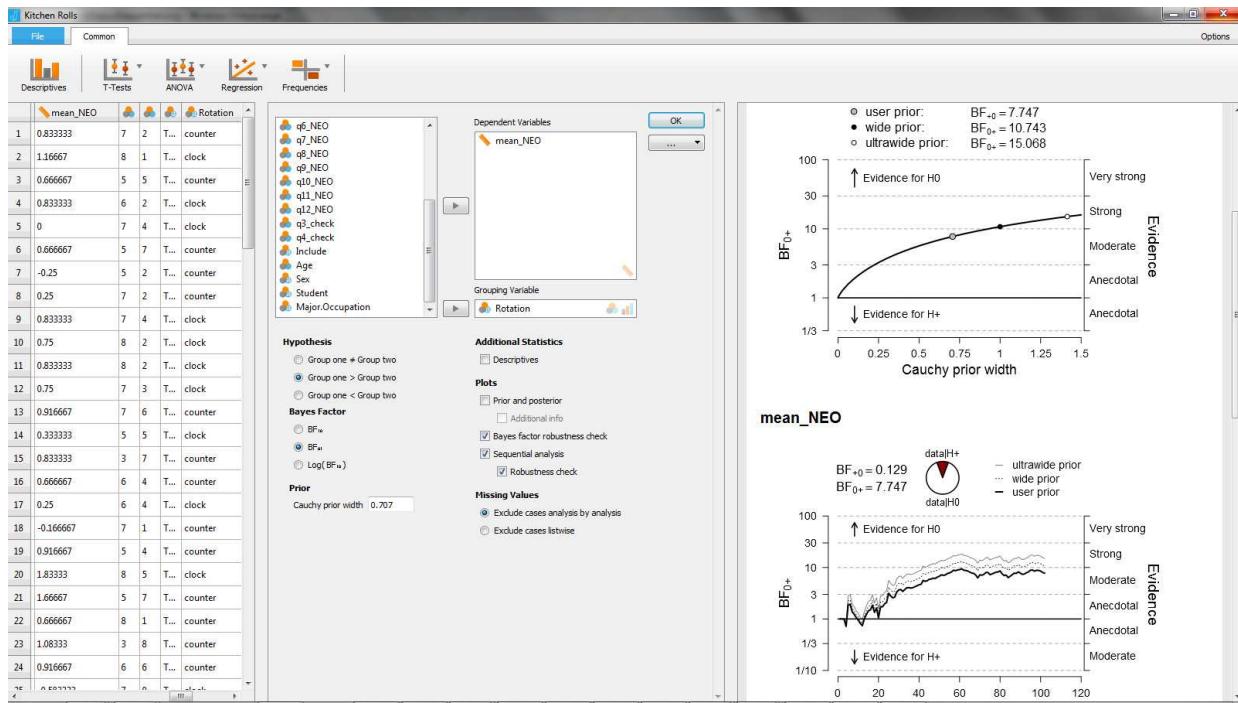


Figure 3.7: JASP screenshot for the one-sided test of the kitchen roll replication experiment (Wagenmakers et al., 2015). The right panel shows the analysis output: the upper plot is a robustness analysis, and the bottom plot is a sequential analysis combined with a robustness analysis.

for the “ultrawide” prior $r = \sqrt{2} \approx 1.41$, the Bayes factor $\text{BF}_{0+} = 15.04$. Thus, over a wide range of plausible values for the prior width r , the data provide moderate to strong evidence in favor of the null hypothesis \mathcal{H}_0 .

Finally, the middle panel of Figure 3.7 also shows that the options “Sequential analysis” and “robustness check” are ticked, and these together produce the lower plot in the right panel of Figure 3.7. The sequential analysis is of interest here because it was part of the experiment’s sampling plan, and because it underscores how researchers can monitor and visualize the evidential flow as the data accumulate. Closer examination of the plot reveals that for the preregistered value of $r = 1$, Wagenmakers et al. (2015) did not adhere to their preregistered sampling plan to stop data collection as soon as $\text{BF}_{0+} > 10$ or $\text{BF}_{+0} > 10$: after about 55 participants, the dotted line crosses the threshold of $\text{BF}_{0+} > 10$ but data collection nonetheless continued. Wagenmakers et al. (2015, p. 3) explain: “This occurred because data had to be entered into the analysis by hand and this made it more difficult to monitor the Bayes factor continually. In practice, the Bayes factor was checked every few days. Thus, we continued data collection until we reached our predetermined stopping criterion at the point of checking.”

One of the advantages of the sequential robustness plot is that it provides a visual impression of when the Bayes factors for the different priors have converged, in the sense that their difference on the log scale is constant (e.g., Gronau & Wagenmakers, in press). For the current situation, the convergence has occurred after testing approximately 35 partici-

pants. To understand why the difference between the log Bayes factors becomes constant after an initial number of observations, consider data y that consists of two batches, y_1 and y_2 . As mentioned above, from the law of conditional probability we have $\text{BF}_{0+}(y) = \text{BF}_{0+}(y_1) \times \text{BF}_{0+}(y_2 | y_1)$. Note that this expression highlights that Bayes factors for different batches of data (e.g., participants, experiments) may not be multiplied blindly; the second factor, $\text{BF}_{0+}(y_2 | y_1)$, equals the relative evidence from the second batch y_2 , after the prior distributions have been properly updated using the information extracted from the first batch y_1 (Jeffreys, 1961, p. 333). Rewriting the above expression on the log scale we obtain $\log \text{BF}_{0+}(y) = \log \text{BF}_{0+}(y_1) + \log \text{BF}_{0+}(y_2 | y_1)$. Now assume y_1 contains sufficient data such that, regardless of the value of prior width r under consideration, approximately the same posterior distribution is obtained. In most situations, this posterior convergence happens relatively quickly. This posterior distribution is then responsible for generating the Bayes factor for the second component, $\log \text{BF}_{0+}(y_2 | y_1)$, and it is therefore robust against differences in r .⁶ Thus, models with different values of r will make different predictions for data from the first batch y_1 . However, after observing a batch y_1 that is sufficiently large, the models have updated their prior distribution to a posterior distribution that is approximately similar; consequently, these models then start to make approximately similar predictions, resulting in a change in the log Bayes factor that is approximately similar as well.

In the first example we noted that the Bayes factor grades the evidence provided by the data on an unambiguous and continuous scale. Nevertheless, the sequential analysis plots in JASP make reference to discrete categories of evidential strength. These categories were inspired by Jeffreys (1961, Appendix B). Table 3.1 shows the classification scheme used by JASP. We replaced Jeffreys's labels "worth no more than a bare mention" with "anecdotal" (i.e., weak, inconclusive), "decisive" with "extreme", and "substantial" with "moderate" (Lee & Wagenmakers, 2013); the moderate range may be further subdivided by using "mild" for the 3-6 range and retaining "moderate" for the 6-10 range.⁷ These labels facilitate scientific communication but should be considered only as an approximate descriptive articulation of different standards of evidence. In particular, we may paraphrase Rosnow and Rosenthal (1989) and state that, surely, God loves the Bayes factor of 2.5 nearly as much as he loves the Bayes factor of 3.5.

⁶This also suggests that one can develop a Bayes factor that is robust against plausible changes in r : first, sacrifice data y_1 until the posterior distributions are similar; second, monitor and report the Bayes factor for the remaining data y_2 . This is reminiscent of the idea that underlies the so-called intrinsic Bayes factor (Berger & Pericchi, 1996), a method that also employs a "training sample" to update the prior distributions before the test is conducted using the remaining data points. The difference is that the intrinsic Bayes factor selects a training sample of minimum size, being just large enough to identify the model parameters.

⁷The present authors are not all agreed on the usefulness of such descriptive classifications of Bayes factors. All authors agree, however, that the advantage of Bayes factors is that –unlike for instance p values which are dichotomized into "significant" and "non-significant"– the numerical value of the Bayes factor can be interpreted directly. The strength of the evidence is not dependent on any conventional verbal description, such as "strong".

Table 3.1: A descriptive and approximate classification scheme for the interpretation of Bayes factors BF_{10} (Lee and Wagenmakers 2013; adjusted from Jeffreys 1961).

Bayes factor	Evidence category
> 100	Extreme evidence for \mathcal{H}_1
30 - 100	Very strong evidence for \mathcal{H}_1
10 - 30	Strong evidence for \mathcal{H}_1
3 - 10	Moderate evidence for \mathcal{H}_1
1 - 3	Anecdotal evidence for \mathcal{H}_1
1	No evidence
$1/3 - 1$	Anecdotal evidence for \mathcal{H}_0
$1/10 - 1/3$	Moderate evidence for \mathcal{H}_0
$1/30 - 1/10$	Strong evidence for \mathcal{H}_0
$1/100 - 1/30$	Very strong evidence for \mathcal{H}_0
$< 1/100$	Extreme evidence for \mathcal{H}_0

Example 3: “A Bayesian One-Way ANOVA to Test Whether Pain Threshold Depends on Hair Color”

An experiment conducted at the University of Melbourne in the 1970s suggested that pain threshold depends on hair color (McClave & Dietrich II, 1991, Exercise 10.20). In the experiment, a pain tolerance test was administered to 19 participants who had been divided into four groups according to hair color: light blond, dark blond, light brunette, and dark brunette.⁸ Figure 3.8 shows the boxplots and the jittered data points. There are visible differences between the conditions, but the sample sizes are small.

The data may be analyzed with a classical one-way ANOVA. This yields a p -value of .004, suggesting that the null hypothesis of no condition differences may be rejected. But how big is the evidence in favor of an effect? To answer this question we now analyze the data in JASP using the Bayesian ANOVA methodology proposed by Rouder et al. (2012) (see also Rouder, Morey, et al., in press). As was the case for the t -test, we assign Cauchy priors to effect sizes. What is new is that the Cauchy prior is now multivariate, and that effect size in the ANOVA model is defined in terms of distance to the grand mean.⁹ The analysis requires that the user opens the data file containing 19 pain tolerance scores in one column and 19 hair colors in the other column. As before, each row corresponds to a participant. The user then selects “ANOVA” from the ribbon, followed by “Bayesian ANOVA”. In the associated analysis menu, the user drags the variable “Pain Tolerance” to the input field

⁸The data are available at <http://www.statsci.org/data/oz/blonds.html>.

⁹The Cauchy prior width r_t for the independent samples t -tests yields the same result as a two-group one-way ANOVA with a fixed effect scale factor r_A equal to $r_t/\sqrt{2}$. With the default setting $r_t = 1/2 \cdot \sqrt{2}$, this produces $r_A = 0.5$. In sum, for the default prior settings in JASP the independent samples t -test and the two-group one-way ANOVA yield the same result. For examples see <https://cran.r-project.org/web/packages/BayesFactor/vignettes/priors.html>.

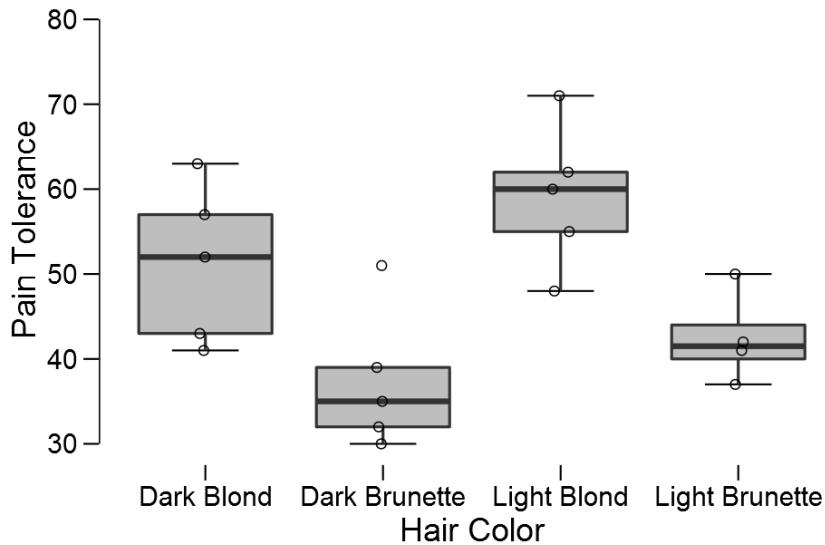


Figure 3.8: Boxplots and jittered data points for the hair color experiment. Figure created with JASP.

labeled “Dependent Variable” and drags the variable “Hair Color” to the input field “Fixed Factors”. The resulting output table with Bayesian results is shown in Figure 3.9.

The first column of the output table, “Models”, lists the models under consideration. The one-way ANOVA features only two models: the “Null model” that contains the grand mean, and the “Hair Color” model that adds an effect of hair color. The next point of interest is the “ BF_{10} ” column; this column shows the Bayes factor for each row-model against the null model. The first entry is always 1 because the null model is compared against itself. The second entry is 11.97, which means that the model with hair color predicts the observed data almost 12 times as well as the null model. As was the case for the output of the t -test, the right-most column, “% error”, indicates the size of the error in the integration routine relative to the Bayes factor; similar to a coefficient of variation, this means that small variability is more important when the Bayes factor is ambiguous than when it is extreme.

Column “ $P(M)$ ” indicates prior model probabilities (which the current version of JASP sets to be equal across all models at hand); column “ $P(M|data)$ ” indicates the updated probabilities after having observed the data. Column “ BF_M ” indicates the degree to which the data have changed the prior model odds. Here the prior model odds equals 1 (i.e., 0.5/0.5) and the posterior model odds equals almost 12 (i.e., 0.923/0.077). Hence, the Bayes factor equals the posterior odds. JASP offers the user “Advanced Options” that can be used to change the prior width of the Cauchy prior for the model parameters. As the name suggests, we recommend that the user exercises this freedom only in the presence of substantial knowledge of the underlying statistical framework.

Currently JASP does not offer post-hoc tests to examine pairwise differences in one-way ANOVA. Such post-hoc tests have not yet been developed in the Bayesian ANOVA framework. In future work we will examine whether post-hoc tests can be constructed by applying a Bayesian correction for multiple comparisons (i.e., Scott & Berger, 2006, 2010;

Bayesian ANOVA ▾

Model Comparison - Pain Tolerance ▾

Models	P(M)	P(M data)	BFM	BF10	% error
Null model	0.500	0.077	0.084	1.000	
Hair Color	0.500	0.923 <small>oe kig</small>	11.969	11.969	0.004

The data are about 12 times more likely under the model that includes hair color.

Figure 3.9: JASP output table for the Bayesian ANOVA of the hair color experiment. The blue text underneath the table shows the annotation functionality that can help communicate the outcome of a statistical analysis.

Stephens & Balding, 2009). Discussion of this topic would take us too far afield.

Example 4: “A Bayesian Two-Way ANOVA for Singers’ Height as a Function of Gender and Pitch”

The next data set concerns the heights in inches of the 235 singers in the New York Choral Society in 1979 (Chambers, Cleveland, Kleiner, & Tukey, 1983).¹⁰ The singers’ voices were classified according to voice part (e.g., soprano, alto, tenor, bass) and recoded to voice pitch (i.e., very low, low, high, very high). Figure 3.10 shows the relation between pitch and height separately for men and women.

Our analysis concerns the extent to which the dependent variable “height” is associated with gender (i.e., male, female) and/or pitch. This question can be examined statistically using a 2×4 ANOVA. Consistent with the visual impression from Figure 3.10, a classical analysis yields significant results for both main factors (i.e., $p < .001$ for both gender and pitch) but fails to yield a significant result for the interaction (i.e., $p = .52$). In order to assess the extent to which the data support the presence and absence of these effects we now turn to a Bayesian analysis.

In order to conduct this analysis in JASP, the user first opens the data set and then navigates to the “Bayesian ANOVA” input panel as was done for the one-way ANOVA. In the associated analysis menu, the user then drags the variable “Height” to the input field labeled “Dependent Variable” and drags the variables “Gender” and “Pitch” to the input field “Fixed Factors”. The resulting output table with Bayesian results is shown in Figure 3.11.

¹⁰Data available at <https://stat.ethz.ch/R-manual/R-devel/library/lattice/html/singer.html>.

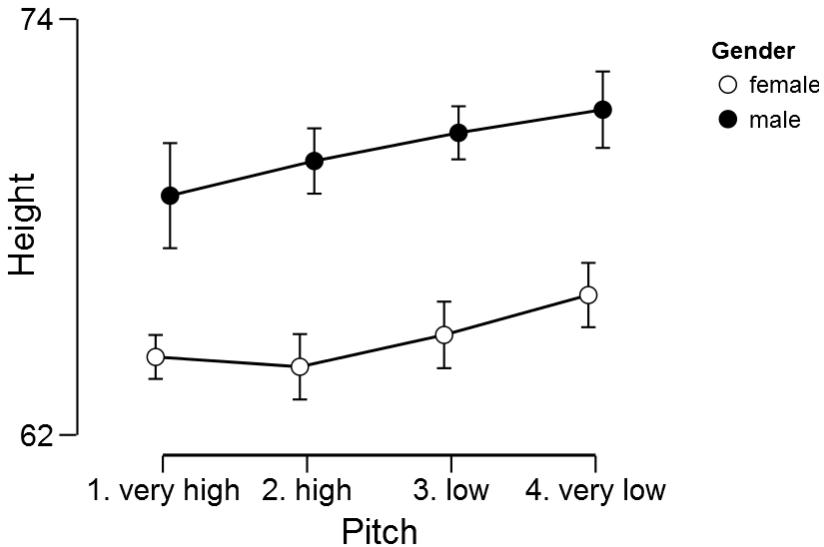


Figure 3.10: Relation between voice pitch, gender, and height (in inches) for data from 235 singers in the New York Choral Society in 1979. Error bars show 95% confidence intervals. Figure created with JASP.

The first column of the output table, “Models”, lists the five models under consideration: the “Null model” that contains only the grand mean, the “Gender” model that contains the effect of gender, the “Pitch” model that contains the effect of Pitch, the “Gender + Pitch” model that contains both main effects, and finally the “Gender + Pitch + Gender \times Pitch” model that includes both main effects and the interaction. Consistent with the principle of marginality, JASP does not include interactions in the absence of the component main effects; for instance, the interaction-only model “Gender \times Pitch” may not be entertained without also adding the two main effects (for details, examples, and rationale see Bernhardt & Jung, 1979; Grieppentrog, Ryan, & Smith, 1982; McCullagh & Nelder, 1989; Nelder, 1998, 2000; Peixoto, 1987, 1990; Rouder, Morey, et al., in press; Rouder, Engelhardt, et al., in press; Venables, 2000).

Now consider the BF_{10} column. All models (except perhaps for Pitch) receive overwhelming evidence in comparison to the Null model. The model that outperforms the Null model the most is the two main effects model, Gender + Pitch. Adding the interaction makes the model less competitive. The evidence against including the interaction is roughly a factor of ten. This can be obtained as $8.192\text{e+}39 / 8.864\text{e+}38 \approx 9.24$. Thus, the data are 9.24 times more likely under the two main effects model than under the model that adds the interaction.

Column “P(M)” indicates the equal assignment of prior model probability across the five models; column “P(M|data)” indicates the posterior model probabilities. Almost all posterior mass is centered on the two main effects model and the model that also includes the interaction. Column “ BF_M ” indicates the change from prior to posterior model odds. Only the two main effects model has received support from the data in the sense that the data have increased its model probability.

Bayesian ANOVA

Model Comparison - Height

Models	P(M)	P(M data)	BFM	BF10	% error
Null model	0.200	1.097e -40	4.388e -40	1.000	
Gender	0.200	0.004	0.017	3.807e +37	1.392e -44
Pitch	0.200	8.911e -39	3.564e -38	81.238	0.003
Gender + Pitch	0.200	0.899	35.447	8.192e +39	1.821
Gender + Pitch + Gender * Pitch	0.200	0.097	0.431	8.864e +38	2.441

Figure 3.11: JASP output table for the Bayesian ANOVA of the singers data. Note that JASP uses exponential notation to represent large numbers; for instance, “3.807e +37” represents 3.807×10^{37} .

The screenshot shows the JASP Bayesian ANOVA interface. On the left, the 'Model' panel displays 'Components' (Gender, Pitch) and 'Model Terms' (Gender, Pitch, Gender * Pitch). Under 'Is Nuisance', 'Gender' and 'Pitch' are checked, while 'Gender * Pitch' is not. On the right, the 'Bayesian ANOVA ▾' panel shows the 'Model Comparison - Height ▾' table. The table includes columns for Models, P(M), P(M|data), BFM, BF10, and % error. The table data is as follows:

Models	P(M)	P(M data)	BFM	BF10	% error
Null model (incl. Gender, Pitch)	0.500	0.902	9.251	1.000	
Gender * Pitch	0.500	0.098	0.108	0.108	2.849

Note: All models include Gender, Pitch.

Figure 3.12: JASP screenshot and output table for the Bayesian ANOVA of the singers data, with Gender and Pitch added as nuisance factors.

Above we wished to obtain the Bayes factor for the main effects only model versus the model that adds the interaction. We accomplished this objective by comparing the strength of the Bayes factor against the Null model for models that exclude or include the critical interaction term. However, this Bayes factor can also be obtained directly. As shown in Figure 3.12, the JASP interface allows the user to specify Gender and Pitch as nuisance variables, which means that they are included in every model, including the Null model. The Bayes factor of interest is $BF_{10} = 0.108$; when inverted, this yields $BF_{01} = 1/0.108 = 9.26$, confirming the result obtained above through a simple calculation. The fact that the numbers are not identical is due to the numerical approximation; the error percentage is indicated in the right-most column.

In sum, the Bayesian ANOVA reveals that the data provide strong support for the two main effects model over any of the simpler models. The data also provide good support against including the interaction term.

Finally, as described in Cramer et al. (2016), the multiway ANOVA harbors a multiple comparison problem. As for the one-way ANOVA, this problem can be addressed by applying the proper Bayesian correction method (i.e., Scott & Berger, 2006, 2010; Stephens & Balding, 2009). This correction has not yet been implemented in JASP.

			
Assassin Bug	Cicada Killer	Diving Beetle <small>James L. Castner, U. Fla. Ent. Dept.</small>	Scorpion
			
June Beetle	Wasp	Cockroach	Ant Lion
Low Disgusting & Low Frightening	Low Disgusting & High Frightening	High Disgusting & Low Frightening	High Disgusting & High Frightening

Figure 3.13: The arthropod stimuli used in Ryan et al. (2013). Each cell in the 2×2 repeated measures design contains two arthropods. The original stimuli did not show the arthropod names. Figure adjusted from Ryan et al. (2013).

Example 5: “A Bayesian Two-Way Repeated Measures ANOVA for People’s Hostility Towards Arthropods”

In an online experiment, Ryan et al. (2013) presented over 1300 participants with pictures of eight arthropods. For each arthropod, participants were asked to rate their hostility towards that arthropod, that is, “...the extent to which they either wanted to kill, or at least in some way get rid of, that particular insect” (p. 1297). The arthropods were selected to vary along two dimensions with two levels: disgustingness (i.e., low disgusting and high disgusting) and frighteningness (i.e., low frighteningness and high frighteningness). Figure 3.13 shows the arthropods and the associated experimental conditions. For educational purposes, we ignore the gender factor, we ignore the fact that the ratings are not at all normally distributed, we analyze data from a subset of 93 participants, and we side-step the nontrivial question of whether to model the item-effects. The pertinent model is a linear mixed model, and the only difference with respect to the previous example is that we now require a prior for

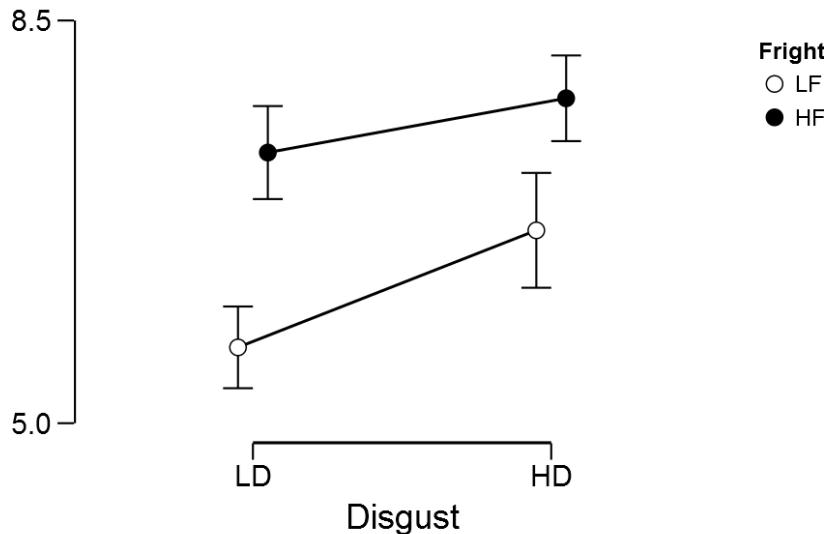


Figure 3.14: Hostility ratings for arthropods that differ in disgustingness (i.e., LD for low disgusting and HD for high disgusting) and frighteningness (i.e., LF for low frighteningness and HF for high frighteningness). Error bars show 95% confidence intervals. Data kindly provided by Ryan et al. (2013). Figure created with JASP.

the new random factor –in this case, participants– which is set a little wider because we assume a priori that participants are variable in the main effect (for an in-depth discussion see Rouder, Morey, et al., *in press*).

Our analysis asks whether and how people’s hostility towards arthropods depends on their disgustingness and frighteningness. As each participant’s rated all eight arthropods, these data can be analyzed using a repeated measures 2×2 ANOVA. A classical analysis reveals that the main effects of disgustingness and frighteningness are both highly significant (i.e., p ’s $< .001$) whereas the interaction is not significant ($p = 0.146$). This is consistent with the data as summarized in Figure 3.14: arthropods appear to be particularly unpopular when they are high rather than low in disgustingness, and when they are high rather than low in frighteningness. The data do not show a compelling interaction. To assess the evidence for and against the presence of these effects we now turn to a Bayesian analysis.

To conduct the Bayesian analysis the user first needs to open the data set in JASP.¹¹ Next the user selects the “Bayesian Repeated Measures ANOVA” input panel that is nested under the ribbon option “ANOVA”. Next the user needs to name the factors (here “Disgust” and “Fright”) and their levels (here “LD”, “HD”, and “LF”, “HF”). Finally the input variables need to be dragged to the matching “Repeated Measures Cells”.

The analysis produces the output shown in the top panel of Figure 3.15. As before, the column “Models” lists the five different models under consideration. The BF_{10} column shows that compared to the Null model, all other models (except perhaps the Disgust-only model) receive overwhelming support from the data. The model that receives the most

¹¹The data set is available on the project OSF page and from within JASP (i.e., File → Open → Examples → Bugs).

Bayesian Repeated Measures ANOVA

Model Comparison - dependent

Models	P(M)	P(M data)	BF _M	BF ₁₀	% error
Null model (incl. subject)	0.200	2.198e-10	8.793e-10	1.000	
Disgust	0.200	4.497e-9	1.799e-8	20.458	0.928
Fright	0.200	0.014	0.057	6.408e+7	0.946
Disgust + Fright	0.200	0.712	9.902	3.240e+9	1.895
Disgust + Fright + Disgust * Fright	0.200	0.274	1.507	1.245e+9	3.431

Note. All models include subject.

Analysis of Effects - dependent

Effects	P(incl)	P(incl data)	BF _{Inclusion}
Disgust	0.600	0.986	46.659
Fright	0.600	1.000	1.413e+8
Disgust * Fright	0.200	0.274	1.507

Figure 3.15: JASP screenshot for the output tables of the Bayesian ANOVA for the arthropod experiment. The top table shows the model-based analysis, whereas the bottom panels shows the analysis of effects, averaging across the models that contain a specific factor. See text for details.

support against the Null model is the two main effects model, Disgust + Fright. Adding the interaction decreases the degree of this support by a factor of $3.240/1.245 = 2.6$. This is the Bayes factor in favor of the two main effects model versus the model that also includes the interaction. The same result could have been obtained directly by adding “Disgust” and “Fright” as nuisance variables, as was illustrated in the previous example.

The “P(M)” column shows the uniform distribution of prior model probabilities across the five candidate models, and the “P(M|data)” column shows the posterior model probabilities. Finally, the “BF_M” column shows the change from prior model odds to posterior model odds. This Bayes factor also favors the two main effects model, but at the same time indicates mild support in favor of the interaction model. The reason for this discrepancy (i.e., a Bayes factor of 2.6 against the interaction model versus a Bayes factor of 1.5 in favor of the interaction model) is that these Bayes factors address different questions: The Bayes factor of 2.6 compares the interaction model against the two main effects model (which happens to be the model that is most supported by the data), whereas the Bayes factor of 1.5 compares the interaction model against all candidate models, some of which receive almost no support from the data. Both analyses are potentially of interest. Specifically, when the two main effects model decisively outperforms the simpler candidate models then it may be appropriate to assess the importance of the interaction term by comparing the two main effects model

against the model that adds the interaction. However, it may happen that the simpler candidate models outperform the two main effects model – in other words, the two main effects model has predicted the data relatively poorly compared to the Null model or one of the single main effects models. In such situations it is misleading to test the importance of the interaction term by solely focusing on a comparison to the poorly performing two main effects model. In general we recommend radical transparency in statistical analysis; an informative report may present the entire table shown in Figure 3.15. In this particular case, both Bayes factors (i.e., 2.6 against the interaction model, and 1.5 in favor of the interaction model) are “not worth more than a bare mention” (Jeffreys, 1961, Appendix B); moreover, God loves these Bayes factors almost an equal amount, so it may well be argued that the discrepancy here is more apparent than real.

As the number of factors grows, so does the number of models. With many candidate models in play, it may be risky to base conclusions on a comparison involving a small subset. In Bayesian model averaging (BMA; e.g., Etz & Wagenmakers, *in press*; Haldane, 1932; Hoeting, Madigan, Raftery, & Volinsky, 1999) the goal is to retain model selection uncertainty by averaging the conclusions from each candidate model, weighted by that model’s posterior plausibility. In JASP this is accomplished by ticking the “Effects” input box, which results in an output table shown in the bottom panel of Figure 3.15.

In our example, the averaging in BMA occurs over the models shown in the Model Comparison table (top panel of Figure 3.15). For instance, the factor “Disgust” features in three models (i.e., Disgust only, Disgust + Fright, and Disgust + Fright + Disgust * Fright). Each model has a prior model probability of 0.2, so the summed prior probability of the three models that include disgust equals 0.6; this is known as the prior inclusion probability for Disgust (i.e., the column P(incl)). After the data are observed we can similarly consider the sum of the posterior model probabilities for the models that include disgust, yielding $4.497e-9 + 0.712 + 0.274 = 0.986$. This is the posterior inclusion probability (i.e., column P(incl|data)). The change from prior to posterior inclusion odds is given in the column “BF_{Inclusion}”. Averaged across all candidate models, the data strongly support inclusion of both main factors Disgust and Fright. The interaction only receives weak support. In fact, the interaction term occurs only in a single model, and therefore its posterior inclusion probability equals the posterior model probability of that model (i.e., the one that contains the two main effects and the interaction).

It should be acknowledged that the analysis of repeated measures ANOVA comes with a number of challenges and caveats. The development of Bayes factors for crossed-random effect structures is still a topic of ongoing research. And in general, JASP currently does not feature an extensive suite of estimation routines to assess the extent to which generic model assumptions (e.g., sphericity) are violated.

Future Directions for Bayesian Analyses in JASP

The present examples provides a selective overview of default Bayesian inference in the case of the correlation test, *t*-test, one-way ANOVA, two-way ANOVA, and two-way repeated measures ANOVA. In JASP, other analyses can be executed in similar fashion (e.g., for

contingency tables, Jamil, Ly, et al., in press; Jamil, Marsman, et al., in press; Scheibehenne, Jamil, & Wagenmakers, in press; or for linear regression Rouder & Morey, 2012). A detailed discussion of the entire functionality of JASP is beyond the scope of this article.

In the near future, we aim to expand the Bayesian repertoire of JASP, both in terms of depth and breadth. In terms of depth, our goal is to provide more and better graphing options, more assumption tests, more nonparametric tests, post-hoc tests, and corrections for multiplicity. In terms of breadth, our goal is to include modules that offer the functionality of the BAS package (i.e., Bayesian model averaging in regression, Clyde, 2016), the informative model comparison approach (e.g., Gu, Mulder, Decović, & Hoijtink, 2014; Gu, 2016; Mulder, 2014, 2016), and a more flexible and subjective prior specification approach (e.g., Dienes, 2011, 2014, 2016; Gronau et al., 2017). By making the additional functionality available as add-on modules, beginning users are shielded from the added complexity that such options add to the interface. In the short-term we also aim to develop educational materials that make JASP output easier to interpret and to teach to undergraduate students. This entails writing a JASP manual, developing course materials, writing course books, and designing a Massive Open Online Course.

Our long-term goal is for JASP to facilitate several aspects of statistical practice. Free and user-friendly, JASP has the potential to benefit both education and research. By featuring both classical and Bayesian analyses, JASP implicitly advocates a more inclusive statistical approach. JASP also aims to assist with data preparation and aggregation; currently, this requires that JASP launches and interacts with an external editor (see our data-editing video at <https://www.youtube.com/watch?v=1dT-iAU9Zuc&t=70s>); in the future, JASP will have its own editing functionality including filtering and outlier exclusion. Finally, by offering the ability to save, annotate, and share statistical output, JASP promotes a transparent way of communicating one's statistical results. An increase in statistical transparency and inclusiveness will result in science that is more reliable and more replicable.

As far as the continued development of JASP is concerned, our two main software developers and several core team members of the JASP team have tenured positions. The Psychological Methods Group at the University of Amsterdam is dedicated to long-term support for JASP, and in 2017 we have received four million euro to set up projects that include the development of JASP as a key component. The JASP code is open-source and will always remain freely available online. In sum, JASP is here to stay.

Concluding Comments

In order to promote the adoption of Bayesian procedures in psychology, we have developed JASP, a free and open-source statistical software program with an interface familiar to users of SPSS. Using JASP, researchers can obtain results from Bayesian techniques easily and without tears. Dennis Lindley once said that "Inside every Non-Bayesian, there is a Bayesian struggling to get out" (Jaynes, 2003). We hope that software programs such as JASP will act to strengthen the resolve of one's inner Bayesian and pave the road for a psychological science in which innovative hypotheses are tested using coherent statistics.

References

- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2, 317–352.
- Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91, 109–122.
- Bernhardt, I., & Jung, B. S. (1979). The interpretation of least squares regression with interaction or polynomial terms. *The Review of Economics and Statistics*, 61, 481–483.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. New York: Chapman & Hall.
- Clyde, M. (2016). *BAS: Bayesian adaptive sampling for Bayesian model averaging*. (R package version 1.4.1)
- Costa, P. T., & McCrae, R. R. (1992). *NEO Personality Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., et al. (2016). Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies. *Psychonomic Bulletin & Review*, 23, 640–647.
- 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. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781.
- Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78–89.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Etz, A., & Wagenmakers, E.-J. (in press). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*.
- Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, 57, 153–169.
- Griepentrog, G. L., Ryan, J. M., & Smith, L. D. (1982). Linear transformations of polynomial regression models. *The American Statistician*, 36, 171–174.
- Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian *t*-tests. *Manuscript submitted for publication*.
- Gronau, Q. F., & Wagenmakers, E.-J. (in press). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*.
- Gu, X. (2016). *Bayesian evaluation of informative hypotheses*. Unpublished doctoral dissertation, Utrecht University.
- Gu, X., Mulder, J., Decović, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality

- constrained hypotheses. *Psychological Methods*, 19, 511–527.
- Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, 61, 545–557.
- Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 55–61.
- Hoekstra, H. A., Ormel, J., & de Fruyt, F. (1996). *Handleiding bij de NEO persoonlijkheids vragenlijsten NEO-PIR, NEO-FFI [manual for the NEO personality inventories NEO-PI-R and NEO-FFI]*. Lisse, the Netherlands: Swets & Zeitlinger.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.
- Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (in press). Default “Gunel and Dickey” Bayes factors for contingency tables. *Behavior Research Methods*.
- Jamil, T., Marsman, M., Ly, A., Morey, R. D., & Wagenmakers, E.-J. (in press). What are the odds? Modern relevance and Bayes factor solutions for MacAlister’s problem from the 1881 *Educational Times*. *Educational and Psychological Measurement*.
- JASP Team. (2017). *JASP (Version 0.8.1)[Computer software]*.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
- Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19313–19317.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.
- Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute Monographs*, 25, 149–163.
- Ly, A., Marsman, M., & Wagenmakers, E.-J. (in press). Analytic posteriors for Pearson’s correlation coefficient. *Statistica Neerlandica*.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72, 43–55.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- 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.
- McClave, J. T., & Dietrich II, F. H. (1991). *Statistics*. San Francisco: Dellen Publishing.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11-1*. Comprehensive R Archive Network.
- Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, 71, 448–463.
- Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, 72, 104–115.
- Nelder, J. A. (1998). The selection of terms in response-surface models—how strong is the weak-heredity principle? *The American Statistician*, 52, 315–318.
- Nelder, J. A. (2000). Functional marginality and response-surface fitting. *Journal of Applied Statistics*, 27, 109–112.
- O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.)*. London: Arnold.
- Overstall, A. M., & King, R. (2014a). conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58, 1–27.
- Overstall, A. M., & King, R. (2014b). A default prior distribution for contingency tables with dependent factor levels. *Statistical Methodology*, 16, 90–99.
- Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41, 311–313.
- Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, 44, 26–30.
- R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-00-3)
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–1284.
- Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (in press). Model comparison in ANOVA. *Psychonomic Bulletin & Review*.
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877–903.
- 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., Morey, R. D., Verhagen, A. J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs. *Psychological Methods*.
- 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.
- Ryan, R. S., Wilde, M., & Crist, S. (2013). Compared to a small, supervised lab experiment, a large, unsupervised web-based experiment on a previously unknown effect has benefits that outweigh its potential costs. *Computers in Human Behavior*, 29, 1295–1301.
- Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (in press). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*.
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (in press). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.

- Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136, 2144–2162.
- Scott, J. G., & Berger, J. O. (2010). Bayes and empirical–Bayes multiplicity adjustment in the variable–selection problem. *The Annals of Statistics*, 38, 2587–2619.
- Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10, 681–690.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, 24, 159–171.
- Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science*, 3, 308–314.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76, 31–48.
- van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (in press). Bayesian inference for Kendall’s rank correlation coefficient. *The American Statistician*.
- Venables, W. N. (2000). *Exegeses on linear models*. Paper presented to the S-PLUS User’s Conference.
- Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., et al. (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., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., et al. (in press). Bayesian statistical inference for psychological science. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855t tests. *Perspectives on Psychological Science*, 6, 291–298.
- 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.
- Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press.

Appendix: Visualizing the Strength of Evidence



A dart board analogy to intuit the strength of evidence that a Bayes factor provides. Figure available at <https://osf.io/m6bi8/> under a CC-BY license.

Parameter estimation in nonstandard models

Dora Matzke, Udo Boehm, and Joachim Vandekerckhove

Introduction

In this special issue, Dienes (this issue) has argued that Bayesian methods are to be preferred over classical methods, Kruschke (this issue) and Etz and Vandekerckhove (this issue) have introduced the Bayesian philosophy and associated mathematics, and Love et al. (this issue; see also Love et al., 2015) and Wagenmakers et al. (this issue) described software that implements standard hypothesis tests within the Bayesian framework. In the present paper, we demonstrate the use of three popular software packages that enable psychologists to estimate parameters in formal models of varying complexity.

The mathematical foundations of Bayesian parameter estimation are not especially difficult—all that is involved are the elementary laws of probability theory to determine the posterior distribution of parameters given the data. Once the posterior distribution has been defined, the final hurdle of Bayesian parameter estimation is to compute descriptive statistics on the posterior. In order to obtain these descriptive statistics, one widely applicable strategy is to draw random samples from the posterior distribution using Markov chain Monte Carlo methods (MCMC; van Ravenzwaaij, this issue)—with sufficient posterior samples, descriptives on the sample set can substitute for actual quantities of interest.

In this article, we describe the use of three popular, general-purpose MCMC engines that facilitate the sampling process. We will focus on WinBUGS, JAGS, and Stan, and illustrate their use for parameter estimation in two popular models in psychology. The development of these software packages has greatly contributed to the increase in the prevalence of Bayesian methods in psychology over the past decade (e.g., Lee & Wagenmakers, 2013). The packages owe their popularity to their flexibility and usability; they allow researchers to build a large number of models of varying complexity using a relatively small set of sampling statements and deterministic transformations. Moreover, the packages have a smooth learning curve, are well documented, and are supported by a large community of users both within and outside of psychology. Their popularity notwithstanding, WinBUGS, JAGS, and Stan represent only a subclass of the many avenues to Bayesian analysis; the different avenues implement

a trade-off between flexibility and accessibility. At one end of the spectrum, researchers may use off-the-shelf Bayesian software packages, such as JASP (Love et al. this issue; see also Love et al., 2015). JASP has an attractive and user-friendly graphical user interface, but presently it only supports standard hypothesis tests (see also Morey, Rouder, & Jamil, 2015). At the other end of the spectrum, researcher may implement their own MCMC sampler, one that is tailored to the peculiarities of the particular model at hand (e.g., van Ravenzwaaij, this issue; Rouder & Lu, 2005). This approach provides tremendous flexibility, but it is time-consuming, labor-intensive, and requires expertise in computational methods. General-purpose MCMC engines—such as WinBUGS, JAGS, and Stan—are the middle-of-the-road alternatives to Bayesian analysis that provide a large degree of flexibility at a relatively low cost.

We begin with a short introduction of formal models as generative processes using a simple linear regression as an example. We then show how this model can be implemented in WinBUGS, JAGS, and Stan, with special emphasis on how the packages can be interacted with from R and MATLAB. We then turn to a more complex model, and illustrate the basic steps of Bayesian parameter estimation in a multinomial processing tree model for a false-memory paradigm. The WinBUGS, JAGS, and Stan code for all our examples is available in the Supplemental Materials at <https://osf.io/ucmaz/>. The discussion presents a comparison of the strengths and weaknesses of the packages and provides useful references to hierarchical extensions and Bayesian model selection methods using general-purpose MCMC software.

An Introduction with Linear Regression

Specification of Models as Generative Processes

Before we continue, it is useful to consider briefly what we mean by a formal *model*: A formal model is a set of formal statements about how the data come about. Research data are the realizations of some stochastic process, and as such they are draws from some random number generator whose properties are unknown. In psychology, the random number generator is typically a group of randomly selected humans who participate in a study, and the properties of interest are often differences in group means between conditions or populations (say, the difference in impulsivity between schizophrenia patients and controls) or other invariances and systematic properties of the data generation process. A formal model is an attempt to emulate the unknown random number generator in terms of a network of basic distributions.

Consider, for example, simple linear regression, with its three basic assumptions of *normality*, *linearity*, and *homoskedasticity*. This common technique implies a stochastic process: the data are assumed to be random draws from a normal distribution (normality), whose mean is a linear function of a predictor (linearity), and whose variance is the same (homoskedasticity) for all units, where “units” can refer to participants, items, conditions, and so on. A regression model in which we predict y from x may be written as follows:

$$y_i | \mu_i, \tau \sim N((\mu)_i, \tau) \quad (4.1)$$

$$\mu_i | \beta_1, \beta_2, x_i = \beta_1 + \beta_2 x_i. \quad (4.2)$$

The tilde (\sim) may be read as “is a random sample from”. These two statements encode the assumptions of normality (Eq. 4.1), homoskedasticity across units i (Eq. 4.1), and linearity (Eq. 4.2). Usually omitted, but implied, is that these statements hold true for all values that the subscript i can take:

$$\forall i, \quad i = 1, \dots, N. \quad (4.3)$$

We use τ to indicate the *precision*—the inverse of the variance—because that is how WinBUGS and JAGS parameterize the Gaussian distribution.

In the Bayesian framework, we must further specify our prior assumptions regarding the model parameters β_1 , β_2 , and τ . Let us use the following¹ forms for the priors:

$$\beta_1 \sim N((0), 0.001) \quad (4.4)$$

$$\beta_2 \sim N((0), 0.001) \quad (4.5)$$

$$\tau \sim \Gamma(0.001, 0.001). \quad (4.6)$$

This simple model also helps to introduce the types of variables that we have at our disposal. Variables can be *stochastic*, meaning that they are draws from some distribution. Stochastic variables can be either observed (i.e., data) or unobserved (i.e., unknown parameters). In this model, y , β_1 , β_2 , and τ are stochastic variables. Variables can also be *deterministic*, which means their values are completely determined by other variables. Here, μ_i is determined as some combination of β_1 , β_2 , and x_i . N is a constant.

Taken together, a Bayesian model can be thought of as a data-generation mechanism that is conditional on parameters: Bayesian models make *predictions*. In particular, the *sampling statements*—including the priors—in Equations 4.1, 4.4, 4.5, and 4.6 and the deterministic transformation in Equation 4.2, fully define a *generative model*; this set of statements fully defines the model because they are all that is needed to generate data from the model. The generative model thus formalizes the presumed process by which the data in an empirical study were generated.

A Toy Data Set

As our introductory example, we will use a small data set containing (a) the observed number of attendees at each session of a recent conference (the data y) and (b) the number of attendees that was expected by the organizers (the predictor x). Table 4.1 shows the data set.

Implementing a Generative Model

The generative specification is the core of the BUGS modeling language (Lunn, Thomas, Best, & Spiegelhalter, 2000) that is used by WinBUGS and dialects of which are used by JAGS

¹We chose values for the parameters of the prior distributions that fit the introductory example. In general, these values should depend on the application at hand (see Vanpaemel & Lee, this issue; and Morey, this issue).

Table 4.1: Example data set for linear regression. Attendance at each session of a conference, as predicted by the organizers (left) and as observed (middle), with the corresponding “S-style” data file (right).

Expected (x)			Observed (y)			
51	24	32	33	35	32	$x \leftarrow c(51, 44, 57, 41, 53, 56,$
44	21	42	55	18	31	49, 58, 50, 32, 24, 21,
57	23	27	49	14	37	23, 28, 22, 30, 29, 35,
41	28	38	56	31	17	18, 25, 32, 42, 27, 38,
53	22	32	58	13	11	32, 21, 21, 12, 29, 14)
56	30	21	61	23	24	$y \leftarrow c(33, 55, 49, 56, 58, 61,$
49	29	21	46	15	17	46, 82, 53, 33, 35, 18,
58	35	12	82	20	5	14, 31, 13, 23, 15, 20,
50	18	29	53	20	16	20, 33, 32, 31, 37, 17,
32	25	14	33	33	7	11, 24, 17, 5, 16, 7)

and Stan. In all of these programs, the model definition consists of a generative specification. In many cases, the model code is almost a point-to-point translation of a suitable generative specification. Consider this BUGS implementation of the linear regression model:

```
model {
  # linear regression
  for (i in 1:N) {                                # Eq. 4.3
    y[i] ~ dnorm(mu[i], tau)                      # Eq. 4.1
    mu[i] <- beta[1] + beta[2] * x[i]             # Eq. 4.2
  }
  # prior definitions
  beta[1] ~ dnorm(0, 0.001)                      # Eq. 4.4
  beta[2] ~ dnorm(0, 0.001)                      # Eq. 4.5
  tau ~ dgamma(0.001, 0.001)                     # Eq. 4.6
}
```

The parameter `beta[1]` denotes the intercept (i.e., observed number of attendees for 0 expected attendees), `beta[2]` denotes the slope of the regression line (i.e., the increase in the observed number of attendees associated with a one-unit increase in the expected number of attendees), and `tau` represents the inverse of the error variance. This short piece of code maps exactly to the generative model for linear regression that we specified. Of course, since there is much more freedom in mathematical expression than there is in computer code, the point-to-point translations will not always be perfect, but it will typically be an excellent starting point.

In the code, deterministic variables are followed by the `<-` assignment operator. For instance, the line `mu[i] <- beta[1] + beta[2] * x[i]` specifies that the `mu` parameters are given by a linear combination of the stochastic `beta` variables and the observed data `x`. The `#` symbol is used for comments. The complete list of distributions, functions, logical operators, and other programming constructs that are available in WinBUGS, JAGS,

and Stan, is listed in their respective user manuals. BUGS is a declarative language, which means that the order of the statements in the model file is largely irrelevant. In contrast, in Stan, the order of statements matters. With the model translated from formal assumptions to BUGS language, the next step is to interact with the software and sample from the posterior distribution of the parameters.

WinBUGS Graphical User Interface

WinBUGS (Bayesian inference Using Gibbs Sampling for Windows; Lunn et al., 2000; Lunn, Spiegelhalter, Thomas, & Best, 2009; Spiegelhalter, Thomas, Best, & Lunn, 2003; for an introduction see Kruschke, 2010, Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012, and Lee & Wagenmakers, 2013) is a stand-alone piece of software that is freely available at <http://www.mrc-bsu.cam.ac.uk/bugs/>. In this section we give a brief description of the WinBUGS graphical user interface (GUI) using the linear regression model introduced above; later we illustrate how WinBUGS can be called from other software, such as R and MATLAB. For a detailed step-by-step introduction to the WinBUGS GUI, the reader is referred to Lee and Wagenmakers (2013).

To interact with WinBUGS via the GUI, users have to create a number of files. First, there is a *model file* that describes the generative specification of the model, second is the *data file* that contains the raw data, and third is an *initial values file* that contains some starting values for the sampling run.

Panel A in Figure 4.1 shows the model file `linreg_model.txt` that describes the generative model for the linear regression example. Panel B shows the data file `data.txt`. The data specification follows S-plus object notation, where vectors are encapsulated in the concatenation operator `c(...)` and matrices are defined as structures with a dimension field, such as `structure(.Data = c(...), .Dim = c(R, C))`, where `R` stands for the number of rows and `C` for the number of columns. In the linear regression example, the data consist of the vector of observations `y` corresponding to the observed number of attendees, a vector of observations `x` corresponding to the predicted number of attendees, and a scalar `N` corresponding to the number of sessions.

The same data format is used to store the (optional, but strongly recommended) set of initial values for the unobserved stochastic variables. If initial values are not supplied, WinBUGS will generate these automatically by sampling from the prior distribution of the parameters. Automatically generated initial values can provide poor starting points for the sampling run and may result in numerical instability. If multiple MCMC chains are run in order to diagnose convergence problems, we encourage users to create a separate file for each set of initial values. As shown in Panel C in Figure 4.1, we will run three chains, each with a different set of initial values, and store these in `inits1.txt`, `inits2.txt`, and `inits3.txt`.

Once the model file, the data file, and the files containing the initial values are created, follow the steps outlined below to sample from the posterior distribution of the parameters.

1. Load the model file and check the model specification. To open the model file, go to `File -> Open` and select `linreg_model.txt` in the appropriate directory. To check

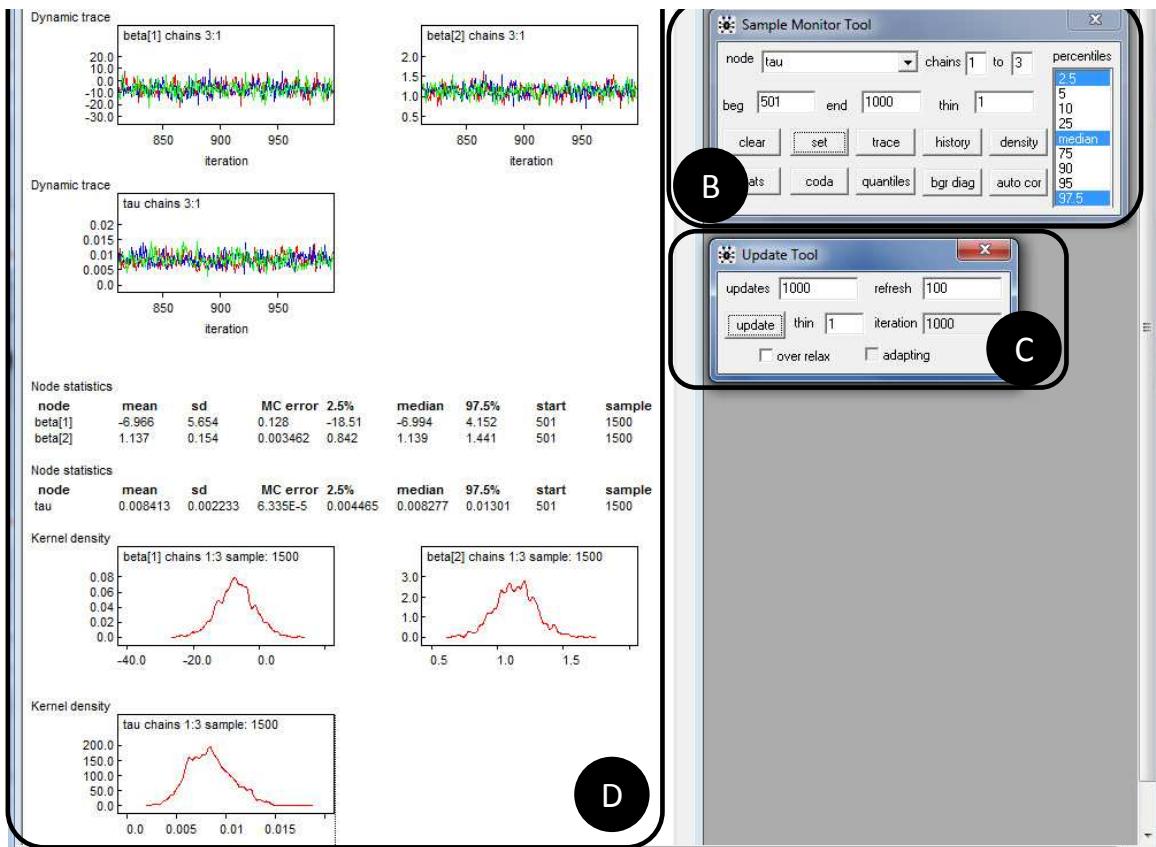


Figure 4.1: The WinBUGS graphical user interface. Panel A shows the model file; Panel B shows the data file; Panel C shows the initial values; Panel D shows the Specification Tool window; Panel E show the status bar.

the syntax of the model specification, go to **Model -> Specification** and open the **Specification Tool** window (Panel D in Figure 4.1), activate the model file by clicking inside `linreg.model.txt`, click on **check model**, and wait for the message “model is syntactically correct” to appear in the status bar.

2. Load the data file. To open the data file, go to **File -> Open** and select `data.txt` in the appropriate directory. To load the data, activate the data file, click on **load data** in the **Specification Tool** window, and wait for the message “data loaded” to appear in the status bar.
3. Compile the model. To compile the model, specify the number of MCMC chains in the box labeled **num of chains** in the **Specification Tool** window, click on **compile**, and wait for the message “model compiled” to appear in the status bar. In the linear regression example, we will run three MCMC chains, so we type “3” in the **num of chains** box.
4. Load the initial values. To open the file that contains the initial values for the first chain, go to **File -> Open** and select `inits1.txt` in the appropriate directory. To

load the first set of initial values, activate `inits1.txt`, click on `load inits` in the **Specification Tool** window, and wait for the message “chain initialized but other chain(s) contain uninitialized variables”. Repeat these steps to load the initial values for the second and third MCMC chain. After the third set of initial values is loaded, wait for the message “model is initialized” to appear in the status bar (Panel E in Figure 4.1).

5. Choose the output type. To ensure that WinBUGS pastes all requested output in a single user-friendly log file, go to **Output -> Output options**, open the **Output options** window, and select the `log` option (Panel A in Figure 4.2).
6. Specify the parameters of interest. To specify the parameters that you want to draw inference about, go to **Inference -> Samples**, open the **Sample Monitor Tool** window, type one by one the name of the parameters in the box labeled `node`, and click on `set` (Panel B in Figure 4.2). In the linear regression example, we will monitor the `beta[1]`, `beta[2]`, and `tau` parameters. To request dynamic trace plots of the progress of the sampling run, select the name of the parameters in the drop-down menu in the **Sample Monitor Tool** window and click on `trace`. WinBUGS will start to display the dynamic trace plots once the sampling has begun.
7. Specify the number of recorded samples. To specify the number of recorded samples per chain, fill in the boxes labeled `beg`, `end`, and `thin` in the **Sample Monitor Tool** window. In our linear regression example, we will record 500 posterior samples for each parameter. We will discard the first 500 samples as burn-in and start recording samples from the 501th iteration (`beg=501`); we will draw a total of 1,000 samples (`end=1000`); and we will record each successive sample without thinning the chains (`thin=1`).
8. Sample from the posterior distribution of the parameters. To sample from the posteriors, go to **Model -> Update**, open the **Update Tool** window (Panel C in Figure 4.2), fill in the total number of posterior samples per chain (i.e., 1,000) in the box labeled `updates`, specify the degree of thinning (i.e., 1) in the box labeled `thin`, click on `update`, and wait for the message “model is updating” to appear in the status bar.
9. Obtain the results of the sampling run. To obtain summary statistics and kernel density plots of the posterior distributions, select the name of the parameters in the drop-down menu in the **Sample Monitor Tool** window and click on `stat` and `density`. WinBUGS will print all requested output in the log file (Panel D in Figure 4.2). The figures labeled “Dynamic trace” show trace plots of the monitored parameters; the three MCMC chains have mixed well and look identical to one another, indicating that the chains have converged to the stationary distribution and that the successive samples are largely independent. The table labeled “Node statistics” shows summary statistics of the posterior distribution of the parameters computed based on the sampled values. For each monitored parameter, the table displays the mean, the median, the standard deviation, and the upper-and lower bound of the central 95% credible interval of the

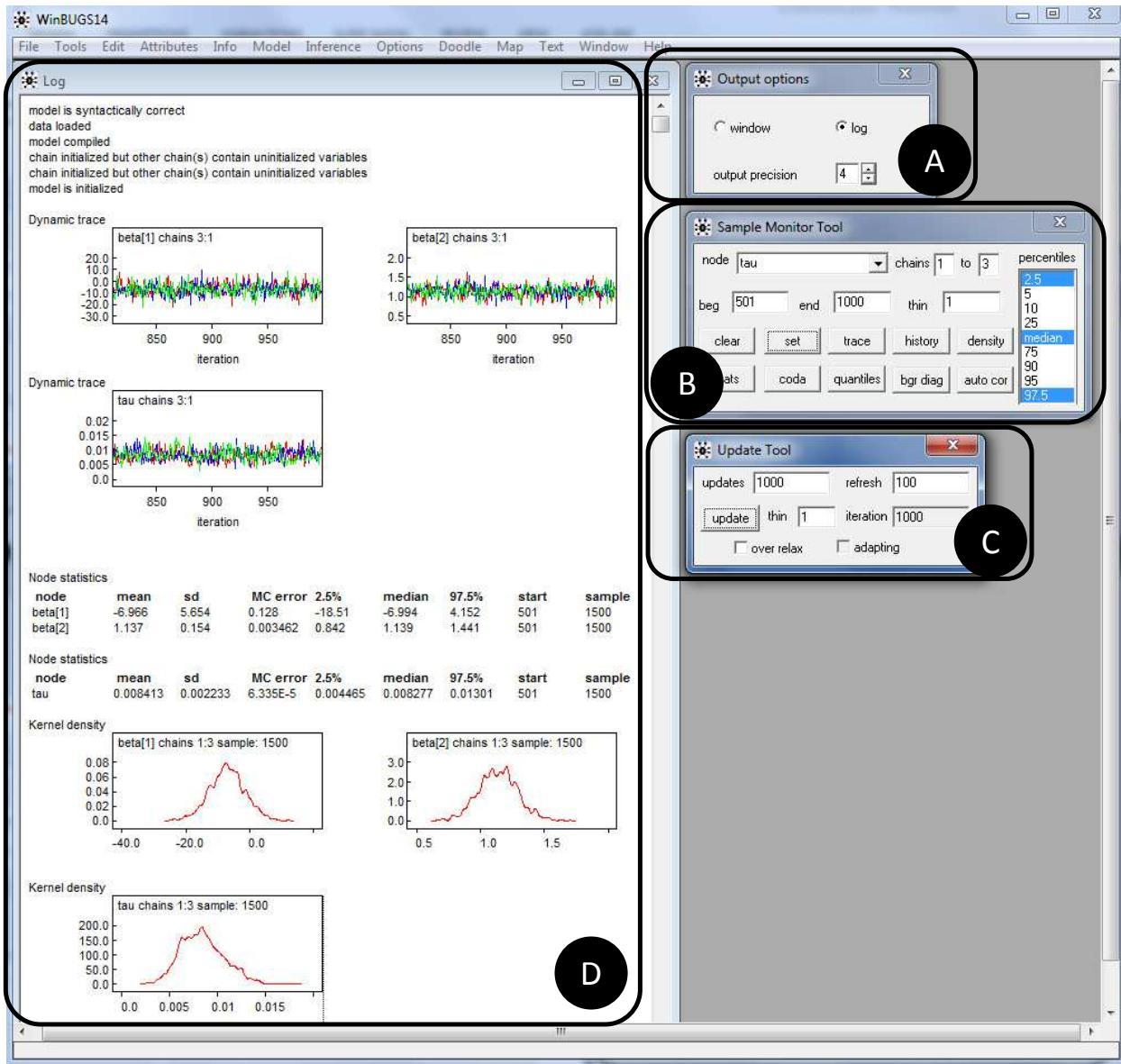


Figure 4.2: The WinBUGS graphical user interface continued. Panel A shows the Output Options window; Panel B shows the Sample Monitor Tool window; Panel C shows the Update Tool window; Panel D shows the log file.

posterior distribution. The central tendency of the posterior, such as the mean, can be used as a point estimate for the parameter. This 95% credible interval ranges from the 2.5th to the 97.5th percentile of the posterior and encompasses a range of values that contains the true value of the parameter with 95% probability; the narrower this 95% credible interval, the more precise the parameter estimate. The figures labeled “Kernel density” show density plots of the posterior samples for each parameter.

As the reader might have noticed by now, running analyses via the GUI is inflexible and labor-intensive; the GUI does not allow for data manipulation and visualization and requires

users to click through a large number of menus and options. Later we therefore illustrate how WinBUGS can be called from standard statistical software, such as R and MATLAB.

JAGS and Stan Command-Line Interface

Both JAGS and Stan are based on a command-line interface. Although this type of interface has fallen out of fashion, and it is strictly speaking not required to use either of these programs, we introduce this low-level interface here—using JAGS as the example—in order to provide the reader with an appreciation of the inner workings of other interfaces. Readers who are not interested in this can skip to either one of the next two sections.

Before launching the program, it is again useful to make a set of text files containing the model, data, and initial values. The model file should contain the code in the listing above; for this example, we saved the model in `linreg_model.txt`.

The data file should contain the data, formatted as in the right column of Table 4.1. The data format in Table 4.1 is sometimes referred to as “S-style”; each variable name is given in double quotation marks, followed by the assignment operator `<-` and the value to be assigned to the variable. Vectors are encapsulated in the concatenation operator `c(...)` and matrices are defined as structures with a dimension field: `struct(c(...), .Dim=c(R,C))`, where the $R \times C$ matrix is entered in column-major order. Our data file is called `linreg_data.txt`.

The same data format is used to store the (optional, but strongly recommended) set of initial values. For at least some of the unknowns nodes (i.e., nodes which in the BUGS code are followed by the sampling operator `~`), initial values should be provided. If multiple chains will be run, one unique file for each chains is recommended. Our initial values files are called `inits1.txt`, `inits2.txt`, and `inits3.txt`.

Once all these files are in place, start JAGS by opening a command window and typing `jags`. Below is the complete interaction with JAGS, in which user input is preceded by the period (.) prompt. Comments are preceded by a pound sign `#`.

```
~$ jags
Welcome to JAGS 3.4.0 on Mon Jul 20 14:02:50 2015
JAGS is free software and comes with ABSOLUTELY NO WARRANTY
Loading module: basemod: ok
Loading module: bugs: ok
. model in "linreg_model.txt"          # loads the model
. data in "linreg_data.txt"            # loads the data
Reading data file linreg_data.txt
. compile, nchains(3)                 # compiles the model for 3 chains
Compiling model graph
  Resolving undeclared variables
  Allocating nodes
  Graph Size: 117
. parameters in "inits1.txt"          # loads initial values for chain 1
Reading parameter file inits1.init
. parameters in "inits2.txt"          # loads initial values for chain 2
```

```

Reading parameter file inits2.txt
. parameters in "inits3.txt"          # loads initial values for chain 3
Reading parameter file inits3.txt
. initialize                          # sets up the sampling algorithms
Initializing model
. update 500                           # draws 500 samples for burn-in
Updating 500
-----| 500
***** 100%
. monitor set beta, thin(1)          # indicates a variable to save
. monitor set tau, thin(1)           # indicates a variable to save
. update 500                           # draws 500 samples from posterior
Updating 500
-----| 500
***** 100%
. coda *, stem('samples_')          # saves posterior samples to files
. exit                                # exits

```

This will produce a set of files starting with `samples_chain` and an index file starting with `samples_index`. These files can be loaded into a spreadsheet program like Microsoft Excel or LibreOffice Calc (or command line tools like awk and perl) to compute summary statistics and do inference. However, this approach is both tedious and labor-intensive, so there exist convenient interfaces from programming languages such as R, MATLAB, and Python.

Working from MATLAB

MATLAB is a commercial software package that can be obtained via <http://www.mathworks.com/>. Just like Python or R, MATLAB can be used to format data, generate initial values, and visualize and save results of a sampling run. In this section we outline how users can interact with WinBUGS, JAGS, and Stan using MATLAB. R users can skip this section; in the next sections, we will describe how to use R for the same purposes.

To interact with the three computational engines from MATLAB, we will use the Trinity toolbox (Vandekerckhove, 2014), which is developed as a unitary interface to the Bayesian inference engines WinBUGS, JAGS, and Stan. Trinity is a work-in-progress that is (and will remain) freely available via <http://tinyurl.com/matlab-trinity>. The MATLAB code needed to call these three engines from Trinity is essentially identical.

To start Trinity, download the toolbox, place it in your MATLAB path, and then call:

```
>> trinity install
>> trinity new
```

The first line will cause the Trinity files to be detected by MATLAB and the second line will create a bare-bones MATLAB script with programming instructions. For example, one line

reads:²

```
% Write the model into a variable (cell variable)
model = {
    %% MODEL GOES HERE $%
};
```

The user can then enter the model code directly into the MATLAB script, using cell string notation (note the single quotes around each line):

```
model = {
    'model {'
    '    # linear regression'
    '    for (i in 1:N) {'
    '        y[i] ~ dnorm(mu[i], tau)'
    '        mu[i] <- beta[1] + beta[2] * x[i]'
    '    }'
    '    # prior definitions'
    '    beta[1] ~ dnorm(0, 0.001)'
    '    beta[2] ~ dnorm(0, 0.001)'
    '    tau ~ dgamma(0.001, 0.001)'
    '}',
};
```

It is also possible to write the model in a separate file and provide the file name here instead of the model code. One advantage of writing model code directly into the MATLAB script is that the script can be completely self-contained. Another is that the model code, when treated as a MATLAB variable, could be generated on-the-fly if tedious or repetitive code is required to define a model or if the model file needs to be adapted dynamically (e.g., if variable names need to change from one run to another).

Next, we need to list the parameters of interest (i.e., for which variables we should save posterior samples). For our current application we could list all variables but choose to omit `mu` (which is particularly useful if `N` is large and the vector `mu` takes up much memory):

```
% List all the parameters of interest (cell variable)
params = {
    'beta' 'tau'
};
```

Next, we collect the data variables that MATLAB will send to the computational engine. Again, it is possible to do this by providing the name to a properly formatted data file, but it is more practical to make a MATLAB variable that contains the data. To collect the data, make a *structure variable* as follows (for the example, `x` and `y` should first be defined with the values given in Table 4.1):

```
% Make a structure with the data (note that the name of the field needs to
```

²It is likely that the exact appearance of this code will vary a little over successive versions of the Trinity toolbox, but the requirements will remain broadly the same.

```
% match the name of the variable in the model)
data = struct(...  
    'x', x, ...  
    'y', y, ...  
    'N', numel(x) ...  
);
```

Each field name (in single quotes) is the name of the variable as it is used in the model definition.³ Note that Trinity will not permit the model code to have variable names containing underscores, as this symbol is reserved for internal use. Following each field name is the value that this variable will take; this value is taken from the MATLAB workspace, so it can be an existing variable with any name, or it can be a MATLAB expression that generates the correct value, as we did here with `N`. Of course, before making this data structure, the data may need to be parsed, read into MATLAB, and possibly pre-processed (outliers removed, etc.).

The final block to complete is a little more involved and requires understanding of MATLAB's *anonymous functions* construct. Anonymous functions are in-line function definitions that are saved as variables. A typical command to define an anonymous function has the following structure:

$$\underbrace{\text{anonfun}}_{(1)} = \underbrace{@(\text{a},\text{b})}_{(2)} \underbrace{3*\text{a} + \text{sqrt}(\text{b})}_{(3)};$$

In this example, `anonfun` (part (1)) is the name given to the new function—this can be anything that is a valid MATLAB variable name. Part (2) indicates the start of an anonymous function with the `@` symbol and lists the input variables of the function between parentheses. Part (3) is a single MATLAB expression that returns the output variable, computed from inputs `a` and `b`. This anonymous function could be invoked with: `anonfun(1,4)`, which would yield 5.

It is possible for an anonymous function to take no (zero) input arguments. For example, `nrand = @() - rand` will create a function called `nrand` that generates uniformly distributed variates between -1 and 0 . In order to supply the computational engine with initial values for the sampling process, we will define an anonymous function that draws a sample from the prior distribution of all or part of the parameter set. An example is:

```
% Write a function that generates a structure with one random value for
```

³Because MATLAB does not differentiate between vectors and single-column or single-row matrices, but some of the computational engines do, it is sometimes convenient to pass variables explicitly as a matrix or explicitly as a vector. For this situation, Trinity allows the flags `AS_MATRIX_` and `AS_VECTOR_` to be prepended to any variable name. A common situation in which this is useful is when matrix multiplication is applied in the model, but one of the matrices has only one column. JAGS, for example, will treat that matrix as a vector and throw a “dimension mismatch” error unless the flag is applied. In our example, the data structure would then be defined as `struct('AS_MATRIX_x', x)`.

```
% each parameter in a field
generator = @()struct...
    'beta' , randn(2, 1) * 10 + 0, ...
    'tau' , rand * 5 ...
);
```

Here, a structure is generated with one field for each parameter, and a random initial value for each. The initial value for each of the two `betas` is generated from a normal distribution with mean 0 and standard deviation 10, and `tau` is generated from a uniform distribution between 0 and 5. The function `generator()` can now be called from MATLAB:

```
>> generator()
ans =
    beta: [2x1 double]
    tau: 0.6349
```

Note that either of these variables can be validly omitted, but at least one must be given. If one of the random number generators draws a value that is not allowed by the model (e.g., where the prior or likelihood is zero), the engines will throw errors (e.g., JAGS will call them “invalid parent values”). If no initial values are given, both the engine and Trinity will proceed without error, *but in some engines all MCMC chains will have the same starting point*, rendering any convergence statistics invalid. It is always prudent to provide at least some initial values. Initial values can be scalars, vectors, or matrices, as needed.

Once all of these variables are prepared, they can be handed off to the main function of Trinity, `callbayes`. This function can take a large number of input fields to control the behavior of the engine, which can be WinBUGS, JAGS, or Stan (WinBUGS is currently limited to Windows operating systems, and Stan is limited to unix-based systems). To select the computational engine, set `engine` to `'bugs'`, `'jags'`, or `'stan'`. (Note that if Stan is selected, the model code should be changed to the Stan code provided in the next section.) More detail regarding the use of `callbayes` can be found in its help documentation (`doc callbayes`). These default inputs are generally sufficient:

```
1 [stats, chains, diagnostics, info] = callbayes(engine, ...
2     'model' , model , ... % the model as a cell
3     'data' , data , ... % the data as a struct
4     'outputname' , 'samples' , ... % any character string
5     'init' , generator , ... % an anonymous function
6     'datafilename' , proj_id , ... % any character string
7     'initfilename' , proj_id , ... % any character string
8     'scriptfilename' , proj_id , ... % any character string
9     'logfile' , proj_id , ... % any character string
10    'nchains' , 3 , ... % the number of chains
11    'nburnin' , 1000 , ... % the burnin period
12    'nsamples' , 10000 , ... % how many saved samples?
13    'monitorparams' , params , ... % the cell string
14    'thin' , 1 , ... % the thinning factor
```

```

15   'workingdir'      ,     ['/tmp/' proj_id] , ... % a temp dir
16   'verbosity'       ,      0 , ... % higher is more verbose
17   'saveoutput'      ,     true , ... % save JAGS log file?
18   'parallel'        ,    false , ... % use multiple cores?
19   'modules'         , { 'dic' } );      % use extra modules?

```

Often, many of these settings can be omitted, and Trinity will choose default values that are appropriate for the engine and operating system. The first input selects the engine. The various '`*filename`' inputs on lines 6–9 serve to organize the temporary files in a readable fashion, so that the user can easily access them for debugging or reproduction purposes.⁴

The input values on lines 10–14 determine how many independent chains should be run, how many samples should be used for burn-in, how many samples should be saved per chain, which parameters should be saved, and by how much the chains should be thinned (n means every n^{th} sample is saved). Line 15 determines a working directory, which is currently set to a value that will work well on unix systems; Windows users might want to change this. Line 16 determines how much output Trinity gives while it is running. Line 17 decides whether the text output given by the engine should be saved.

Line 18 determines if parallel processing should be used—if this is set to `true`, all the chains requested on line 10 will be started simultaneously.⁵ Note that for complex models, this may cause computers to become overburdened as all the processing power is used up by Trinity. Users who want to run multiple chains than they have computing cores available can use the optional input pair '`numcores`', C , ..., where C is the maximum number of cores Trinity is allowed to use. Finally, line 19 lists optional extra modules (JAGS only). By default, the `dic` module is called because this facilitates tracking of the model deviance as a variable. Users with programming experience can create their own modules for inclusion here (e.g., '`wiener`'; see Wabersich & Vandekerckhove, 2014).

A successful `callbayes` call will yield up to four output arguments. `stats` contains summary statistics for each saved parameter (mean, median, standard deviation, and the mass of the posterior below 0). These can be used for easy access to parameter estimates. `chains` contains all the posterior samples saved. The usefulness of this is discussed below. `diagnostics` provides quick access to the convergence metric \hat{R} and the number of effective samples (Gelman & Rubin, 1999). `info` gives some more information, in particular the model variable and a list of all the options that were set for the analysis (combining the user-provided and automatically generated settings).

The most important output variable is `chains`, which contains the saved posterior sam-

⁴When using JAGS or Stan, the working directory will contain a file with a cryptic name that starts with `tp` and ends in a sequence of random characters, with no file extension. This is the entry point script that Trinity uses to call the engine. It can be used to reproduce the analysis outside of MATLAB, if desired—the files in that directory that do not have the `.txt` extension are all that is needed for reproduction. The `*.txt` files are output, containing the posterior samples and the log file. When using WinBUGS, data files, initial values files, and model files will be available in the working directory where they can be accessed with the WinBUGS GUI.

⁵On unix systems, this requires the installation of the free program GNU parallel (Tange, 2011). On Windows systems, it currently requires the MATLAB Parallel Computing Toolbox, but we are working to resolve this dependency.

ples that are the immediate goal of the MCMC procedure. This variable is used by practically all functions in Trinity that do post-processing, summary, and visualization. The default Trinity script contains the line `grtable(chains, 1.05)`. The `grtable` function prints a table with a quick overview of the sampling results, such as the posterior mean, the number of samples drawn, the number of effective samples (`n_eff`) and the \hat{R} convergence metric. The second input to `grtable` can be either a number, in which case only parameters which an \hat{R} larger than that number will be printed (or a message that no such parameters exist); or it can be a string with a *regular expression*, in which case only parameters fitting that pattern will be shown.⁶

Another useful function that relies on the `chains` variable and on regular expressions is `codatable`, which prints a table with user-selected statistics for selected parameters. For example, to see the posterior mean and standard deviation of the `beta` parameters:

```
>> codatable(chains, 'beta', @mean, @std)
Estimand      mean      std
beta_1      -6.937    5.624
beta_2       1.139    0.1554
```

Finally, Trinity contains a set of functions for visualizing MCMC chains and posterior distributions, but for the present application, a simple scatter plot and regression line suffice (Figure 4.3):

```
scatter(x, y)
line(xlim, stats.mean.beta_1 + stats.mean.beta_2 * xlim)
```

Note that the posterior distributions of the regression parameters contain the first bisector ($\beta_1 \approx 0$, $\beta_2 \approx 1$).

Working from R

R (R Development Core Team, 2004) is a free statistical software package that can be downloaded from <http://www.r-project.org/>. In this section, we outline how users can interact with WinBUGS, JAGS, and Stan using R. As with MATLAB, using R to run analyses increases flexibility compared to working with these Bayesian engines directly; users can use R to format the data, generate the initial values, and visualize and save the results of the sampling run using simple R commands.

Interacting with WinBUGS: R2WinBUGS

To interact with WinBUGS, users have to install the `R2WinBUGS` package (Sturtz, Ligges, & Gelman, 2005). The `R2WinBUGS` package allows users to call WinBUGS from within R

⁶Regular expressions are an extremely powerful and flexible programming constructs. To give some examples: if the expression is '`beta`', all parameters with the string `beta` in their name will be shown. If it is '`^beta`', only parameters starting with that string will be shown. '`beta$`' will show only those ending in that string. '`.`' will match any variable, and '`be|ta`' will match anything containing `be` or `ta`. A complete overview to regular expressions in MATLAB can be found via the documentation for the function `regexp`.

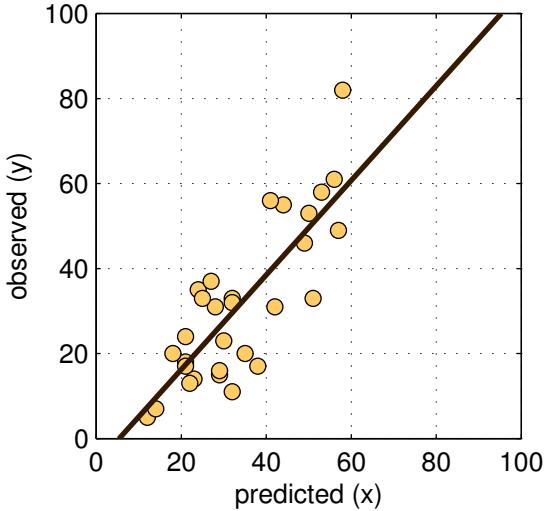


Figure 4.3: Results of the linear regression example. The best fitting regression line is very close to the first bisector $y = x$.

and pass on the model specification, the data, and the initial values to WinBUGS using the `bugs()` function. WinBUGS then samples from the posterior distribution of the parameters and returns the MCMC samples to R.

The following R code can be used to sample from the posterior distribution of the model parameters in the linear regression example using WinBUGS.

```
# set working directory
setwd("C:/Dropbox/My Documents/Bayesian_estimation/WinBUGS")
# load R2WinBUGS package
library(R2WinBUGS)
```

The `setwd()` function specifies the working directory where R will look for the model file and will save the results. The `library()` function loads the `R2WinBUGS` package.

```
# create vector that contains the expected number of attendees
x <- c(51, 44, 57, 41, 53, 56, 49, 58, 50, 32,
       24, 21, 23, 28, 22, 30, 29, 35, 18, 25,
       32, 42, 27, 38, 32, 21, 21, 12, 29, 14)
# create vector that contains the observed number of attendees
y <- c(33, 55, 49, 56, 58, 61, 46, 82, 53, 33,
       35, 18, 14, 31, 13, 23, 15, 20, 20, 33,
       32, 31, 37, 17, 11, 24, 17, 5, 16, 7)
# create a scalar that contains the number of sessions
N <- 30
# create a list that contains the data and will be passed on to WinBUGS
mydata <- list("y", "x", "N")
```

Here we create a list named `mydata` that contains the data (i.e., `x`, `y`, and `N`) and will be passed on to WinBUGS.

```
# create the initial values for the unobserved stochastic nodes
myinits=function(){
  list(beta=rnorm(2, 0, 10), tau=runif(1, 0, 5))
}
```

Here we create the initial values for the unobserved stochastic nodes. The initial values for `beta[1]` and `beta[2]` are random deviates from a zero-centered normal distribution with a standard deviation of 10 generated using the `rnorm()` function. The initial values for `tau` are generated from a uniform distribution with lower bound of 0 and upper bound of 5 using the `runif()` function. The code generates a unique set of initial values for each chain.

```
# specify parameters of interest
myparameters <- c("beta", "tau")
```

Here we create a vector that contains the names of the model parameters that we want to draw inference about.

```
# call WinBUGS
samples <- bugs(data=mydata, inits=myinits, parameters=myparameters,
  model.file="linreg_model.txt", n.chains=3, n.iter=1000, n.burnin=500,
  n.thin=1, DIC=FALSE, bugs.directory="C:/WinBUGS14", codaPkg=FALSE,
  debug=FALSE)
```

The `bugs()` function calls WinBUGS and passes on the model specification, the data, and the start values using the following arguments:

- `data` specifies the list object that contains the data.
- `inits` specifies the list object that contains the initial values.
- `parameters` specifies the vector that lists the names of the parameters of interest.
- `model.file` specifies the text file that contains the model specification. The `model.file` argument can also refer to an R function that contains the model specification that is written to a temporary file.
- `n.chain` specifies the number of MCMC chains.
- `n.iter` specifies the total number of samples per chain.
- `n.burnin` specifies the number of samples per chain that will be discarded at the beginning of the sampling run.
- `n.thin` specifies the degree of thinning.
- `DIC` specifies whether WinBUGS should return the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) measure of model comparison.

- `bugs.directory` specifies the location of `WinBUGS14.exe`.
- `codaPkg` specifies the output that is returned from WinBUGS. Here `codaPkg` is set to `FALSE` to ensure that WinBUGS returns the posterior samples in the `samples` object. If `codaPkg` is set to `TRUE`, WinBUGS returns the paths to a set of files that contains the WinBUGS output.
- `debug` specifies whether WinBUGS will be automatically shut down after sampling. Here `debug` is set to `FALSE` to ensure that WinBUGS shuts down immediately after sampling and returns the results to R. If `debug` is set to `TRUE`, WinBUGS will not shut down after sampling and will display summary statistics and trace plots of the monitored parameters. As the name suggests, setting `debug` to `TRUE` can also provide—often cryptic—cues for debugging purposes.

For more details on the use of `bugs()`, the reader is referred to the help documentation.

Once WinBUGS has finished sampling, it returns the posterior samples to R in the `samples` object. The results of the sampling run can be accessed, visualized, and summarized using, for instance, the following code:

```
# display the first 15 samples in the first chain for tau
samples$sims.array[1:15,1,"tau"]
# plot a histogram of the posterior distribution of tau
hist(samples$sims.array[,,"tau"])
# display summary statistics of the posterior distributions
print(samples)
```

The posterior samples for `beta[1]`, `beta[2]`, and `tau` are stored in `samples$sims.array` (or `samples$sims.list`). The `hist()` function can be used to plot histograms of the posterior distribution of the parameters based on the samples values. The `print(samples)` command displays a useful summary of the posterior distribution of each model parameter, including the mean, the standard deviation, and the quantiles of the posteriors, and (if multiple chains are run) the \hat{R} convergence metric.

Interacting with JAGS: R2jags

To interact with JAGS, users have to install the `R2jags` package (Su & Yajima, 2012). The `R2jags` package allows users to call JAGS from within R and pass on the model specification, the data, and the start values to JAGS using the `jags()` function. JAGS then samples from the posterior distribution of the parameters and returns the MCMC samples to R.

The R code for running the MCMC routine for the linear regression example in JAGS is similar to the R code for running the WinBUGS analysis outlined in the previous section, with the following modifications. Instead of loading the `R2WinBUGS` package, load the `R2jags` package by typing:

```
# load R2jags
library(R2jags)
```

Once the `mydata`, `myinits`, and `myparameters` objects are created in R, use the `jags()` function to call JAGS and sample from the posterior distribution of the parameters:

```
# call JAGS
samples <- jags(data=mydata, inits=myinits,
  parameters.to.save=myparameters, model.file="linreg_model.txt",
  n.chains=3, n.iter=1000, n.burnin=500, n.thin=1, DIC=FALSE)
```

The `jags()` function takes as input the following arguments:

- `data` specifies the list object that contains the data.
- `inits` specifies the list object that contains the initial values.
- `parameters.to.save` specifies the vector that lists the names of the parameters of interest.
- `model.file` specifies the file that contains the model specification. The `model.file` argument can also refer to an R function that contains the model specification that is written to a temporary file.
- `n.chains` specifies the number of MCMC chains.
- `n.iter` specifies the total number of samples per chain.
- `n.burnin` specifies the number of samples per chain that will be discarded at the beginning of the sampling run.
- `n.thin` specifies the degree of thinning.
- `DIC` specifies whether JAGS should return the DIC.

For more details on the use of `jags()`, the reader is referred to the help documentation.

Once JAGS has finished sampling, it returns the posterior samples to R in the `samples` object. The results of the sampling run can be accessed, visualized, and summarized using, for instance, the following code:

```
# display the first 15 samples in the first chain for tau
samples$BUGSoutput$sims.array[1:15,1,"tau"]
# plot a histogram of the posterior distribution of tau
hist(samples$BUGSoutput$sims.array[,,"tau"])
# display summary statistics of the posterior distributions
print(samples)
# plot traceplot; press ENTER for page change
traceplot(samples)
```

The posterior samples for `beta[1]`, `beta[2]`, and `tau` are stored in `samples$BUGSoutput$sims.array` (or `samples$BUGSoutput$sims.list`), and can be visualized and summarized using the `hist()` and `print()` functions, respectively. As the name suggests, the `traceplot(samples)` command displays trace plots of the model parameters, which provide useful visual aids for convergence diagnostics.

Interacting with Stan: rstan

To interface R to Stan, users need to install the `rstan` package (Guo et al., 2015). The `rstan` package allows users to call Stan from within R and pass the model specification, data, and starting values to Stan using the `stan()` function. The MCMC samples from the posterior distribution generated by Stan are then returned and can be further processed in R.

There are a few differences between WinBUGS/JAGS and Stan that are worth noting when specifying Stan models. While JAGS and WinBUGS simply interpret the commands given in the model, Stan compiles the model specification to a C++ program. Consequently, Stan differentiates between a number of different variable types, and variables in a model need to be declared before they can be manipulated. Moreover, model code in Stan is split into a number of blocks, such as “*data*” and “*model*”, each of which serves a specific purpose. Finally, unlike in WinBUGS and JAGS, the order of statements in a Stan model matters and statements cannot be interchanged with complete liberty.

To run the R code for the linear regression example in Stan, begin by loading the `rstan` package:

```
# load rstan package
library(rstan)
```

The `mydata`, `myinits`, and `myparameters` are created in R as illustrated before. However, as Stan relies on a somewhat different syntax than WinBUGS and JAGS, we need to rewrite the model file so it can be parsed by Stan. Here we chose to specify the Stan model as a vector string in R and pass it directly to Stan’s sampling function. Note, however, that we could get the same result by simply saving the code as, say, `linreg_model.stan`.

```
# specify Stan model as a string vector
linreg_model <- "data{
  int<lower=1> N;
  vector[N]      x;
  vector[N]      y;
}

parameters{
  vector[2]      beta;
  real<lower=0>  sigma2;
}

transformed parameters{
  real<lower=0>  tau;
  tau <- pow(sigma2, -1);

  real<lower=0>  sigma;
  sigma <- pow(sigma2, 0.5);
}

model{
  // prior definitions
  beta[1] ~ normal(0, sqrt(1000));                                // Eq. 4.4
  beta[2] ~ normal(0, 100);
  y ~ normal(beta[1] + beta[2]*x, sigma);
}
```

```

beta[2] ~ normal(0, sqrt(1000));                                // Eq. 4.5
// inverse gamma prior for the variance
sigma2 ~ inv_gamma(0.001, 0.001)                                // Eq. 4.6
// linear regression
for(i in 1:N){                                                 // Eq. 4.3
    y[i] ~ normal(beta[1] + beta[2] * x[i], sigma);           // Eqs. 4.1-
    4.2
}
}

```

There are a number of very obvious ways in which this model specification differs from that in WinBUGS and JAGS. The model code is split into four blocks and all variables that are mentioned in the “model” block are defined in the preceding blocks. The “data” block contains the definition of all observed data that are provided by the user. The “parameters” block contains the definition of all stochastic variables, and the “transformed parameters” block contains the definition of all transformations of the stochastic variables. The difference between these latter two parts of the code is rather subtle and has to do with the number of times each variable is evaluated during the MCMC sampling process; a more elaborate explanation can be found in the Stan reference manual (Stan Development Team, 2015).

We will not discuss the specifics of all the variable definitions here (see Stan Development Team, 2015, for details) but will rather illustrate a few important points using as example the `tau` variable. As in the model specification for WinBUGS and JAGS, `tau` is the precision of the Gaussian distribution. Defining a variable for the precision of the Gaussian is, strictly speaking, not necessary because distribution functions in Stan are parameterized in terms of their standard deviation. Nevertheless, we retain `tau` for easy comparability of the Stan MCMC samples with the output of WinBUGS or JAGS. The first line of the definition of `tau` states that it is a real number that is not smaller than 0, and Stan will return an error message should it encounter a negative value for `tau` during the sampling process. The next line states that `tau` is the inverse of the variance of the Gaussian. If we were to reverse the order of these last two lines, due to Stan’s line-by-line evaluation of the code, we would get an error message stating that the variable `tau` is not defined.

The specification of the actual sampling statements in the “model” block begins, in line with Stan’s line-by-line evaluation style, with the prior distributions for the regression coefficients `beta[1]` and `beta[2]` and the variance of the Gaussian. Note that the prior for `sigma2` is an inverse gamma distribution—this is equivalent to the prior specification in the WinBUGS/JAGS model where the inverse of the variance was given a gamma prior. Finally, we summarized equations 4.1 and 4.2 into a single line, which is another way in which the Stan model specification differs from the WinBUGS/JAGS code. While WinBUGS does not allow users to nest statements within the definition of a stochastic node, Stan (and also JAGS) users can directly specify the mean of the Gaussian to be a function of the regression coefficients and observed data `x`, without needing to define `mu[i]`.

To sample from the posterior distribution of the parameters, call the `stan()` function:

```
# call Stan
samples <- stan(data = mydata, init = myinits, pars = myparameters,
```

```
model_code = linreg_model,
chains = 3, iter = 1000, warmup = 500, thin = 1)
```

The `stan()` function takes as input the following arguments:

- `data` specifies the list object that contains the data.
- `init` specifies the list object that contains the initial values.
- `pars` specifies the vector that lists the names of the parameters of interest.
- `model_code` specifies the string vector that contains the model specification. Alternatively, the name of a `.stan` file that contains the model specification can be passed to Stan using the `file` argument.
- `chains` specifies the number of MCMC chains.
- `iter` specifies the total number of samples per chain.
- `warmup` specifies the number of samples per chain that will be discarded at the beginning of the sampling run.
- `thin` specifies the degree of thinning.

For more details on the use of `stan()`, we refer readers to the corresponding R help file.

Once sampling is finished, Stan returns the posterior samples to R in the `samples` object. The results of the sampling run can be accessed, visualized, and summarized using the following code:

```
# display the first 15 samples for tau
extract(samples, pars="tau", inc_warmup=F)$tau[1:15]
```

The posterior samples in the `samples` object can most easily be accessed using the `extract()` function, which takes as input arguments:

- `samples` object containing the posterior samples from Stan.
- `pars` character vector with the names of the parameters for which the posterior samples should be accessed.
- `inc_warmup` logical value indicating whether warm-up samples should be extracted too.

```
# plot a histogram of the posterior distribution of tau
hist(extract(samples, pars="tau")$tau)
# display summary statistics of the posterior distributions
print(samples)
# plot traceplot; press ENTER for page change
traceplot(samples)
```

The posterior samples for `beta[1]`, `beta[2]`, and `tau` can be visualized and summarized using the `hist()` and `print()` functions, respectively. As the name suggests, the `traceplot(samples)` command displays trace plots of the model parameters, which provide useful visual aids for convergence diagnostics.

Example: Multinomial Processing Tree for Modeling False-Memory Data

In this section, we illustrate the use of WinBUGS, JAGS, and Stan for Bayesian parameter estimation in the context of multinomial processing trees, popular cognitive models for the analysis of categorical data. As an example, we will use data reported in Wagenaar and Boer (1987). The data result from an experiment in which misleading information was given to participants who were asked to recall details of a studied event. The data were previously revisited by Vandekerckhove, Matzke, and Wagenmakers (2015), and our discussion of Wagenaar and Boer's experiment and their three possible models of the effect of misleading postevent information on memory closely follows that of Vandekerckhove et al..

The experiment proceeded in four phases. Participants were first shown a sequence of drawings involving a pedestrian–car collision. In one particular drawing, a car was shown at an intersection where a traffic light was either red, yellow, or green. In the second phase, participants were asked questions about the narrative, such as whether they remembered a pedestrian crossing the road as the car approached the “traffic light” (in the consistent-information condition), the “stop sign” (in the inconsistent-information condition) or the “intersection” (the neutral group). In the third phase, participants were given a recognition test. They were shown pairs of pictures from Phase I, where one of the pair had been slightly altered (e.g., the traffic light had been replaced by a stop sign), and asked to pick out the unaltered version. In the final phase, participants were informed that there had indeed been a traffic light, and were then asked to recall the color of the light.

The data consist of the frequency with which participants' responses fall into each of the four response categories, where each response category is characterized by a distinct response pattern: both Phase III and Phase IV answers are correct (Correct–Correct), Phase III answer is correct but Phase IV answer is incorrect (Correct–Incorrect), Phase III answer is incorrect but Phase IV answer is correct (Incorrect–Correct), and both Phase III and Phase IV answers are incorrect (Incorrect–Incorrect). The data from the Wagenaar and Boer (1987) experiment are shown in Figure 4.5; the figure shows the frequency of participants in each of the four response categories in the consistent, inconsistent, and neutral conditions.

The first theoretical account on the effect of misleading postevent information is Loftus' *destructive-updating* model. This model predicts that when conflicting information is presented, it replaces and destroys the original information. Second is the *coexistence* model, under which the initial memory is suppressed by an inhibition mechanism. However, the suppression is temporary and can revert. The third model is the *no-conflict* model, under which misleading postevent information cannot replace or suppress existing information, so that it only has an effect if the original information is somehow missing (i.e., was not encoded

or is forgotten).

Multinomial Processing Tree Models

The three theoretical accounts can be cast as *multinomial processing tree models* (MPT), which translate a decision tree like the one in Figure 4.4 into a multinomial distribution (Batchelder & Riefer, 1980; Chechile, 1973; Riefer & Batchelder, 1988). Figure 4.4 shows the tree associated with the no-conflict model. In Phase I of the experiment, the presence of the traffic light is correctly stored with probability p . If this phase is successful, the color is encoded next, with success probability c . In Phase II, the false presence of the stop sign is stored with probability q . In Phase III, the answer is either known or guessed correctly with probability $1/2$, and in Phase IV the answer is either known or guessed correctly with probability $1/3$.

To calculate the probability of the four possible response patterns (i.e., correct vs. error in Phase III and correct vs. error in Phase IV), we add together the probabilities of each branch leading to that response pattern. The probability of each branch being traversed is given by the product of the individual probabilities encountered on the path. For example, under the no-conflict model, the probability (and hence, expected proportion) of getting Phase III correct but Phase IV wrong is (adding the paths in Figure 4.4 from left to right and starting at the bottom from those cases where Phase III was correct but Phase IV was not): $\frac{2}{3} \times q \times (1 - c) \times p + \frac{2}{3} \times (1 - q) \times (1 - c) \times p + \frac{2}{3} \times \frac{1}{2} \times (1 - q) \times (1 - p)$.

The two competing models both add one parameter to the no-conflict model. In the case of the destructive-updating model, we add one parameter d for the *probability that the traffic light information is destroyed upon encoding the stop sign*. In the case of the coexistence model, we instead add one parameter s for the *probability that the stop sign encoding causes the traffic light information to be suppressed, not destroyed*, so that it remains available in Phase IV.

Here we focus on the no-conflict model, but implementing the other models would involve only small changes to our code. The generative specification of the no-conflict model for the consistent (cons), inconsistent (inco) and neutral (neut) conditions is as follows:

$$\text{cons} \sim \mathcal{M}(\theta_{(1,:)}, N_1) \quad (4.7)$$

$$\text{inco} \sim \mathcal{M}(\theta_{(2,:)}, N_2) \quad (4.8)$$

$$\text{neut} \sim \mathcal{M}(\theta_{(3,:)}, N_3), \quad (4.9)$$

where \mathcal{M} denotes that the data follow a multinomial distribution and N refers to the number of participants in the n^{th} , $n = 1, 2, 3$, condition. The 3×4 matrix θ contains the category probabilities of the multinomial distributions in the three conditions, where $\theta_{(n,:)}$ refers to the n^{th} row of θ . For each condition, the four category probabilities are expressed in terms of the three model parameters p , q , and c . As shown in Figure 4.4, the category probabilities map onto the the four response categories and the corresponding response patterns, and are obtained by following the paths in the tree representation of the model. In particular, the category probabilities in the three conditions are given by:

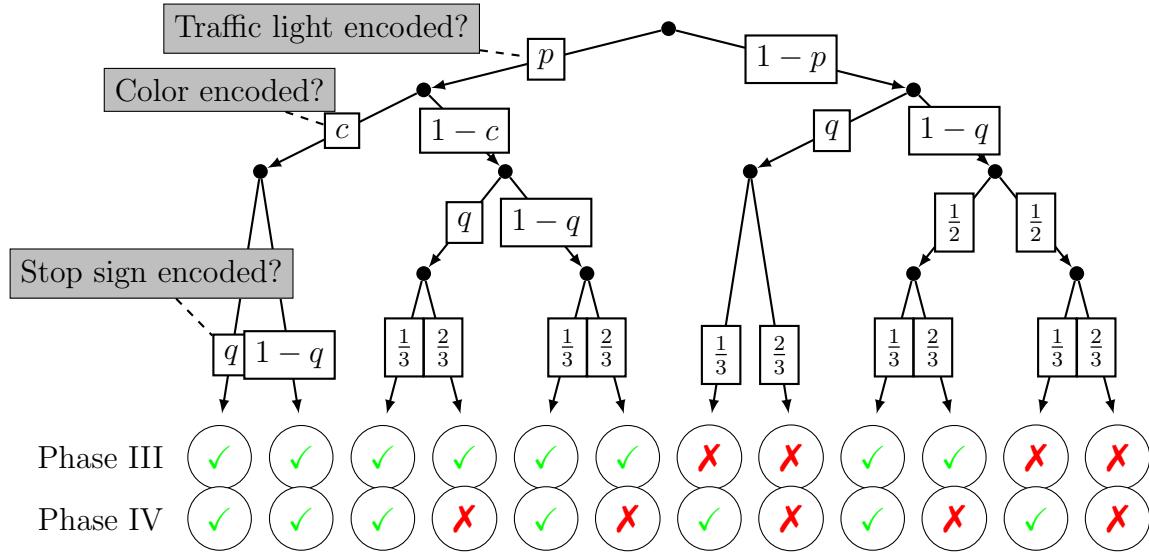


Figure 4.4: Multinomial processing tree representation of the inconsistent condition according to the no-conflict model (adapted from Wagenaar & Boer, 1987). The probability of each of the four response patterns (i.e., correct vs. error in Phase III and correct vs. error in Phase IV) is given by adding the probabilities of each branch leading to that data response pattern. The probability of each branch is given by the product of the individual probabilities encountered on the path.

$$\theta_{(1,1)} = (1 + p + q - pq + 4pc)/6 \quad (4.10)$$

$$\theta_{(1,2)} = (1 + p + q - pq - 2pc)/3 \quad (4.11)$$

$$\theta_{(1,3)} = (1 - p - q + pq)/6 \quad (4.12)$$

$$\theta_{(1,4)} = (1 - p - q + pq)/3 \quad (4.13)$$

$$\theta_{(2,1)} = (1 + p - q + pq + 4pc)/6 \quad (4.14)$$

$$\theta_{(2,2)} = (1 + p - q + pq - 2pc)/3 \quad (4.15)$$

$$\theta_{(2,3)} = (1 - p + q - pq)/6 \quad (4.16)$$

$$\theta_{(2,4)} = (1 - p + q - pq)/3 \quad (4.17)$$

$$\theta_{(3,1)} = (1 + p + 4pc)/6 \quad (4.18)$$

$$\theta_{(3,2)} = (1 + p - 2pc)/3 \quad (4.19)$$

$$\theta_{(3,3)} = (1 - p)/6 \quad (4.20)$$

$$\theta_{(3,4)} = (1 - p)/3 \quad (4.21)$$

Finally, our priors are flat beta distributions $\mathcal{B}(1, 1)$; these distributions imply equal prior probability for all values between 0 and 1 (i.e., $\mathcal{B}(1, 1)$ is the same as a standard uniform

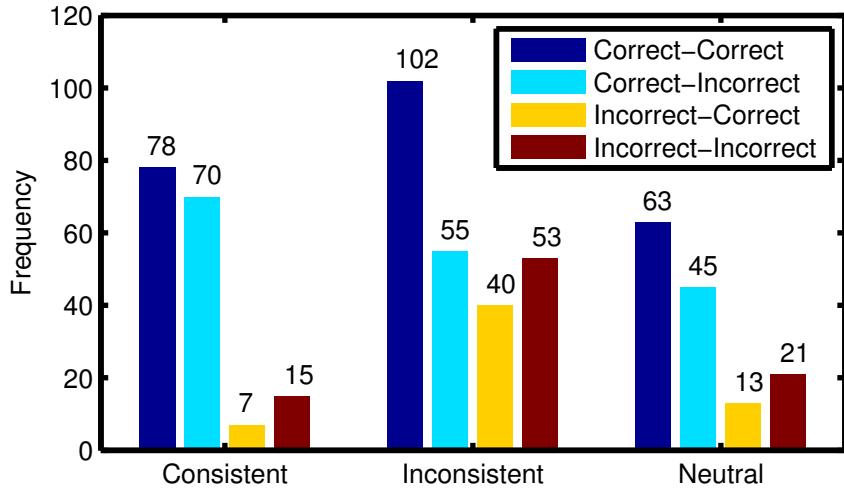


Figure 4.5: The data from the Wagenaar and Boer (1987) experiment. Correct–Correct: Both Phase III and Phase IV answers are correct; Correct–Incorrect: Phase III answer is correct but Phase IV answer is incorrect; Incorrect–Correct: Phase III answer is incorrect but Phase IV answer is correct; Incorrect–Incorrect: Both Phase III and Phase IV answers are incorrect. The data are grouped by condition.

distribution):

$$p \sim \mathcal{B}(1, 1) \quad (4.22)$$

$$q \sim \mathcal{B}(1, 1) \quad (4.23)$$

$$c \sim \mathcal{B}(1, 1) \quad (4.24)$$

We will now fit the no–conflict model to the Wagenaar and Boer (1987) data using WinBUGS, JAGS and Stan in combination with both MATLAB and R. Obtaining parameter estimates for the destructive–updating and the coexistence models requires only minor modifications to the code. In particular, we would have to modify the category probabilities (Equations 4.10–4.21) to reflect the tree architecture of the alternative models and define an additional parameter (i.e., parameter d for the destructive–updating and parameter s for the coexistence model) with the corresponding uniform prior distribution. As an illustration, the Supplemental Material presents the WinBUGS, JAGS and Stan model files and the corresponding R code that allows users to estimate the parameters of the no–conflict as well as the destructive–updating and coexistence models.

Working from R using R2WinBUGS

The WinBUGS code for the generative specification of the no–conflict model is given below. Note here that since the generative model specification is just a list of declarative statements, the order of statements does not matter for the specification. We write the statements here in the order in which they appear in the text. This intentionally violates the usual “programmer

logic” in which variables need to be declared before they are used. We emphasize that such restriction is not needed in WinBUGS code.

```

model {
  # ---- Data ----- #
  cons[1:4] ~ dmulti(theta[1,1:4], N[1])          # Eq. 4.7
  inco[1:4] ~ dmulti(theta[2,1:4], N[2])          # Eq. 4.8
  neut[1:4] ~ dmulti(theta[3,1:4], N[3])          # Eq. 4.9

  # ---- Consistent condition ----- #
  theta[1,1] <- ( 1 + p + q - pq + 4 * pc ) / 6    # Eq. 4.10
  theta[1,2] <- ( 1 + p + q - pq - 2 * pc ) / 3    # Eq. 4.11
  theta[1,3] <- ( 1 - p - q + pq ) / 6            # Eq. 4.12
  theta[1,4] <- ( 1 - p - q + pq ) / 3            # Eq. 4.13

  # ---- Inconsistent condition ----- #
  theta[2,1] <- ( 1 + p - q + pq + 4 * pc ) / 6    # Eq. 4.14
  theta[2,2] <- ( 1 + p - q + pq - 2 * pc ) / 3    # Eq. 4.15
  theta[2,3] <- ( 1 - p + q - pq ) / 6            # Eq. 4.16
  theta[2,4] <- ( 1 - p + q - pq ) / 3            # Eq. 4.17

  # ---- Neutral condition----- #
  theta[3,1] <- ( 1 + p + 4 * pc ) / 6            # Eq. 4.18
  theta[3,2] <- ( 1 + p - 2 * pc ) / 3            # Eq. 4.19
  theta[3,3] <- ( 1 - p ) / 6                      # Eq. 4.20
  theta[3,4] <- ( 1 - p ) / 3                      # Eq. 4.21

  # ---- Priors ----- #
  p ~ dbeta(1,1)                                    # Eq. 4.22
  q ~ dbeta(1,1)                                    # Eq. 4.23
  c ~ dbeta(1,1)                                    # Eq. 4.24

  # ---- Some useful transformations ----- #
  pq <- p * q
  pc <- p * c
}

```

Once the model specification is saved to a text file (e.g., `noconflict.txt`), the following R code can be used to create the data and the initial values, and call WinBUGS using the `R2WinBUGS` package:

```

# load R2WinBUGS package
library(R2WinBUGS)

# create the data

```

```

cons <- c( 78, 70, 7, 15)
inco <- c(102, 55, 40, 53)
neut <- c( 63, 45, 13, 21)
N <- c(170, 250, 142)
mydata <- list("cons", "inco", "neut", "N")

# create the initial values
myinits <- function(){
  list(p=runif(1), q=runif(1), c=runif(1))
}

# specify the parameters of interest
myparameters <- c("p", "q", "c")

# call WinBUGS
samples <- bugs(data=mydata, inits=myinits, parameters=myparameters,
  model.file="noconflict.txt",
  n.chains=3, n.iter=3500, n.burnin=500, n.thin=5,
  DIC=FALSE, bugs.directory="C:/WinBUGS14",
  codaPkg=FALSE, debug=TRUE)

```

Note that we ran 3500 iterations per chain (`n.iter=3500`) and retained only every 5th sample (`n.thin=5`). As the parameters in cognitive models are often strongly correlated, it is typically necessary to run relatively long MCMC chains and thin the chains to reduce auto-correlation. When the sampling run has finished, WinBUGS returns the posterior samples for the three model parameters in the `samples` object. The posterior distribution of the parameters—plotted using the sampled values—is shown in the first column of Figure 4.6.

Working from R using R2jags

The JAGS code for the generative specification of the no-conflict model is identical to the WinBUGS code presented in the previous section, and so is the R code for creating the data and generating the initial values. Once the `R2jags` package is loaded by typing `library(R2jags)`, the following R code can be used to call JAGS and sample from the posterior distribution of the parameters:

```

samples <- jags(data=mydata,inits=myinits,parameters.to.save=myparameters,
  model.file ="noconflict.txt",
  n.chains=3, n.iter=3500, n.burnin=500, n.thin=5, DIC=FALSE)

```

JAGS returns the posterior samples for the three model parameters in the `samples` object. The posterior distribution of the parameters is shown in the second column of Figure 4.6. The posteriors obtained with JAGS are essentially indistinguishable from the ones obtained with WinBUGS.

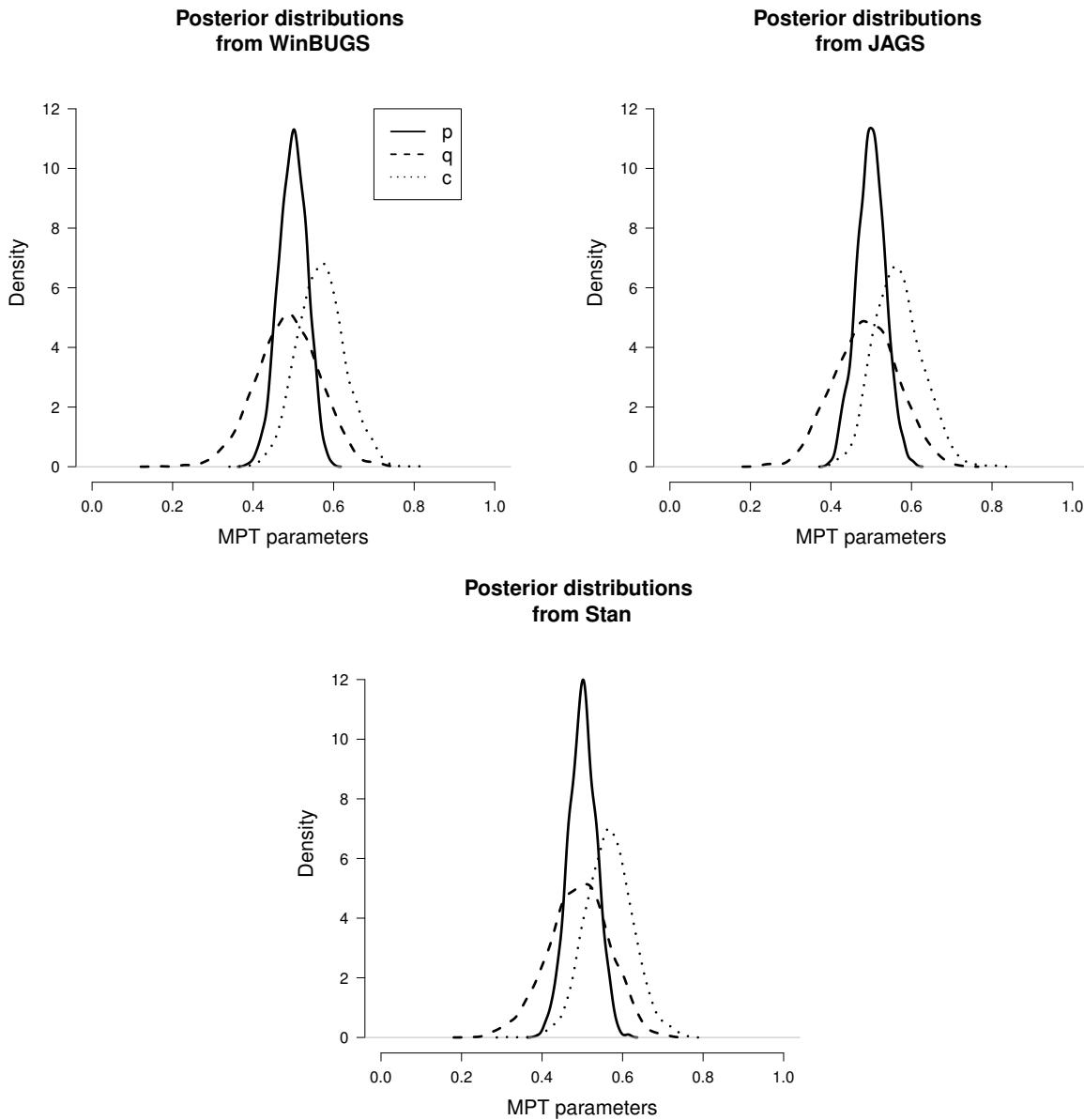


Figure 4.6: The posterior distribution of the parameters of the no-conflict MPT model obtained from WinBUGS, JAGS, and Stan in combination with R. The solid, dashed, and dotted lines show the posterior distribution of the p , q , and c parameters, respectively.

Working from R using rstan

The Stan code for the non-conflict model again differs somewhat from the WinBUGS/JAGS code:

```
data{
  int cons[4];
  int inco[4];
```

```

    int neut[4];
}

parameters{
    real<lower=0,upper=1> p;
    real<lower=0,upper=1> q;
    real<lower=0,upper=1> c;
}

transformed parameters{
    real<lower=0,upper=1> pq;
    real<lower=0,upper=1> pc;
    simplex[4] theta1;
    simplex[4] theta2;
    simplex[4] theta3;

    pq <- p * q;
    pc <- p * c;

    // consistent condition
    theta1[1] <- ( 1 + p + q - pq + 4 * pc ) / 6; // Eq. 4.10
    theta1[2] <- ( 1 + p + q - pq - 2 * pc ) / 3; // Eq. 4.11
    theta1[3] <- ( 1 - p - q + pq ) / 6; // Eq. 4.12
    theta1[4] <- ( 1 - p - q + pq ) / 3; // Eq. 4.13

    // inconsistent condition
    theta2[1] <- ( 1 + p - q + pq + 4 * pc ) / 6; // Eq. 4.14
    theta2[2] <- ( 1 + p - q + pq - 2 * pc ) / 3; // Eq. 4.15
    theta2[3] <- ( 1 - p + q - pq ) / 6; // Eq. 4.16
    theta2[4] <- ( 1 - p + q - pq ) / 3; // Eq. 4.17

    // neutral condition
    theta3[1] <- ( 1 + p + 4 * pc ) / 6; // Eq. 4.18
    theta3[2] <- ( 1 + p - 2 * pc ) / 3; // Eq. 4.19
    theta3[3] <- ( 1 - p ) / 6; // Eq. 4.20
    theta3[4] <- ( 1 - p ) / 3; // Eq. 4.21
}

model{
    // priors
    p ~ beta(1,1); // Eq. 4.22
    q ~ beta(1,1); // Eq. 4.23
    c ~ beta(1,1); // Eq. 4.24
    cons ~ multinomial(theta1); // Eq. 4.7
    inco ~ multinomial(theta2); // Eq. 4.8
    neut ~ multinomial(theta3); // Eq. 4.9
}

```

Once the model specification is saved as `noconflict.stan`, the `rstan` package has been loaded by typing `library(rstan)`, and R objects have been created that contain the data, initial values, and parameters of interest, the following code can be used to obtain samples from the posterior distributions of the parameters:

```
samples <- stan(data = mydata, init = myinits, pars = myparameters,
                 file = 'noconflict.stan',
                 chains = 3, iter = 3500, warmup = 500, thin = 5)
```

The posterior samples for the three model parameters are returned in the `samples` object. The third column of Figure 4.6 shows estimates of the posterior densities based on the sampled values; the posteriors closely resemble those obtained with WinBUGS and JAGS.

Working from MATLAB using Trinity

The code to fit the no-conflict model from MATLAB using Trinity is again very formulaic, and differs very little between the three computational engines. In the bare-bones script automatically generated by `trinity new`, we first enter the data:

```
cons = [ 78 , 70 , 7 , 15 ] ;
inco = [ 102 , 55 , 40 , 53 ] ;
neut = [ 63 , 45 , 13 , 21 ] ;
N = [sum(cons) sum(inco) sum(neut)];
```

After the data are entered, the model definition needs to be provided as a cell string. We omit the model specification here because both the WinBUGS/JAGS and Stan versions are fully given in the previous sections.

Next, we list the parameters of interest in a cell variable:

```
parameters = {
    'c' 'p' 'q'
};
```

and we write a function that generates a structure containing one random value for each parameter in a field:

```
generator = @()struct...
    'c', rand, ...
    'p', rand, ...
    'q', rand ...
);
```

We also enter the data into a structure where we match the names of the fields to the variable names in the model definition:

```
data = struct...
    'cons', cons, ...
    'inco', inco, ...
    'neut', neut, ...
```

```
'N'      , N      ...
);
```

After selecting an engine, the `callbayes` function is called with mostly default settings:

```
%% Run Trinity with the CALLBAYES() function
tic
[stats, chains, diagnostics, info] = callbayes(engine, ...
    'model'           , model , ...
    'data'            , data , ...
    'outputname'     , 'samples' , ...
    'init'             , generator , ...
    'modelfilename'   , proj_id , ...
    'datafilename'    , proj_id , ...
    'initfilename'    , proj_id , ...
    'scriptfilename'  , proj_id , ...
    'logfilename'     , proj_id , ...
    'nchains'         , 4 , ...
    'nburnin'         , 1e4 , ...
    'nsamples'        , 1e4 , ...
    'monitorparams'   , parameters , ...
    'thin'             , 5 , ...
    'refresh'          , 1000 , ...
    'workingdir'       , ['/tmp/' proj_id] , ...
    'verbosity'        , 0 , ...
    'saveoutput'       , true , ...
    'parallel'         , isunix() , ...
    'modules'          , {'dic'} );
```

The engine will return, among others, the `chains` variable containing posterior samples for all three parameters of interest. We can inspect the results, and we can use the `codatable` function to give qualitative feedback about the convergence of the MC chains:

```
if any(codatable(chains, @gelmanrubin) > 1.1)
    grtable(chains, 1.1)
    warning('Some chains were not converged!')
else
    disp('Convergence looks good.')
end
```

Finally, we can inspect the posterior means by chain using the `stats` structure:

```
disp('Posterior means by chain:')
disp(stats.mean)
```

as well as check some basic descriptive statistics averaged over all chains:

```
disp('Descriptive statistics for all chains:')
codatable(chains)
```

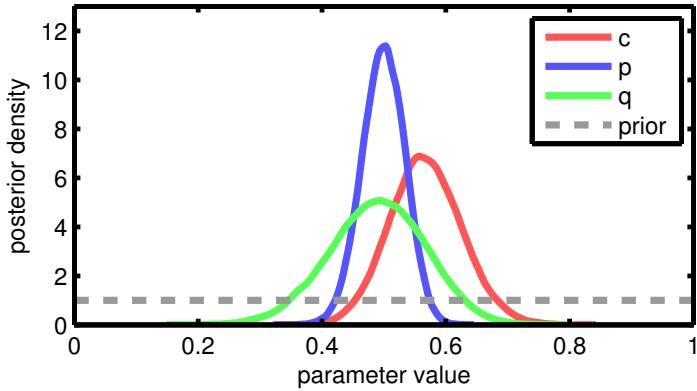


Figure 4.7: The posterior distribution of the parameters of the no-conflict MPT model obtained from JAGS in combination with Trinity.

and visually inspect the posterior distributions using the `smhist` function:

```
smhist(chains, '^c$|^q$|^p$');
```

where the regular expression may be read as “match only variables whose name is exactly `c` or exactly `q` or exactly `p`”. The output of the last command—the posterior distribution of the parameters—is shown in Figure 4.7.

Testing hypotheses

After the posterior samples have been drawn, and posterior distributions possibly visualized as above, there remains the issue of testing hypotheses relating to parameters. With the current false-memory data set, one hypothesis of interest might be that the probability p of encoding the traffic light is greater (versus lower) than chance (Hypothesis 1). The same question might be asked of the probability c of encoding the light color (Hypothesis 2).

Given samples from the posterior, a convenient way of computing the posterior probability that a hypothesis is true is by computing the proportion of posterior samples in which the hypothesis holds. To test Hypothesis 1, we would calculate the proportion of cases in which $p > 0.5$. To test Hypothesis 2, we calculate the proportion of cases in which $c > 0.5$.

The `codatable` command is useful in this regard. Custom statistics on the posterior samples can be computed by providing anonymous functions as secondary input variables. A quick way of counting the proportion of cases in which a condition is true is to make use of the fact that MATLAB represents logically true statements as 1 and false statements as 0. Hence, the anonymous function `@(x)mean(x>.5)` will return the proportion of cases where the input is greater than 0.5:

```
>> codatable(chains, '^c$|^p$', @(x)mean(x>.5))
Estimand      mean  @(x)mean(x>.5)
    c          0.558   0.8441
    p          0.4956   0.4547
```

As it turns out, the probability of Hypothesis 1 given the data is about 84% and that of Hypothesis 2 is about 45%. In other words, neither of the hypotheses is strongly supported by the data. In fact, as Figure 4.7 shows, most of the posterior mass is clustered near 0.5 for all parameters.

Conclusion

Bayesian methods are rapidly rising from obscurity and into the mainstream of psychological science. While Bayesian equivalents of many standard analyses, such as the t test and linear regression, can be conducted in off-the-shelf software such as JASP (Love et al., 2015), custom models will continue to require a flexible programming framework and, unavoidably, some degree of software MacGyverism. To implement specialized models, researchers may write their own MCMC samplers, a process that is time-consuming and labor-intensive, and does not come easy to investigators untrained in computational methods. Luckily, general-purpose MCMC engines—such as WinBUGS, JAGS, and Stan—provide easy-to-use alternatives to custom MCMC samplers. These software packages hit the sweet spot for most psychologists; they provide a large degree of flexibility at a relatively low time cost.

In this tutorial, we demonstrated the use of three popular Bayesian software packages in conjunction with two scientific programming languages, R and MATLAB. This combination allows researchers to implement custom Bayesian analyses from already familiar environments. As we illustrated, models as common as a linear regression can be easily implemented in this framework, but so can more complex models, such as multinomial processing trees (MPT; Batchelder & Riefer, 1980; Chechile, 1973; Riefer & Batchelder, 1988).

Although the tutorial focused exclusively on non-hierarchical models, the packages may also be used for modeling hierarchical data structures (e.g, Lee, 2011). In hierarchical modeling, rather than estimating parameters separately for each unit (e.g., participant), we model the between-unit variability of the parameters with group-level distributions. The group-level distributions are used as priors to “shrink” extreme and poorly constrained estimates to more moderate values. Hierarchical estimation can provide more precise and less variable estimates than non-hierarchical estimation, especially in data sets with relatively few observations per unit (Farrell & Ludwig, 2008; Rouder, Lu, Speckman, Sun, & Jiang, 2005). Hierarchical modeling is rapidly gaining popularity in psychology, largely by virtue of the availability of accessible MCMC packages. The WinBUGS, JAGS, and Stan implementation of most hierarchical extensions is very straightforward and often does not require more than a few additional lines of code. For the hierarchical WinBUGS implementation of regression models, the reader is referred to Gelman and Hill (2007). For the hierarchical implementation of custom models, such as multinomial processing trees, signal detection, or various response time models, the reader is referred to Lee and Wagenmakers (2013), Matzke, Dolan, Batchelder, and Wagenmakers (2015), Matzke and Wagenmakers (2009), Nilsson, Rieskamp, and Wagenmakers (2011), Rouder, Lu, Morey, Sun, and Speckman (2008) and Vandekerckhove, Tuerlinckx, and Lee (2011).

Although the goal of our tutorial was to demonstrate the use of general-purpose MCMC software for Bayesian *parameter estimation*, our MPT-example has also touched on Bayesian

hypothesis testing. Various other Bayesian methods are available that rely on MCMC–output to test hypotheses and formally compare the relative predictive performance of competing models. For instance, Wagenmakers, Lodewyckx, Kuriyal, and Grasman (2010) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009) discuss the use of the Savage–Dickey density ratio, a simple procedure that enables researchers to compute Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995) for nested model comparison using the height of the prior and posterior distributions obtained from WinBUGS. Vandekerckhove et al. (2015) shows how to use posterior distributions obtained from WinBUGS and JAGS to compute Bayes factors for non–nested MPTs using importance sampling. Lodewyckx et al. (2011) outline a WinBUGS implementation of the product–space method, a transdimensional MCMC approach for computing Bayes factors for nested and non–nested models. Most recently, Gronau et al. (2017) provide a tutorial on bridge sampling—a new, potentially very powerful method that is under active development. It is important to note, however, that all of these methods are almost all quite difficult to use and can be unstable, especially for high-dimensional problems.

Throughout the tutorial, we have advocated WinBUGS, JAGS, and Stan as flexible and user-friendly alternatives to homegrown sampling routines. Although the MCMC samplers implemented in these packages work well for the majority of models used in psychology, they may be inefficient and impractical for some. For instance, models of choice and response times, such as the linear ballistic accumulator (Brown & Heathcote, 2008) or the lognormal race (Rouder, Province, Morey, Gomez, & Heathcote, 2015), are notoriously difficult to sample from using standard MCMC software. In these cases, custom-made MCMC routines may be the only solution. For examples of custom-made and non–standard MCMC samplers, the reader is referred to Rouder and Lu (2005) and Turner, Sederberg, Brown, and Steyvers (2013), respectively.

Their general usefulness notwithstanding, the three packages all have their own set of limitations and weaknesses. WinBUGS, as the name suggests, was developed specifically for Windows operating systems. Although it is possible to run WinBUGS under OS X and Linux using emulators such as Darwine and CrossOver or compatibility layers such as Wine, user experience is often jarring. Even under Windows, software installation is a circuitous process and requires users to decode a registration key and an upgrade patch via the GUI. Once installed, users typically find the GUI inflexible and labor–intensive. In interaction with R, user experience is typically more positive. Complaints focus mostly on WinBUGS’ cryptic error messages and the limited number of built–in functions and distributions. Although the WinBUGS Development Interface (WBDev; Lunn, 2003) enables users to implement custom–made functions and distributions, it requires experience with Component Pascal and is poorly documented. Matzke, Dolan, Logan, Brown, and Wagenmakers (2013) provide WBDev scripts for the truncated–normal and ex–Gaussian distributions; Wetzels, Lee, and Wagenmakers (2010) provide an excellent WBDev tutorial for psychologists, including a WBDev script for the shifted–Wald distribution. Importantly, the BUGS Project has shifted development away from WinBUGS; development now focuses on OpenBUGS (<http://www.openbugs.net/w/FrontPage>).

Stan comes equipped with interfaces to various programming languages, including R, Python and MATLAB, and only requires the installation of the specific interface package,

which is easy and straightforward under most common operating systems. In terms of computing time, Stan seems a particularly suitable choice for complex models with many parameters and large posterior sample sizes. This advantage in computing time is due to the fact that Stan compiles the sampling model to a C++ program before carrying out the sampling process. The downside of this compilation step is that, particularly for small models as used in the present tutorial, compilation of the model might require more time than the sampling process itself, in which case WinBUGS or JAGS seem a more advantageous choice.

Finally, we will highlight two advantages of JAGS over Stan. First, as illustrated in our example code, Stan code requires variable declaration and as a result can be somewhat more complicated than JAGS code. Second, as a consequence of Stan's highly efficient Hamiltonian Monte Carlo sampling algorithm, some model specifications are not allowed—in particular, Stan does not easily allow model specifications that require inference on discrete parameters, which reduces its usefulness if the goal is model selection rather than parameter estimation.

We demonstrated the use of three popular Bayesian software packages that enable researchers to estimate parameters in a broad class of models that are commonly used in psychological research. We focused on WinBUGS, JAGS, and Stan, and showed how they can be interfaced from R and MATLAB. We hope that this tutorial can serve to further lower the threshold to Bayesian modeling for psychological science.

References

- Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, 87, 375–397.
- Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, 57, 153–178.
- Chechile, R. A. (1973). *The relative storage and retrieval losses in short-term memory as a function of the similarity and amount of information processing in the interpolated task*. Unpublished doctoral dissertation, University of Pittsburgh.
- Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, 15, 1209–1217.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Gelman, A., & Rubin, D. B. (1999). Evaluating and using statistical methods in the social sciences. *Sociological Methods & Research*, 27, 403–410.
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *arXiv preprint arXiv:1703.05984*.
- Guo, J., Lee, D., Goodrich, B., de Guzman, J., Niebler, E., Heller, T., et al. (2015). *rstan: R interface to stan*.
- Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Burlington, MA: Academic Press.

- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55, 1–7.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge: Cambridge University Press.
- Lodewyckx, T., Kim, W., Tuerlinckx, F., Kuppens, P., Lee, M. D., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, 55, 331–347.
- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., et al. (2015). *JASP [computer software]*. <https://jasp-stats.org/>.
- Lunn, D. J. (2003). WinBUGS Development Interface (WBDev). *ISBA Bulletin*, 10, 10–11.
- Lunn, D. J., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton (FL): Chapman & Hall/CRC.
- Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28, 3049–3067.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.
- Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80, 205–235.
- Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, 142, 1047–1073.
- Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, 16, 798–817.
- Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Package “Bayes factor”. <http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf>.
- Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, 55, 84–93.
- R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-00-3)
- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95, 318–399.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12, 573–604.
- Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, 137, 370–389.
- Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12, 195–223.

- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80, 491–513.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64, 583–639.
- Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS version 1.4 user manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.
- Stan Development Team. (2015). *Stan modeling language: User's guide and reference manual. version 2.7.0*.
- Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12, 1–16.
- Su, Y.-S., & Yajima, M. (2012). *R2jags: A package for running JAGS from R*.
- Tange, O. (2011, Feb). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1), 42-47.
- Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, 18, 368–384.
- Vandekerckhove, J. (2014). *Trinity: A MATLAB interface for Bayesian analysis*. <http://tinyurl.com/matlab-trinity>.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–317). Oxford, UK: Oxford University Press.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16, 44–62.
- Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, 46, 15–28.
- Wagenaar, W. A., & Boer, J. P. A. (1987). Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychologica*, 66, 291–306.
- 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.
- Wetzels, R., Lee, M. D., & Wagenmakers, E.-J. (2010). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods*, 42, 884–897.
- 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.

5

Parameter estimation and Bayes factors

Jeffrey N. Rouder, Julia Haaf, and Joachim Vandekerckhove

Bayesian analysis has become increasing popular in many fields including psychological science. There are many advantages to the Bayesian approach. Some champion its clear philosophical underpinnings where probability is treated as a statement of belief or information and the focus is on updating beliefs rationally in face of new data (Finetti, 1974; Edwards, Lindman, & Savage, 1963). Others note the practical advantages—Bayesian analysis often provides a tractable means of solving difficult problems that remain intractable in more conventional frameworks (Gelman, Carlin, Stern, & Rubin, 2004). This practical advantage is especially pronounced in psychological science where substantive models are designed to account for mental representation and processing. As a consequence, the models tend to be complex and nonlinear, and may include multiple sources of variation (Kruschke, 2011b; Lee & Wagenmakers, 2013; Rouder & Lu, 2005). Bayesian analysis, especially Bayesian nonlinear hierarchical modeling, has been particularly successful at providing straightforward analyses in these otherwise difficult settings (e.g., Rouder, Sun, Speckman, Lu, & Zhou, 2003; Vandekerckhove, Tuerlinckx, & Lee, 2011).

Bayesian analysis is not a unified field, and Bayesian analysts disagree with one another in important ways.¹ We highlight here two popular Bayesian approaches that seem incompatible and discuss them in the context of the simple problem of determining whether performance in two experimental conditions differs. In one approach, termed here the **posterior-estimation approach**, the difference between the conditions is represented by a parameter, and posterior distributions about this parameter are updated using Bayes' Rule. From these posterior distributions, researchers may observe directly which parameter values are plausible, and, as importantly, which are implausible. Two examples are provided in Figure 5.1. In Figure 5.1A the posterior distribution is compact in extent and localized away from zero, and this localization serves as evidence for a statistically substantial difference between the two conditions. In Figure 5.1B, in contrast, the value of zero is well within the belly of

¹Perhaps such disagreements should be expected given the contentious history of academic statistics. Even null hypothesis significance testing is a contentious hybrid of Fisherian and Neyman-Pearson schools of thought (Gigerenzer et al., 1989; Lehmann, 1993).

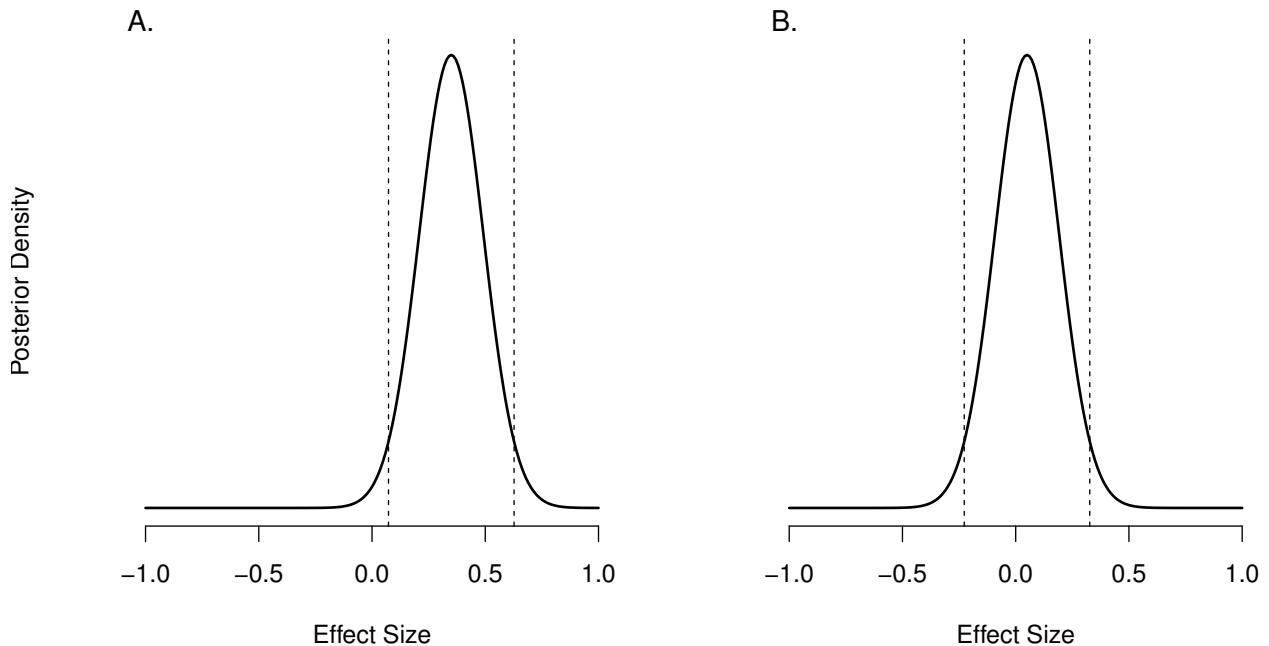


Figure 5.1: In the posterior-estimation approach, the posterior distribution describes which parameter values are plausible and implausible. **A.** The posterior localizes the effect away from zero perhaps providing evidence for an effect. **B.** The posterior localizes the effect around zero perhaps providing a lack of evidence for an effect. Dashed lines indicate the 95% credible intervals on posteriors.

posterior, indicating that there is little evidence for such a difference. Perhaps the leading advocate of the posterior-estimation approach in psychology is Kruschke (Kruschke, 2011a, 2012). Although the posterior-estimation approach seems straightforward, it is not recommended by a number of Bayesian psychologists including Dienes (2014), Gallistel (2009), Rouder, Speckman, Sun, Morey, and Iverson (2009), and Wagenmakers (2007). These authors instead advocate a **Bayes factor** approach. In Bayesian analysis, it is possible to place probability on models themselves without recourse to parameter estimation. In this case, a researcher could construct two models: one that embeds no difference between the conditions and one that embeds some possible difference. The researcher starts with prior beliefs about the models and then updates these rationally with Bayes' rule to yield posterior beliefs. Evidence from data is how beliefs about the models themselves change in light of data; there may be a favorable revision for either the effects or null-effects model.

Posterior estimation and Bayes factor approaches do not necessarily lead to the same conclusions. Consider for example the posterior in Figure 5.1A where the posterior credible interval does not include zero. This posterior seemingly provides positive evidence for an effect. Yet, the Bayes factor, which is discussed at length subsequently, is 2.8-to-1 in favor of the effect. If we had started with 50-50 beliefs about an effect (vs. a lack of an effect), we end up with just less than 75-25 beliefs in light of data. While this is some revision of belief, this small degree is considered rather modest (Jeffreys, 1961; Raftery, 1995) rather than substantial.

This divergence leaves the nonspecialist in a quandary about whether to use posterior estimation or Bayes factors. We fear this may lead some to ignore Bayesian analysis altogether. Here we address this quandary head-on: We will first draw a sharp contrast between the two approaches and show that they provide for quite different views of evidence. Then, to help understand these differences, we provide a unification. We show that the Bayes factor may be represented as estimation under a certain model specification known in the statistics literature as a *spike-and-slab* model (George & McCulloch, 1993). With this demonstration, one difference between estimation and a Bayes factor approach comes into full view—it is a difference in model specification rather than any deep difference in the Bayesian machinery. These spike-and-slab models entail different commitments than more conventional models. Our own view is that the commitments underlying spike-and-slab are the correct ones for most testing questions. Once researchers understand these commitments, they can make informed and thoughtful choices about which are most appropriate for specific research applications.

Posterior Estimation

Bayesian estimation is performed straightforwardly through updating by Bayes' rule. Let us take a simple example where a set of participants provide performance scores in each of two conditions. For example, consider a priming task where the critical variable is the response time, and participants provide a mean response time in a primed and unprimed condition. Each participant's data may be expressed as a difference score, namely the difference between mean response times. Let $Y_i, i = 1, \dots, n$ be these difference scores for n participants. In the usual analysis, researchers would perform a *t*-test to assess whether these difference scores are significantly different from zero.

Bayesian analysis begins with consideration of a model, and in this case, we assume that each difference score is a draw from a normal with mean μ and variance σ^2 :

$$Y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (5.1)$$

In the following development, we will assume that σ^2 is known to simplify the exposition, but it is straightforward to dispense with this assumption. It is helpful to consider the model in terms of effect sizes, δ , where $\delta = \mu/\sigma$ is the true effect size and is the parameter of interest.

Bayesian analysis proceeds by specifying beliefs about the effect-size parameter δ . The beliefs are expressed as a prior distribution on parameters. In this article, we use the term *prior* and *model* interchangeably as a prior is nothing more than a model of parameters. Model \mathcal{M}_1 provides prior beliefs on δ .

$$\mathcal{M}_1 : \quad \delta \sim \text{Normal}(0, \sigma_0^2). \quad (5.2)$$

The centering of the distribution at zero is interpreted as a statement of prior equivalence about the direction of any possible effect—negative and positive effects are *a priori* equally likely. The prior variance, σ_0^2 must be set before analysis, and it is helpful to explore how the value of this setting affects estimation. Figure 5.2A shows this effect. Ten hypothetical

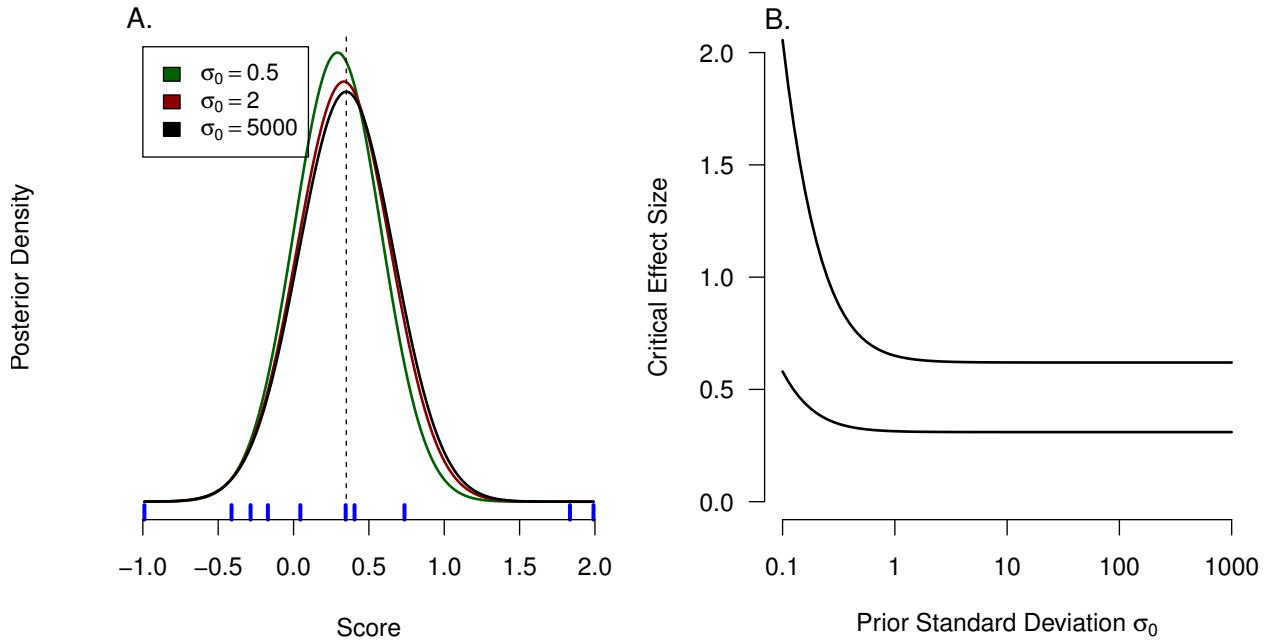


Figure 5.2: The dependence of the posterior estimation on prior setting σ_0 . **A.** Posterior distributions on effect size δ for $N = 10$ and for a sample effect size of .35. for three settings of σ_0 **B.** Minimum observed effect sizes needed such that the posterior 95% credible interval excludes zero. The two lines are for sample sizes of 10 (top) and 40 (bottom). The results show a robustness to the prior setting of σ_0 .

values of Y_i , the difference scores, are shown as line segments across the bottom of the plot. The sample mean of these ten is shown as the vertical line. The posterior distributions of δ are shown for three different prior settings. The first prior setting, $\sigma_0 = .5$, codes an *a priori* belief that δ is not much different than zero. The second prior setting, $\sigma_0 = 2$, is a fairly wide setting that allows for a large range of reasonable effect sizes without mass on exceedingly large values. The third prior setting, $\sigma_0 = 5000$ indicates that researcher is unsure of the effect size, and holds the possibility that it can be exceedingly large. Even though the priors are quite different, the posterior distribution are quite similar. We may say that the posterior is robust to wide variation in prior settings. In fact, it is possible to set $\sigma_0 = \infty$ to equally weight all effect sizes *a priori*, and in this case, the posterior would be indistinguishable from that for $\sigma_0 = 5000$. This robustness to prior settings may be viewed by some as an advantage of Model \mathcal{M}_1 on δ . As will be shown, while this robustness holds for \mathcal{M}_1 , it is not a general property of Bayesian estimation. We will introduce a useful model in the next section where it does not hold.

There are many ways to use the posterior distributions to state conclusions. One could simply inspect them and interpret them as needed (Gelman & Shalizi, 2013 and Rouder et al., 2008 take variants of this approach). Alternatively, one could make a set of inferential rules. In his early career, Lindley (1965) recommended inference by *highest-density credible intervals* (HDCIs). These highest-density credible intervals contain a fixed proportion of the mass, say 95%, and posterior values inside the interval are greater than those outside the

interval. Examples of these HDCIs are shown in Figure 5.1 with the dashed vertical lines. Values outside the intervals may be considered sufficiently implausible to be untenable. By this reasoning, there is evidence for an effect in Figure 5.1A as zero is outside the 95% credible interval. Figure 5.2B shows that inference by credible intervals does not depend heavily on the prior setting σ_0^2 . Shown is the minimal effect size needed such that zero is excluded from the lower end of the credible interval. As can be seen, this value stabilizes quickly and varies little.

Kruschke (2012) takes a similar approach. A posterior interval may be compared to a pre-established region, called a *region of practical equivalence* or ROPE. ROPES are small intervals around zero that are considered to be practically the same as zero. An example of a ROPE might be the interval on effect sizes from $-.2$ to $.2$. In Kruschke's approach, one concludes that the null hypothesis is false if the HDCI falls completely outside of the ROPE. If the HDCI falls completely inside of the ROPE, one concludes that the null hypothesis is (for all practical purposes) true. If the HDCI partly overlaps with the ROPE, Kruschke recommends one reserve judgment. Inferences drawn this way are robust to the prior setting of σ_0 , and arbitrarily large (even infinite) values may be chosen.

Bayes Factors

The contrasting approach is inference by Bayes factors. In Bayesian analysis, it is possible to place beliefs directly onto models themselves and update these beliefs with Bayes' rule. Let \mathcal{M}_A and \mathcal{M}_B denote any two models. Let $Pr(\mathcal{M}_A)$ and $Pr(\mathcal{M}_B)$ be *a priori* beliefs about the plausibility of these two models. It is more desirable to state relative beliefs about the two models as odds. The ratio $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ is the relative plausibility of the models, and for example, the statement $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B) = 3$ indicates that Model \mathcal{M}_A is three times as plausible as Model \mathcal{M}_B . Odds such as $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ are called *prior odds* because they are stipulated before seeing data. They may be contrasted to *posterior odds*, which are the same odds in light of the data and denoted $Pr(\mathcal{M}_A | \mathbf{Y})/Pr(\mathcal{M}_B | \mathbf{Y})$. Be the prior and posterior odds, respectively. Bayes rule for updating to posterior odds from prior odds is

$$\frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)} = \frac{f(\mathbf{Y} | \mathcal{M}_A)}{f(\mathbf{Y} | \mathcal{M}_B)} \times \frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)}. \quad (5.3)$$

The updating factor, $f(\mathbf{Y} | \mathcal{M}_A)/f(\mathbf{Y} | \mathcal{M}_B)$, is called the *Bayes factor*, and it describes how the data have led to a revision of beliefs about the models. Several authors including (Jeffreys, 1961) and Morey, Romeijn, and Rouder (2016) refer to the Bayes factors as the *strength of evidence from data about the models* precisely because the strength of evidence should refer to how data lead to revision of beliefs. The Bayes factor has a second meaning stemming from it being the relative probability of data under models. The probability of data under a model may be thought of as the predictive accuracy of that model – the degree to which the model predicted the data. The data in the equation is the observed date we obtain in an experiment, and if the probability of observed data is high, then the model predicted the observed data to be where they were observed. If the probability of data is low, then the model did not predict the observations well. The Bayes factor is the relative predictive

accuracy of one model relative to another. The deep meaning of Bayes' rule is that the strength of evidence is the *relative predictive accuracy*, and this is captured by the Bayes factor in Equation 5.3.

We denote the Bayes factor by B_{AB} , where the subscripts indicate which two models are being compared. A Bayes factor of $B_{AB} = 10$ means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_A ; likewise, a Bayes factor of $B_{AB} = .1$ means that prior odds should be updated by a factor of 10 in favor of model \mathcal{M}_B . Bayes factors of $B_{AB} = \infty$ and $B_{AB} = 0$ correspond to infinite—total—support of one model over the other with the former indicating infinite support for model \mathcal{M}_A and the latter indicating infinite support for model \mathcal{M}_B .

For the simple example of comparing performance in two experimental conditions, we need one model for an effect a different model for a lack of an effect (which is also called an invariance). A suitable model for an effect is the previous model, \mathcal{M}_1 given in (5.2). A model for an invariance is given by

$$\mathcal{M}_0 : \quad \delta = 0.$$

With this setup, the Bayes factor is straightforward to compute.²

Inference by Bayes factor is more dependent on the prior setting σ_0^2 than is inference by the preceding posterior-estimation approach. Figure 5.3A shows the effects of increasing σ_0 . As can be seen, the Bayes factor B_{10} favors the alternative when σ_0 is small (say, near 1) but decreases toward zero as σ_0 becomes increasingly large. Of note is the limit as σ_0 gets increasingly large. These diffuse priors on effect size in the alternative leads to total support for the null model over the alternative (Lindley, 1957), and this result contrasts to that for inference with credible intervals where inference reflects the data even when σ_0 becomes increasingly large. This result occurs because the Bayes factor is sensitive to the complexity of the model, and when the $\sigma_0^2 = \infty$, the alternative can account for all data equally well, without constraint. Consequently, it is penalized completely. Figure 5.3B provides an different view of the effect of prior setting σ_0 . It shows the minimum positive effect size need to support a Bayes factor of 3-to-1 in favor of Model \mathcal{M}_1 over \mathcal{M}_0 and is comparable to Figure 5.2B. As can be seen, inference by Bayes factor is more sensitive to prior settings than inference by estimation.

At first glance, this dependence of the Bayes factors on the prior settings may seem undesirable. One fear is that researchers can seemingly obtain different results by adjusting the prior settings perhaps undermining the integrity of their conclusions. This dependence seems all the more undesirable when contrasted to the the robustness of posterior intervals to prior settings as shown in Figure 5.2. However, the situation is far more nuanced, and we believe researchers should not worry too much about prior dependence or lack thereof. Indeed both Bayesian parameter estimation and Bayes factor model selection are supported

²The Bayes factor between Model \mathcal{M}_1 and \mathcal{M}_0 is

$$B_{10} = \frac{1}{\sqrt{n\sigma_0^2 + 1}} \exp\left(\frac{n^2 d^2}{2(n + 1/\sigma_0^2)}\right) \quad (5.4)$$

where d is the observed effect size given by \bar{Y}/σ .

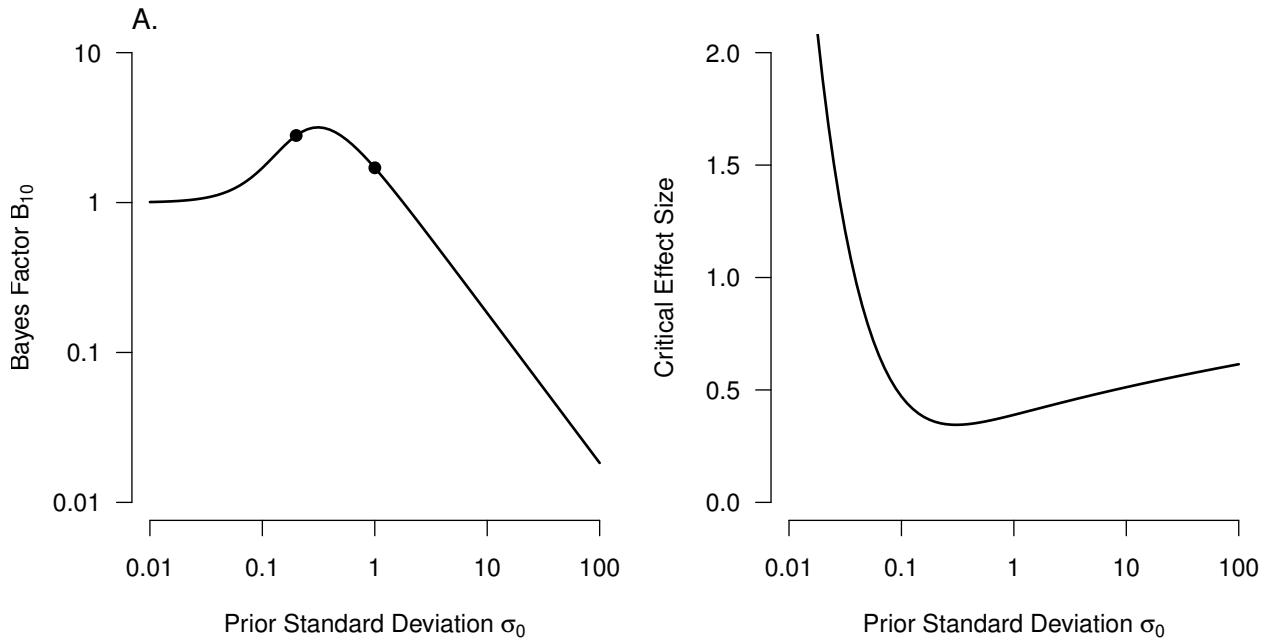


Figure 5.3: The dependence of Bayes factor on prior setting σ_0 . **A.** Bayes factor as a function of σ_0 for $N = 40$ and for an observed effect size of .35 and a sample size of 40 observations. **B.** Minimum observed effect sizes needed such that Bayes factor favors the alternative by 3-to-1. The filled circles show the lower and upper bounds of reasonable variation in prior standard deviation.

by the same rules of probability (see Etz & Vandekerckhove, this issue), and the differences are more subtle and perhaps even more interesting than they first appear. In the next section we provide a unification, and with this unification can pinpoint the differences and make recommendations for researchers.

Unification

The differences between the estimation and Bayes factor approach can be understood by combining models \mathcal{M}_0 and \mathcal{M}_1 . Figure 5.4A shows the combination, which is expressed as a mixture. One component of the mixture is the usual normal model on effect size (Model \mathcal{M}_1), and this component is denoted by the curve in Figure 5.4A. The other component is a placing mass on the point of zero, and this component is denoted by the arrow. In this case, the arrow is half-way up its scale, shown in dashed line, indicating that half of the total mass is placed at zero, and the other half is distributed around zero. This model is well known in the statistics literature as a *spike-and-slab model* (Mitchell & Beauchamp, 1988). We denote it by Model \mathcal{M}_s .³ The spike-and-slab model in Figure 5.4 has two parameters: the amount

³The density of a spike-and-slab model is given by

$$f(\delta) = \rho_0 s(\delta) + (1 - \rho_0)\phi(\delta/\sigma_0),$$

where s is the density of the spike, defined next, ϕ is the density of a standard normal, ρ_0 is the prior mass on the spike, and σ_0^2 is the variance of the slab. The density of the spike, s , is known as a Dirac

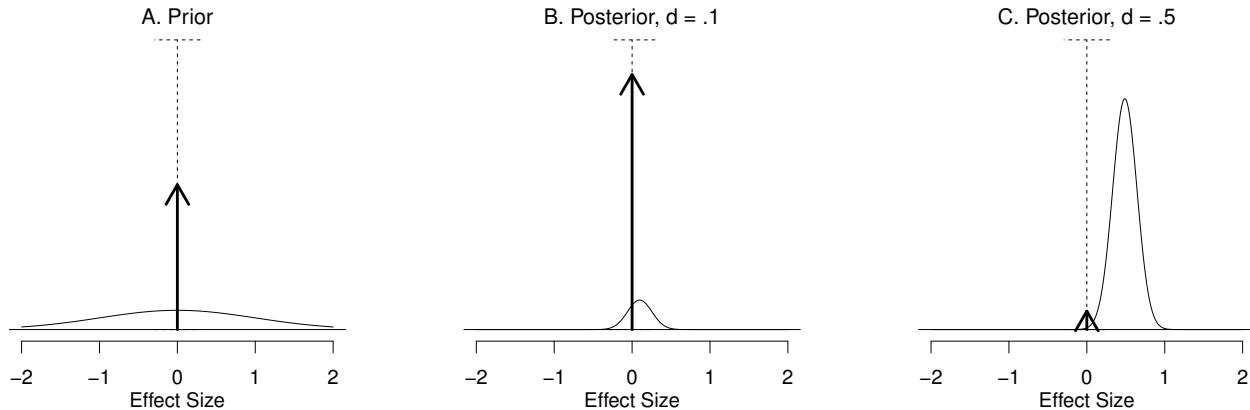


Figure 5.4: The spike-and-slab model is a mixture of a spike, shown as an arrow, and slab, shown as the normal curve. **A.** Prior distribution on effect size with half the mass in the spike, and the slab centered around zero. **B-C.** The posterior on effect size δ for observed effect sizes of $d = .1$ and $d = .5$, respectively, for a sample size of 40.

of probability in the spike, denoted ρ_0 , and the variance of the slab, denoted σ_0^2 . Figure 5.4A shows the case where $\rho_0 = 1/2$ and $\sigma_0^2 = 1$.

It is straightforward to update beliefs about δ in the spike-and-slab model using Bayes' rule.⁴ Figure 5.4B-C show a few examples for different observed effect sizes. In all cases, the resulting posterior is in the spike-and-slab form, but the spike has changed mass and the slab has shifted and rescaled. Figure 5.4B shows the posterior for a small observed effect size of 0.1. The spike is enhanced as the effect is compatible with a null effect. The slab is attenuated in mass, narrowed, and shifted from 0 to about .1. Figure 5.4B shows the posterior for a large observed effect size of 0.5. The spike is attenuated as the effect is no longer compatible with the null, and the slab is enhanced, narrowed, and shifted from 0 to about .5.

delta function and defined as follows: Consider a normal density centered at zero with standard deviation η , denoted $g(\delta) = \phi(\delta/\eta)$. The Dirac delta function, s , is defined as the density in the limit that $\eta \rightarrow 0$:

$$s(\delta) = \lim_{\eta \rightarrow 0} \phi\left(\frac{\delta}{\eta}\right) = \begin{cases} \infty, & \delta = 0, \\ 0, & \text{otherwise.} \end{cases}$$

⁴The resulting posterior density, $f(\delta|\mathbf{Y})$ is

$$f(\delta|\mathbf{Y}) = \rho_1 s(\delta) + (1 - \rho_1) \phi\left(\frac{\delta - \mu_1}{\sigma_1}\right),$$

where

$$\begin{aligned} \sigma_1^2 &= (n + \sigma_0^{-2})^{-1} \\ \mu_1 &= nd\sigma_1^2 \\ \rho_1 &= \frac{\rho_0}{\rho_0 + (1 - \rho_0)B_{01}}, \end{aligned}$$

where d is the observed effect size and B_{01} is the Bayes factor between Model \mathcal{M}_0 and \mathcal{M}_1 .

There is an intimate relationship between the spike-and-slab posterior distribution and the Bayes factor B_{01} for the comparison between models \mathcal{M}_0 and \mathcal{M}_1 : The Bayes factor describes the change in the spike. The prior probability of the spike, ρ_0 , can be expressed as odds, $\omega_0 = \rho_0/(1 - \rho_0)$. The posterior probability of the spike, ρ_1 , can likewise be expressed as odds, $\omega_1 = \rho_1/(1 - \rho_1)$. The Bayes factor is the change in odds: ω_1/ω_0 . In Figure 5.4B, for example, the initial odds on the spike were 1-to-1, indicating that equal mass was in the spike as was in the slab. In light of data, the posterior odds were 7.4-to-1, or that 88% of the posterior mass was in the spike and 12% of posterior mass was in the slab. Indeed, the Bayes factor for this case is $B_{01} = 7.4$, and this factor describes the change in odds in the spike in light of data (because originally they were 1-to-1).

The spike-and-slab Model \mathcal{M}_s yields posterior estimates of effect size that behave differently, in fact more advantageously, than the slab-only estimates from \mathcal{M}_1 . Figure 5.5A-B shows the comparison. The solid curves are posterior means of δ as a function of observed effect size d . For the slab-only specification (Panel A), the estimated mean follows the observed value, and do so for all prior values of σ_0^2 . But, for the spike-and-slab specification (\mathcal{M}_s , Panel B), there is a pull toward zero. This pull is known as shrinkage. Shrinkage is well known in hierarchical models and often results in estimates that have lower error and perform better out of sample(James & Stein, 1961; Efron & Morris, 1977). The shrinkage from the spike-and-slab model is *adaptive* in that shrinkage toward zero is sizable for small observed values while there is hardly any shrinkage for large values. The dynamics are that small observed effect sizes are more compatible with the hypothesis that there is no effect, and therefore, estimates are more influenced by the zero value. Large effects in contrast are more compatible with the hypothesis that there is an effect, and the estimates are more influenced by the sample effect size.

Adaptive shrinkage is an exceedingly useful part of modern Bayesian analysis. It is a Bayesian approach to variable selection, classification, and smoothing, and, as will be discussed, it has become popular in multivariate settings. The amount of adaptive shrinkage depends on the prior setting σ_0^2 . As σ_0^2 increases, there is more shrinkage to zero as the spike is relatively more salient. In this regard, the prior setting σ_0^2 serves as a tuning parameter.

From the behavior of the effect-size estimate that we observe under the spike-and-slab specification, we can draw a few conclusions: First, estimation is made within the context of a specified model. The model is important, and the obtained *parameter estimates are a function of the choice of model specification*. Model estimates with the spike-and-slab prior show adaptive shrinkage – effect size estimates are attenuated towards zero to an extent that depends on the probability of there being any effect at all.⁵ Second, Bayesian parameter estimation is not in itself more robust to prior settings than the Bayes factor: *Robustness to prior settings is a function of model specification*.

These consequences, that the value of estimates and their robustness to prior settings both depend critically on the model specification, hold for inferences drawn from credible intervals as well. Figure 5.5C-E show the comparison of credible intervals. The shaded

⁵The spike-and-slab model allows for an added level of flexibility in that we can inspect the parameter estimate *in the slab by itself* provided that we are comfortable assuming—after seeing the posterior distribution that includes the spike and deciding that the spike mass can now safely be ignored—that there is an effect. In Figure 5.4C, we would arrive at the estimate $\hat{d} \approx 0.5$.

areas show the 95% CIs as a function of observed effect size. For the slab-only specification, the credible intervals maintain a constant width for all observed effect size values. The key question is when does the credible interval include or exclude the value of zero. The vertical dashed lines show transition points – if the observed values are more extreme, then the 95% CI does not include zero. For the slab-only specification, the CIs include zero for observed effect sizes between $-.325$ and $.325$. Values more extreme exclude zero, and by posterior estimation logic, these values lead to a rejection of the null. The spike-and-slab specifications results in different behavior for the credible intervals. For extreme observed values, say those greater than $.65$ in magnitude, the CIs are quite similar to the slab-only specification. For less extreme values, the spike has influence, and the resulting 95% CI often includes the value of zero. As a result, the transition points are wider – it takes more extreme observed values to localize the 95% CI away from zero. For when $\sigma_0 = 1$, the null may be rejected only if the observed effect size is more extreme in magnitude than $.55$, which is quite a bit larger than the $.325$ for slab-only specification. This value is increased to $.585$ when $\sigma_0 = 10$.

The unification through spike-and-slab priors highlights similarities and differences between inference from posterior estimation and inference from Bayes factors as they are commonly used in psychology. The similarities are obvious, both methods are sibling approaches in the Bayes' rule family lineage. They rely similarly on specification of detailed models including models on parameters (priors), and updating follows naturally through Bayes' rule. There are differences as well, and the difference we highlight here is that from model specification. The recommended methods of inference by estimation, say those from Kruschke, rely on priors that preclude spikes at set points such as points of invariance. The Bayes factor approaches we have developed in Guan and Vandekerckhove (2015), Rouder et al. (2009), Rouder and Morey (2012) and Rouder, Morey, Speckman, and Province (2012), place point-mass on prespecified, theoretically important values. It is this difference in model specification—rather than the difference in inferential statistic—that leads to some of the most salient differences in practice.

Which Model Specification To Use?

A critical question for researchers is then which model specification to use. The answer is that the choice depends on the context of the analysis and the goals of the researcher. As a rule of thumb, if zero is a theoretically meaningful or important quantity of interest it makes sense to consider a point mass on zero. This specification allows one to compute a posterior probability of the theoretically meaningful value. For instance, in the usual testing scenarios, researchers consider the “no-effect” baseline to be qualitatively different than effects. The spike-and-slab model instantiates this qualitative difference, and in the process license the theoretically useful abstractions of “effect” and “no effect.” In the context of this goal, of stating evidence for or against effects, it is reasonable and judicious to use a spike-and-slab estimation approach or, equivalently, a Bayes-factor summary of the change in the spike probability. In some cases, perhaps ones where measurement is a main goal and where the zero value has no special meaning, a slab-only approach may be best. Researchers in these

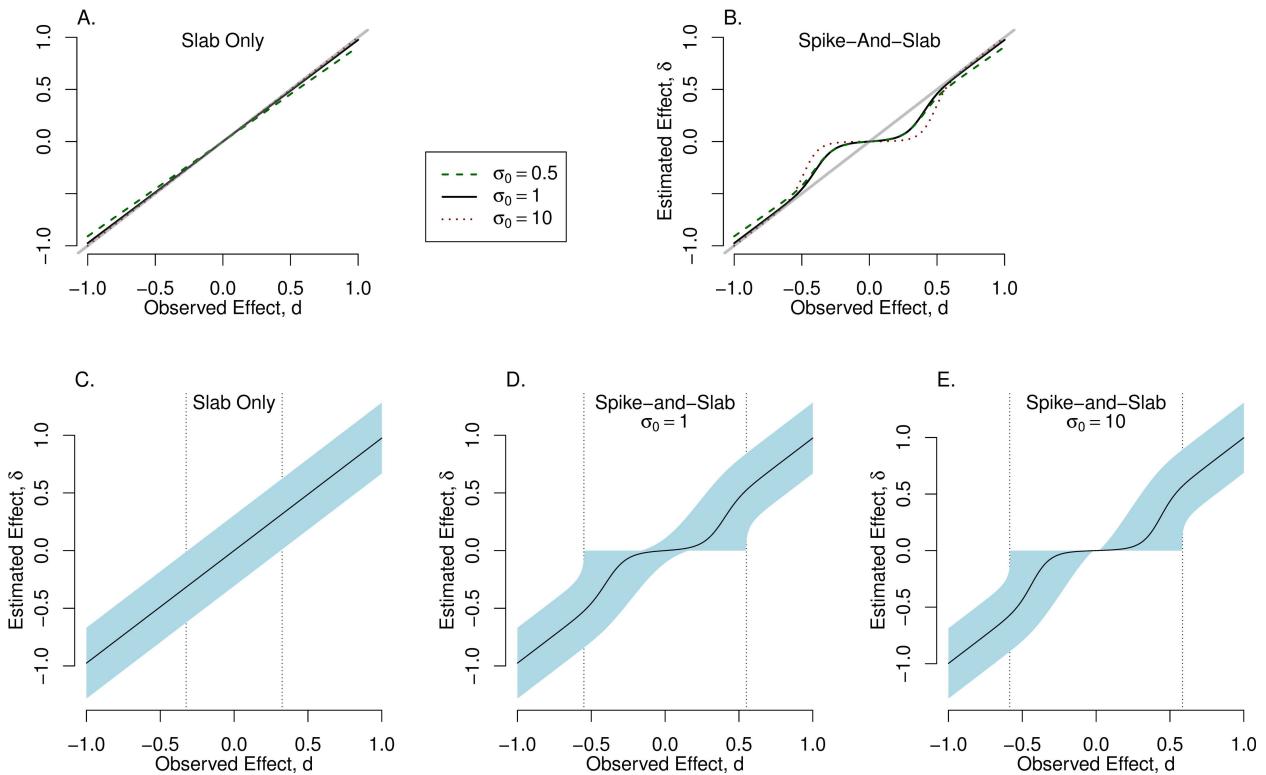


Figure 5.5: A comparison of slab-only (\mathcal{M}_1) and spike-and-slab (\mathcal{M}_s) specifications for a moderate sample size of $N = 40$. **A–B:** Posterior mean of δ as a function of d for a few prior settings of σ_0^2 . The light grey line is the diagonal, and the posterior mean of the slab-only model approaches this diagonal as the prior becomes more diffuse. The posterior mean in the spike-and-slab model shows *adaptive shrinkage* where small values observed values result in greatly attenuated estimates. **C–E:** The posterior means with 95% credible intervals. The vertical lines denote transition points—the credible interval does not include zero when the observed effect size is more extreme than these points. The transition points are more extreme for the spike-and-slab specification than the slab-only specification, and this fact is a direct consequence of the point-mass at zero.

measurement contexts, however, should avoid drawing inferences about whether or not there are effects in the data as the model specification does not capture such abstractions. There will be some differences among researchers as to which specification is best in any given context. These differences should be welcomed as they are part of the richness of adding value in psychological science (Rouder, Morey, & Wagenmakers, 2016). In all cases, however, researchers should justify their choices in the context of these goals.

Researchers who consider Bayes factors may worry about their dependence on prior settings especially when compared to estimation with slab-only models. This worry is assuredly overstated, and a bit of common sense provides for a lot of constraint. It seems to us unreasonable to consider prior settings that are too small or too large as researchers generally know that true effect sizes in psychological experiments are neither arbitrarily small or large. A lower limit of σ_0 is perhaps 0.2 as researchers rarely search for effect sizes smaller than

this value and the practical value of such small effects will often be low.⁶ Likewise, an upper limit is perhaps 1.0 as the vast majority of effect sizes are certainly smaller than this value and effects much larger than that would often be clear in day-to-day life. Within these reasonable—if context-dependent—limits, Bayes factors vary but not arbitrarily so. We have highlighted the Bayes factor values associated with these limits in Figure 5.3A as filled circles. Here the Bayes factors differ from 1.7 to 2.8 or by about 40%. This variation is not too substantial, and in both cases the evidence for an effect is marginal. Such variation strikes us as entirely reasonable and part-and-parcel of the normal variation in research Rouder et al. (2016). It is certainly less than other accepted sources such as variation in stimuli, operationalizations, paradigms, subjects, interpretations and the like.

The Potential of Spike-And-Slab Models In Psychology

We think spike-and-slab priors are going to gain popularity as psychologists develop and adopt new analytic techniques, especially in big-data applications. Consider applications in imaging where there are a great many voxels or in behavioral genetics where there are a great many nucleotide markers in a SNP array. It is desirable to consider the activity in any one voxel or the contribution to behavior of any one marker, and the resulting models necessarily have a large numbers of parameters, say with one parameter for each voxel or each marker. It is in this context, when there are large numbers of parameters especially relative to the sample size, spike-and-slab priors have become an invaluable computational tool for assessing effects, say which voxels are active or which alleles covary with a behavior. The seminal article for assessing covariates in this context is George and McCulloch (1993), and recent conceptual and computational advances, say from Scott and Berger (2010) and Ročková and George (2014), make the approach feasible in increasing large big-data contexts.

As an example of big-data applications in psychology, we highlight the recent work of Sanyal and Ferreira (2012) who used spike-and-slab priors for fMRI analysis. These researchers sought to enhance the spatial precision of imaging by improving the spatial smoothing. Typically, researchers smooth the image by passing a Gaussian filter over it. Sanyal and Ferreira instead performed a wavelet decomposition where activation is represented as having a location and a resolution. In this approach there is a separated wavelet coefficient for each resolution and location pairing, and the upshot is a proliferation of coefficients. Sanyal and Ferreira placed a spike-and-slab prior on these coefficients, and used large values of ρ_0 , the prior probability that a coefficient is zero. In analysis, the posterior for many of these coefficients remained dominated by the spike, and could be removed. When the activation was reconstructed from only the coefficients for which there was substantial mass from the slab, the image had improved quality. The resulting smoothing was adaptive—it was more smooth where activation was spatially homogenous (say within structures) and less smooth where activation was spatially heterogeneous (say at boundaries).

⁶Which is not to say that small effects cannot be *theoretically* meaningful in certain contexts, but we believe interest in very small effects to be generally low.

Conclusions

In this paper we provide a unification between two competing Bayesian approaches—that based on the estimation of posterior intervals and that based on Bayes factors. A salient difference between these two approaches is in model specification. It is common in estimation approaches to place broad priors over parameters that give no special credence to a zero point. Common Bayes factor approaches, such as that from Rouder and Morey and colleagues (Rouder et al., 2009; Rouder & Morey, 2012; Rouder et al., 2012; Guan & Vandekerckhove, 2015) are closely related to estimation with a prior that has some point mass at zero. Which model specification a researcher should choose, whether a broad slab or a spike-and-slab, should depend on the context and goals of the analyst.

References

- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Quantitative Psychology and Assessment*.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236, 119–127.
- Finetti, B. de. (1974). *Theory of probability* (Vol. 1). New York: John Wiley and Sons.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439-453.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman and Hall.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 57-64.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881-889.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. London: Cambridge.
- Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (p. 361-379,).
- Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.
- Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6, 299–312.
- Kruschke, J. K. (2011b). *Doing Bayesian analysis: A tutorial with R and BUGS*. Academic Press.
- Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*.

- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88, 1242-1249.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187-192.
- Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: Inference*. Cambridge, England: Cambridge University Press.
- Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83, 1023-1032.
- Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, -.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111-163.
- Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, 12, 573-604.
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877-903.
- 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., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological sciencecollabra. *Collabra*, 2, 6.
- 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 and Review*, 16, 225-237.
- Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68, 587-604.
- Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, 15(1201-1208).
- Ročková, V., & George, E. L. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, 109.
- Sanyal, N., & Ferreira, M. A. R. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *Neuroimage*, 63, 1519-1531.
- Scott, J. G., & Berger, J. O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38, 2587-2619.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response time. *Psychological Methods*, 16, 44-62.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, 14, 779-804.

II

Teaching resources

Four reasons to prefer Bayesian over orthodox statistical analyses

Zoltan Dienes and Neil McLatchie

Introduction

This paper will present case studies from real research that illustrate how significance testing and Bayesian statistics can lead researchers to draw different conclusions. The question will be, which conclusions are most sensible? First we will discuss the nature of hypothesis testing; then the anatomy of a Bayes factor, focusing on how one models the theory. Finally, the heart of the paper will be a set of five case studies taken from a recent special replication issue of the journal *Social Psychology*.

The nature of hypothesis testing

In using inferential statistics to test a theory of scientific interest, the world is typically first divided into \mathcal{H}_0 (the null hypothesis) and \mathcal{H}_1 (the alternative hypothesis), where one of those hypotheses is a consequence of the theory. Then data are collected in order to evaluate \mathcal{H}_0 and \mathcal{H}_1 . In evaluating whether the theory survived the test, it would often be useful to say whether the data provided good enough evidence for \mathcal{H}_0 ; good enough evidence for \mathcal{H}_1 ; or else failed to discriminate the hypotheses. That is, one might like to make a three-way distinction, as indicated in Figure 6.1a. How could that distinction be made? According to a key intuition, and one that can be readily formalized, evidence is strongest for the theory that most strongly predicted it (Good, 1983; Morey, Romeijn, & Rouder, 2016). Thus, to make the distinction between the three evidential states of affairs, one needs to know what each hypothesis predicts. Explicitly specifying predictions can be described as a 'model'.

In significance testing, one models \mathcal{H}_0 and not \mathcal{H}_1 . A typical model for \mathcal{H}_0 is, for example, the model that there is no population difference in means. Assuming in addition a model of the data (e.g. that the data are normally distributed), the probability of the data given \mathcal{H}_0 can be calculated. Unfortunately modelling \mathcal{H}_0 but not \mathcal{H}_1 does not allow one to make a three-way distinction. How can one know by which hypothesis the data are better predicted,

(a) States of evidence

Evidence for \mathcal{H}_0	No evidence to speak of	Evidence for \mathcal{H}_1
------------------------------	-------------------------	------------------------------

(b) What p -values provide

Evidence for \mathcal{H}_0	No evidence to speak of	Evidence for \mathcal{H}_1
------------------------------	-------------------------	------------------------------

NO MATTER WHAT THE p -VALUE, NO DISTINCTION
MADE WITHIN THIS BOX

(c) What Bayes factors provide

0 ... 1/3	1/3 ... 3	3 ...
-----------	-----------	-------

Evidence for \mathcal{H}_0	No evidence to speak of	Evidence for \mathcal{H}_1
------------------------------	-------------------------	------------------------------

Figure 6.1: (a) States of evidence. (b) What p -values provide. (c) What Bayes factors provide.

if one only knows how well the data are predicted by one of the hypotheses? Thus significance testing only allows a weak form of inference; it tells us something but not all that we want. As shown in Figure 6.1b, p -values only allow one to distinguish evidence against \mathcal{H}_0 from the other two evidential states of affairs (to the extent that p -values allow an evidential distinction at all¹). The p -value, no matter how large it is, in no way distinguishes good evidence for \mathcal{H}_0 from not much evidence at all. (A large p -value may result from a large standard error: A large standard error means the data do not have the sensitivity to discriminate competing hypotheses.)

To remedy the problem, it might seem obvious one needs a model of \mathcal{H}_1 (Dienes, 2016; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2015). The hypothesis testing of Neyman and Pearson (as opposed to the significance testing of Fisher) tries to model \mathcal{H}_1 in a weak way (Dienes, 2008). Hypothesis testing uses power calculations. Typically, when researchers use power they indicate what effect size they expect given their theory, perhaps based on the estimate provided by a past relevant study. Giving a point estimate of the effect size is one way of quantifying \mathcal{H}_1 . But what is the model of \mathcal{H}_1 ? In most contexts the researcher does not believe that that precise effect size is the only possible one. Nor do they

¹A significant effect indicates there is evidence for at least one particular population parameter and against \mathcal{H}_0 ; but it may not be evidence for a specific theory that allows a range of population values, and so it may not be evidence for ones actual theory. This point may not be clear yet; but the examples that follow will illustrate (case study 4 in the text). The equivocation in whether a p -value can even indicate evidence against \mathcal{H}_0 and for \mathcal{H}_1 (i.e., whether it can even make the two-way distinction claimed in the text) arises because only one model is used (only that of \mathcal{H}_0 and not of \mathcal{H}_1).

typically believe that it is the minimal one allowed by the theory. Classic hypothesis testing scarcely models a relevant \mathcal{H}_1 at all.

In fact, to know how well the hypothesis predicts the data, one needs to know the probability of each effect size given the theory (Rouder et al., 2015). This is the inferential step taken in Bayesian statistics but not in classic hypothesis testing. Because classic hypothesis testing does not take this step, it cannot evaluate evidence for \mathcal{H}_1 versus \mathcal{H}_0 , and it cannot make the three-way distinction in Figure 6.1. The case studies below will illustrate.

The anatomy of a Bayes factor

A model, as the term is used here, is a probability distribution of different effects; for example, a distribution of different possible population mean differences. To determine the evidence for \mathcal{H}_1 versus \mathcal{H}_0 , one needs a model of \mathcal{H}_0 and a model of \mathcal{H}_1 . And of course, one needs a model of the data (in the context of a statistical model, this is called the likelihood). Figure 6.2 illustrates the three models needed to calculate a Bayes factor: The model of \mathcal{H}_0 , the model of \mathcal{H}_1 , and the model of the data. In this paper we will assume that \mathcal{H}_0 can be modelled as no difference (it might be a chance value, or a particular difference; conceptually such values can all be translated to “no difference”). The model of \mathcal{H}_1 depends on the theory put to test; it is a model of the predictions of that theory. Finally the model of the data, the likelihood, specifies how probable the data are given different possible population effects. The Dienes (2008) online calculator assumes a normal likelihood (and in that way is similar to many tests that users of significance tests are familiar with where it is assumed that the participants’ data are roughly normally distributed). The first and last models are typically relatively unproblematic in terms of the decisions different researchers might come to (though see, e.g., Morey & Rouder, 2011; Wilcox, 2005). In any case, the first and last models involve decisions of a similar nature in both significance testing and Bayesian statistics: Shall I test against a null hypothesis of no difference; and shall I assume that the process generating the data produces normal distributions? In the Appendix we explore another likelihood distribution one might assume in the same situation. But now we focus on the model of \mathcal{H}_1 , a key feature distinguishing Bayesian from orthodox thinking.

The model of \mathcal{H}_1

Generally, in science predictions are made from a theory via auxiliary assumptions (e.g., Popper, 1963). For example, in testing a theory about extraversion one needs to accept the hypothesis that the scale used measures extraversion. In applying conditioning theory to learning a language, one needs hypotheses about what constitutes the conditioned stimuli. And so on. In general, these auxiliary assumptions should be a) simple, and b) informed by scientific evidence where relevant. Hopefully the latter claim strikes the reader as self-evident. In just the same way, specifying \mathcal{H}_1 is the process of making predictions from a theory via auxiliary assumptions. In general, these assumptions need to be a) simple and b) informed. Hopefully, this claim strikes the reader as equally banal. Science proceeds by deriving predictions from theories in simple and informed ways; indeed in transparent ways open to critical discussion. Of course, different theories and assumptions will lead to different

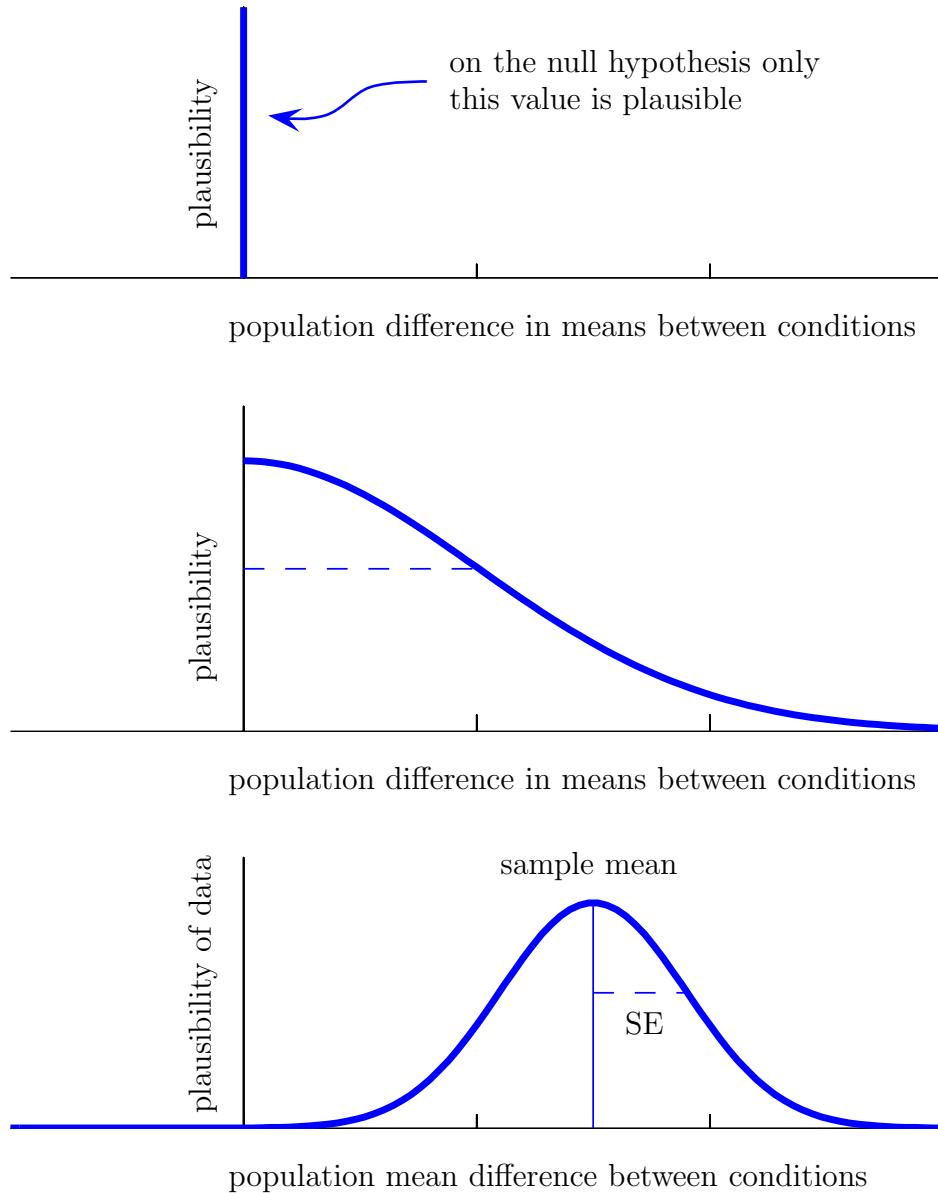


Figure 6.2: **Top:** Model of \mathcal{H}_0 . **Middle:** Model of \mathcal{H}_1 . **Bottom:** Likelihood: model of the data.

predictions. That's not a problem with science; that is how it works. Just so, Bayes factors test particular theories linked to predictions by particular assumptions (cf. Vanpaemel & Lee, 2012). A rational test could not be otherwise.

Specifying \mathcal{H}_1 makes the predictions of a scientific theory explicit. Thus, the relation of \mathcal{H}_1 to the substantial theory can be evaluated according to whether \mathcal{H}_1 was simple and scientifically informed (Dienes, 2014; Vanpaemel, 2010, 2011). One way \mathcal{H}_1 can be scientifically informed is by being based on the sort of effect size the literature claims the type of manipulation in question can produce. This is especially straightforward when the purpose of a second study is to replicate a first study (e.g., Verhagen & Wagenmakers, 2014). In that case, we expect roughly the same order of magnitude of effect as obtained in the first study. But the true population effect in the second study could be larger or smaller than the sample mean difference obtained in the first (due not only to sampling variability but also to unknown changes in conditions, moderating variables, etc.) without much changing the meaning of the first result. How much larger might the effect be? To answer this question, consider the sorts of effect sizes researchers typically investigate. On the one hand, researchers often seem interested in effects sizes with a Cohen's d around 0.5 (the modal effect size in a review of studies in many disciplines within psychology; Kühberger, Fritz, & Scherndl, 2014).² On the other hand, ds greater than about 1 are unlikely for effects that are not trivially true (Simmons, Nelson, & Simonsohn, 2013). That is, twice the expected effect might be a reasonable maximum to consider in a given scientific context. A suggested simple defeasible (i.e., over-turnable) default is: If previous work suggests a raw effect of about E , then regard effects between 0 and twice E plausible. For example, if a past study found a mean difference between conditions of 5 seconds, then for a further study (that is similar to the original), a population mean difference between 0 and 10 seconds may be plausible. (By default, we will work with raw effect sizes, e.g., seconds, because their estimates are less sensitive than standardized effect sizes, e.g., Cohen's d , to theoretically irrelevant factors like number of trials, or other factors affecting error variance alone Baguley, 2009).

We will add one more simplifying assumption about \mathcal{H}_1 . Studies that get published (and perhaps also as yet unpublished studies that catch the eye) in general over-estimate effect sizes (Ioannides, 2008; Open Science Collaboration, 2015). Thus, a defeasible default assumption is: Smaller effect sizes are more plausible than larger ones.

Putting these assumptions together, one way of representing \mathcal{H}_1 when a relevant similar effect size E (ideally in raw units) is available is illustrated in Figure 6.2, as the model for \mathcal{H}_1 . We will consider a case (as in a replication) where a directional prediction is made, that is, one condition is postulated to be greater than another. By convention we will take the difference between groups in the population to be only positive. We model the plausibility of different effects by a Half-Normal distribution (i.e., what was a normal distribution centred on zero, with the bottom half removed; so that only positive mean differences are predicted). The standard deviation of the Half-Normal is set to E . The consequences are that an effective

²Cohen's d is the raw effect size (i.e., mean difference) divided by the within-group standard deviation (Cohen, 1988). Cohen's d is useful as a signal-to-noise measure, that is, it indicates the detectability of an effect. But it should not be misinterpreted as a measure of how big an effect is (or how useful). For example, a slimming pill may have a "large" effect size as measured by Cohen's d (e.g., $d = 1.0$), but if the raw change in weight is 0.2 kg over three months, then the slimming pill may not be useful (Ziliak & McCloskey, 2008).

maximum plausible effect size is about twice E , and smaller effect sizes are more likely than larger ones. Thus the general considerations we mentioned are implemented in a simple way. Further, \mathcal{H}_1 is scientifically informed by being scaled by E . All examples that follow will use this procedure. (See Dienes, 2014, for other ways of setting up \mathcal{H}_1 .) All examples below can be worked out by the reader using the Dienes (2008) online Bayes factor calculator (see Dienes, 2014, for a tutorial; or the Dienes, 2008 website for 5-min YouTube tutorials).

Having constructed an \mathcal{H}_1 , for example by the method just described, there is a crucial final step: The judgment that the model is acceptable for the scientific problem (Good, 1983; Lindley, 2004). While a relatively default procedure is useful for constructing a possible model of \mathcal{H}_1 , in the end \mathcal{H}_1 has to be a good representation of the predictions of the scientific theory. (In the examples that follow, we judged the model of \mathcal{H}_1 generated in this way as consistent with scientific intuitions. Other researchers are free to disagree. Then we will have a scientific debate about what our theories predict and why.) The theory directly tested in each case below is that the second experiment replicated the regularity found by the first (Verhagen & Wagenmakers, 2014). As Popper (1959) pointed out, a 'result' obtained in one experiment is actually a low-level hypothesis concerning the existence of a regularity. Before we can accept that regularity (as counting for or against the substantive theory it was designed to test) we need sufficient evidence for it – as might be provided by direct replications. So the replication tests the low-level hypothesis that defines the 'result' of the first experiment. (In doing so it helps test the more general theory the results of the first experiment were regarded as bearing on, of course.) In using the effect size, E , found in the first experiment we are testing the regularity according to the explicit claims in the first paper of what the regularity is (the stated finding, where the Methods define the hypothesis concerning conditions under which the regularity obtains³).⁴

As a Bayes factor is relative to the model of \mathcal{H}_1 , we will use a subscript to specify the model of \mathcal{H}_1 (a notational convention used in Dienes, 2014, 2015). Specifically $B_{H(0,S)}$ means the Bayes factor obtained when a Half-Normal distribution (hence 'H') is used to model \mathcal{H}_1 with a mode of 0 (which we will always use for a Half-Normal) and a standard deviation of S . (Or, for example, when a Uniform distribution is used to model \mathcal{H}_1 going from a minimum of L and a maximum of M , the notation is $B_{U[L,M]}$.)

In order to illustrate both the flexibility and robustness of Bayes, the Appendix describes a different set of principles for specifying the likelihood and \mathcal{H}_1 which we will use in the examples that follow (where it is appropriate; see Appendix also for notation). This differently specified Bayes factor will be reported in footnotes. Because the scientific intuitions that it instantiates are in the cases discussed similar to the simpler procedure just described, the conclusions that follow from each model turn out to agree fairly closely in the examples that

³For biological and psychological systems, regularities will be context-sensitive. But that in no way undermines the fact that the stated Methods of a paper are a claim about the conditions under which a regularity obtains – which can be shown by the authors treating their finding as unproblematically counting for or against different theories.

⁴One can test different questions. Another relevant question is the extent to which both studies together, the original and the replication, constitute evidence for the low-level regularity. To answer this question, a Bayes factor can be performed on the meta-analytic combination of the two raw effects (cf. van Elk et al., 2015, for Bayesian meta-analyses more generally).

follow. A key difference between the models is that the t -distribution presumes the original study provides a good estimate of the effect and its uncertainty, even when transposed to a different lab; the Half-Normal presumes that the original study likely over-estimated the effect size for replication purposes.

Putting it together: the meaning of a Bayes factor

The Bayes factor provides a continuous measure of evidence for \mathcal{H}_1 over \mathcal{H}_0 . When the Bayes factor is 1, the data is equally well predicted by both models and the evidence does not favour either model over the other. As the Bayes factor increases above 1 (towards infinity) the evidence favours \mathcal{H}_1 over \mathcal{H}_0 (in the convention used in this paper). As the Bayes factor decreases below 1 (towards 0) the evidence favours \mathcal{H}_0 over \mathcal{H}_1 . There are no sharp boundaries or necessary thresholds (unlike the fixed significance levels of the Neyman Pearson approach), just a continuous degree of evidence. Nonetheless, rough guidelines can be provided, in much the same way as Cohen (1988) suggested guidelines for thinking about standardised effect sizes (researchers do not take a Cohen's d of 0.5 as a sharp cut off from small to medium effect size). Jeffreys (1939) suggested a Bayes factor of about 3 often matches the amount of evidence obtained when $p < .05$ (contrast Wetzels et al., 2011). Dienes (2014) also argued that when the raw mean difference matches that used to model \mathcal{H}_1 (a crucial condition, as we will see below), then indeed a Bayes factor of about 3 occurs when a result is just significant. That is, a Bayes factor of 3 corresponds to the amount of evidence we as a scientific community have been used to treating just worth taking note of (when the obtained effect size roughly matches that expected). Whether the scientific community understand what this means as a strength of evidence is a separate empirical question. Jeffreys suggests the label "substantial" for $B > 3$. By symmetry, we automatically get a criterion evidence for \mathcal{H}_0 over \mathcal{H}_1 : When $B < 1/3$, there is substantial evidence for \mathcal{H}_0 over \mathcal{H}_1 . We will follow this convention in reporting results below. "Substantial" means just starting to have some substance; "worth exploring further" might be a better gloss in many contexts. Another discussion worth having is whether this is good enough level of evidence; would it better to default to 6 (cf. Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015) or maybe 10 (cf. Etz & Vandekerckhove, 2016)? Etz and Vandekerckhove recommend calibrating the interpretation of the Bayes factor by studying by how much different degrees of prior belief are swayed by the evidence. This point may help calibrate the scientific community to understand what the evidence actually means. The question of the amount of evidence we should aim for is taken up further in the discussion.

We will illustrate the difference between Bayesian inference and significance testing by taking as case studies papers published in issue 3 of volume 45 of the journal *Social Psychology* (Nosek & Lakens, 2014). These papers were Registered Reports accepted in advance of the results. Thus, the obtained results have not been through a publication filter and allow a range of patterns as may be regularly obtained in research. By the same token, by restricting ourselves to one journal issue, we show the patterns we use are not so hard to find in real research. (Nonetheless, to make a point we will sometimes show what happens when the patterns are changed in instructive ways.)

Case Studies

Often significance testing will provide adequate answers

When a significant result is obtained along with an effect size matching that expected in theory, there will be evidence for \mathcal{H}_1 over \mathcal{H}_0 . Shih, Pittinsky, and Ambady (1999) argued that American Asian women primed with an Asian identity will perform better on a maths test than those primed with a female identity. There was an 11% difference in means, $t(29) = 2.02$, $p = .053$. Gibson, Losee, and Vitiello (2014) replicated the procedure with 83 subjects in the two groups (who were aware of the nature of the race and gender stereotypes); for these selected participants, the difference was 12%, $t(81) = 2.40$, $p = .02$. So there is a significant effect with a raw effect size almost identical to that in the original study. Correspondingly, $B_{H(0,11)} = 4.50$. That is, there is substantial evidence for \mathcal{H}_1 over \mathcal{H}_0 in the replication.⁵

Similarly, when a non-significant result is obtained with large N , it will often be evidence for \mathcal{H}_0 . Williams and Bargh (2008, study 2) asked 53 people to feel a hot or a cold therapeutic pack and then choose between a treat for themselves or for a friend. Seventy-five percent of participants who were exposed to physical cold selected a treat for themselves, but only 46% of the participants who were exposed to warmth did so. The strength of this relation can be expressed as an odds ratio: $OR = (75\% \times 54\%) / (46\% \times 25\%) = 3.52$. The log of the OR is roughly normally distributed; taking natural logs this gives a measure of effect size, that is, $\ln OR = 1.26$. Lynott et al. (2014) attempted a replication with total $N = 861$ people, a sample size a factor of 10 higher than the original study. The results went somewhat in the opposite direction, $OR = 0.77$, so $\ln OR = -0.26$, with a standard error of 0.14.⁶ So $z = 0.26/0.14 = 1.86$, $p = .062$, which is non-significant. Correspondingly, $B_{H(0,1.26)} = 0.04$, indicating substantial evidence for the null hypothesis over the hypothesis defined by the effect obtained in the original study.

In sum, we considered a case where a significant result corresponded with the convention for substantial evidence for \mathcal{H}_1 over \mathcal{H}_0 ; and a case where a non-significant result corresponded to the convention for substantial evidence for \mathcal{H}_0 over \mathcal{H}_1 . Correspondingly, Jeffreys (1939, pp. 323-325) discusses how in the research problems he has investigated, Fisher's methods (i.e., significance testing) and his (using Bayes factors) generally agreed (and hence indicating that the respective conventions were roughly aligned). It is in fact reassuring that the methods will often agree; when different methods with clear rationales converge they support each other. Jeffreys puts the agreement down to Fisher's insight allowing him to patch together solutions that happen to often give the right answer. Jeffreys argues that the advantage of the Bayesian system, on the other hand, is that it is one coherent system that can be derived from first principles. It explains why significance testing is right in those cases where it gives the right answer. But it also tells us why significance

⁵We can also model H1 using the t-distribution method; $B_{t(11,5.4,29),L=t(12,5,81)} = 11.12$, also indicating substantial evidence for the relevant \mathcal{H}_1 over \mathcal{H}_0 .

⁶Lynott et al. (2014) provide a confidence interval for the OR: 95% CI = [.58, 1.02]. Taking natural logs, these limits are [-0.54, 0.02]. Notice these limits are symmetric around the $\ln OR(-0.26)$, spanning 0.28 either side. Because $\ln OR$ is normally distributed, the standard error is thus $0.28/1.96 = 0.14$.

testing is wrong when it gives the wrong answer – or no clear answer at all. We now consider actual cases where Bayesian analyses give a different answer than the conventional analyses. Our aim is to provide the reason why the conventional answer is flawed, so it can be seen why the Bayesian answer is preferable in these cases.

A high powered non-significant result is not necessarily sensitive

Banerjee, Chatterjee, and Sinha (2012, study 2) found that people asked to recall a time that they behaved unethically rather than ethically estimated the room to be darker by 13.30 Watts, $t(72) = 2.70$, $p = .01$. Brandt, IJzerman, and Blanken (2014, lab replication) tried to replicate the procedure as closely as possible, using $N = 121$ participants, sufficient for a power (to pick up the original effect) greater than 0.9.

Brandt et al. (2014) obtained a difference of 5.5 Watts, $t(119) = 0.17$, $p = 0.87$. That is, it was a high-powered non-significant result. By the canons of classic hypothesis testing one should accept the null hypothesis. Yet Brandt et al. sensibly concluded “... we are hesitant to proclaim the effect a false positive based on our null findings, ... Instead we think that scholars interested in how morality is grounded should be hesitant to incorporate the studies reported by BCS into their theories until the effect is further replicated,” (p. 251). Why is this conclusion sensible if the non-significant outcome was high powered? Because a study having high power does not necessitate it has much evidential weight, and researchers should be concerned with evidence (e.g., Dienes, 2016; Wagenmakers et al., 2015). The obtained mean difference by Brandt et al. (5.5 Watts) was almost exactly half-way between the population value based on \mathcal{H}_0 (0 Watts) and the value obtained in the original study (13 Watts, which may therefore be the most likely value expected on \mathcal{H}_1). An outcome half-way between the predictions of two models cannot evidentially favour either model. As a high-powered study can produce a sample mean half between \mathcal{H}_0 and the value highly predicted by \mathcal{H}_1 , it follows that as a matter of general principle, high power does not in itself mean sensitive evidence.

Of course, \mathcal{H}_1 does not really predict just one value. Using our standard representation of plausible effect sizes, a Half-Normal scaled by the original effect size (i.e. allowing effect sizes between very small and twice the original effect), we get $B_{H(0,13.3)} = 0.97$.⁷ That is, the data do not discriminate in any way between \mathcal{H}_0 and \mathcal{H}_1 , despite the fact the study was high powered. Power can be very useful as a meta-scientific concept (e.g. Button et al., 2013; Ioannidis, 2005), but not for evaluating the evidential value of individual studies.

A low-powered non-significant result is not necessarily insensitive

Now we consider a converse case. Shih et al. (1999) argued that American Asian women primed with an Asian identity will perform better on a maths test than unprimed women; indeed, in the sample means priming showed an advantage of 5% more questions answered correctly.⁸ Moon and Roeder (2014) replicated the study, with about 50 subjects in each

⁷We can also model H_1 using the t-distribution method; $B_{t(13.3, 4.93, 72), L=t(5.47, 32.2, 119)} = 0.97$, giving exactly the same answer as the Bayes factor in the text.

⁸This difference was not tested by inferential statistics.

group; power based on the original $d = 0.25$ effect is 24%. Given the low power, perhaps it is not surprising that the replication yielded a non-significant effect, $t(99) = 1.15$, $p = 0.25$. However, it would be wrong to conclude that the data were not evidential. The mean difference was 4% in the wrong direction according to the theory. When the data go in the wrong direction (by a sufficient amount relative to the standard error), they should carry some evidential weight against the theory. Testing the directional theory by modelling \mathcal{H}_1 as a Half-Normal with a standard deviation of 5%, $B_{H(0,5)} = 0.31$, substantial evidence for the null relative to the \mathcal{H}_1 .⁹

Note that a sample difference going in the wrong direction is not necessarily good evidence against the theory (Dienes, 2015). If the standard error is large enough, the sample mean could easily go in the wrong direction by chance even if the population mean is in the theoretically right direction.¹⁰

A high-powered significant result is not necessarily evidence for a theory

Imagine two theories about earthquakes, theory A and theory B, being used to predict whether an earthquake will happen in downtown Tokyo on a certain week. Theory A predicts an earthquake only on Tuesday between 2 and 4 pm of a magnitude between 5 and 6. Theory B predicts earthquakes any time between Monday and Saturday anywhere between 1 (non-existent) to 7 (intense). Theory A makes a precise prediction; theory B is vague and allows just about anything. An earthquake in fact happens on Tuesday around 2:30pm of magnitude 5.1. These data are in the predicted range of both theories. Nonetheless, does this observation count as stronger evidence for one theory rather than the other? Would you rely on one of those theories for future predictions more than the other in the light of these data?

It should be harder to obtain evidence for a vague theory than a precise theory, even when predictions are confirmed. That is, a theory should be punished for being vague. If a theory allows many outcomes, obtaining one of those outcomes should count for less than if the theory allows only some outcomes (Popper, 1959). Thus, a just significant result cannot provide a constant amount of evidence for an \mathcal{H}_1 over \mathcal{H}_0 ; the relative strength of evidence must depend on the \mathcal{H}_1 . For example, a just significant result in the predicted range should count for less for an \mathcal{H}_1 modelled as a normal distribution with a very large rather than small standard deviation. A significant result with a small sample effect size might not be evidence

⁹As before, the effect can also be tested modelling H_1 as a t -distribution with a mean equal to the original mean difference (5%) and SE equal to the original SE of that difference (estimated as 14%). $B_{t(5,14,30),L=t(-4,3.48,99)} = 0.38$. The value is close to the Bayes factor based on the Half-Normal provided in the text. If the original effect had actually been just significant (so setting its SE to 2.5, and keeping everything else the same), then $B_{t(5,2.5,30),L=t(-4,3.48,99)} = 0.18$, sensitive evidence in discriminating \mathcal{H}_0 from \mathcal{H}_1 .

¹⁰Imagine Moon and Roeder (2014) obtained the same mean difference, 4%, but the standard error of this difference was twice as large. (Thus, t would be half the size, i.e., we would have $t(99) = 0.58$, $p = .57$ for the replication.) Now we have $B_{H(0,5)} = 0.63$, with not enough evidence to be worth mentioning one way or the other. Using the t -distribution method, $B_{t(5,14,30),L=t(-4,6.96,99)} = 0.44$. The value is close to the Bayes factor based on the Half-Normal. A mean difference going in the wrong direction does not necessarily count against a theory.

at all for a theory that allows a wide range of effect sizes (see Lindley, 1957; Wagenmakers, Lee, Rouder, & Morey, 2014).

The issue can be illustrated using Lynott et al.'s (2014) replication of Williams and Bargh (2008, study 2). As we described above, Williams and Bargh asked 53 people to feel a hot or a cold therapeutic pack and then choose between a treat for themselves or for a friend. Seventy-five percent of participants exposed to the physical cold selected a treat for themselves, whereas only 46% of participants exposed to the physical warmth did so, with $\ln \text{OR} = 1.26$ (just significant, $p < .05$). Lynott et al. (2014) obtained non-significant results with a larger sample. Imagine that Lynott et al found that 53.5% of people exposed to cold chose the personal reward while only 46.5% of those exposed to warmth did so resulting in an $\ln \text{OR}$ of 0.28, which, given the same standard error as Lynott et al. actually obtained (0.14), gives $p < .05$. However now $B_{H(0,1.26)} = 1.56$, indicating the data are insensitive in discriminating \mathcal{H}_1 from \mathcal{H}_0 .

How can a significant result not count in favour of a theory that predicted a difference? It depends on the theory being tested. The original finding was that 75% of people exposed to cold selected a personal treat (and only 46% exposed to warmth did so); if one could expect an effect size from very small to even larger than this, then a small effect size is not especially probable in itself¹¹. The theory is vague in allowing a wide range of effect sizes. So while 53% compared to 46% choosing a personal reward may be somewhat unlikely on \mathcal{H}_0 , it turned out to be just as unlikely on \mathcal{H}_1 (cf. Lindley, 1993). Vague theories are rightly punished by Bayesian analyses; by contrast, the p -value is indifferent to the inferentially-relevant feature of a theory being vague. So call this model of \mathcal{H}_1 the vague model.

Let us say in the original study 55% of people exposed to cold chose the personal reward whereas 45% of people exposed to warmth did so, and this was significant $p = .049$. Now $\text{OR} = (552/452) = 1.49$, and $\ln \text{OR} = 0.40$. These data render a $\ln \text{OR}$ greater than about twice 0.40 as quite unlikely (in that they fall outside a 95% credibility interval). The theory is more precise (than when effects up to about twice 1.26 were allowed). Call the model of \mathcal{H}_1 based on these counterfactual results the precise model. Finding a replication $\ln \text{OR}$ of 0.28 (with a standard error of 0.14 as before), falls within the range of predictions of this rather precise theory, just as it fell within the range of predictions of the vague theory. Now $B_{H(0,0.40)} = 3.81$, support for the precise \mathcal{H}_1 over \mathcal{H}_0 (the B was 1.56 for the vague \mathcal{H}_1 over \mathcal{H}_0). Bayes factors are sensitive to how vague or precise the theory is; p -values are not. But, normatively, precise theories should be favoured over vague ones when data appear within the predicted range.

Finally, notice that the replication study had less power to distinguish the $\ln \text{OR}$ of 0.40 (the value used for deriving the precise model) from \mathcal{H}_0 than it had to distinguish the $\ln \text{OR}$ of 1.26 (the value used for deriving the vague model) from \mathcal{H}_0 . In this case, the high powered significant result was less good evidence for the theory than the low powered significant result. A high-powered significant result is not necessarily evidence for a theory. How strong the evidence is for a theory all depends on how well the theory predicted the data.

The answer to the question should depend on the question

Jeffreys (1939, p. vi) wrote that “It is sometimes considered a paradox that the answer depends not only on the observations, but also on the question; it should be a platitude.” The point was illustrated in the last case study. The same data provide less evidence for a vague theory than a precise theory when the data fall in the predicted range. Same data, different answers – because the questions are different. Yet although the questions were different, significance testing was only capable of giving one answer. For other examples, Bayes factors can test \mathcal{H}_1 against interval or other non-point null hypotheses (Dienes, 2014; Morey & Rouder, 2011) or one substantial \mathcal{H}_1 against another, instead of against \mathcal{H}_0 (for example, the theories that differences are positive versus negative; or in general theories that allow a different range of effects).

The issue often comes up as a criticism of Bayes factors (e.g. Kruschke, 2013; Kruschke & Liddell, this volume): the answer provided by the Bayes factor is sensitive to the specification of \mathcal{H}_1 , so why should we trust the answer from a Bayes factor? We will illustrate with the following example. Schnall, Haidt, Clore, and Jordan (2008) found that people make less severe judgments on a 1 (perfectly OK) to 7 (extremely wrong) scale when they wash their hands after experiencing disgust (Exp. 2). Of the different problems they investigated, taken individually, the wallet problem was significant, with a mean difference of 1.11, $t(41) = 2.57$, $p = .014$. Johnson, Cheung, and Donnellan (2014, study 2) replicated with an N of 126, giving a power of greater than 99% to pick up the original effect. The obtained mean difference was 0.15, $t(124) = 0.63$, $p = 0.53$. Thus, there is a high-powered non-significant result. But, as is now clear, that still leaves open the question of how much evidence there is, if any, for \mathcal{H}_0 rather than \mathcal{H}_1 .

One could argue that the 1-7 scale used in the replication allows differences between groups between a minimum of 0 and a maximum of 6 (the maximum population mean that one group could have is 7 and the minimum for the other group is 1, giving a maximum difference of 6). The predictions of \mathcal{H}_1 could be represented as a uniform distribution from 0 to 6. That claim has the advantage of simplicity, as it can be posited without reference to data. These considerations give $B_{U[0,6]} = 0.09$. That is, there is substantial evidence for \mathcal{H}_0 over this \mathcal{H}_1 .

We also have our Half-Normal model for representing \mathcal{H}_1 . The original raw effect size was 1.11 rating units; and, $B_{H(0,1.11)} = 0.3712$. That is, the data do not very sensitively distinguish \mathcal{H}_0 from this \mathcal{H}_1 .

So we have one Bayes factor of 0.09 and another of 0.37. Both Bayes factors have a reasonable rationale. Yet they are sufficiently different that they may lead to different conclusions in a Discussion section, and different interpretations of what the replication meant. This situation might seem to be a damning criticism of Bayes factors. In fact, it shows Bayes factor behave as a measure of evidence should.

Each Bayes factor is an indication of the evidence for the \mathcal{H}_1 represented as opposed to \mathcal{H}_0 . The \mathcal{H}_1 s are different, and each Bayes factor appropriately answers a different question. Which Bayes factor answers the question we have been asking in this paper for each case study, namely the extent to which the replication provided evidence for the regularity claimed by the first study? The first Bayes factor is not good at answering this question, because it is

not informed by the first study. The second Bayes factor is informed (and is otherwise simply specified). Therefore, the second Bayes factor is the one that should be used to address this question, and thus guide the corresponding discussion and conclusions in the paper.

The first Bayes factor in effect refers to a different theory, and thus poses a different question of the data. That theory predicted all differences as equally plausible. It is a vague theory and thus was not supported as well as the more precise theory defined by the effect found in the original study. But theories, or models of data, need not differ just in being vague versus precise. Two models could be just as precise but predict different size effects. The Half-Normal model we have been using does not allow this (as predictions are changed only by changing the SD of the distribution and hence its vagueness); but the *t*-distribution described in the Appendix does. One alternative hypothesis, \mathcal{H}_1 , might predict an effect around E_1 and another alternative, \mathcal{H}_2 , an effect just as tightly around E_2 . If the data were close to E_2 and far from E_1 , \mathcal{H}_2 would be supported better than \mathcal{H}_1 – but the *p*-value testing against \mathcal{H}_0 would be the same.

A Bayes factor is a method for comparing two models. Thus there is not one Bayes factor that reflects what the data mean. In comparing \mathcal{H}_1 to \mathcal{H}_0 , the answer depends on what \mathcal{H}_1 and \mathcal{H}_0 are. That's not a problem, any more than in comparing two scientific theories, the answer depends on what the theories are. Further, the use of Bayes factors in no way precludes estimating parameters, or deriving credibility intervals, in order to understand the data. Both model comparison (hypothesis testing) and parameter estimation are important and complementary aspects of the scientific process (Jeffreys, 1939).

Discussion

The aim of the paper is to illustrate how significance testing and Bayesian inference may lead researchers to draw different conclusions in certain cases, and to show why the Bayesian conclusion is the preferred one. Specifically, we considered four types of scenarios. First, researchers may believe that a high-powered non-significant result necessarily means one has good evidence for \mathcal{H}_0 . We showed that in actual situations, high power does not guarantee sensitive evidence for \mathcal{H}_0 rather than \mathcal{H}_1 . Conversely, it might be thought that “power just is not demanding enough; but that means a low-powered non-significant result guarantees the evidence for \mathcal{H}_0 is weak.” But this second intuition turns out to be false as well. A low-powered result may be substantial evidence for \mathcal{H}_0 rather than \mathcal{H}_1 . Thus nothing about the evidential value of a non-significant result follows from the mere fact that study was low or high powered. Thus, classic hypothesis testing does not allow one to distinguish three evidential states of affairs, namely evidence for \mathcal{H}_0 rather than \mathcal{H}_1 , evidence for \mathcal{H}_1 rather than \mathcal{H}_0 , or not much evidence either way. By contrast, Bayes factors do allow this three-way distinction.

The researcher might conclude that she always suspected that non-significant results were problematic anyway. But, she might feel, with significant results we are on firmer ground. However, in the third contradiction, we found that a high-powered significant result may not actually be good evidence for \mathcal{H}_1 rather than \mathcal{H}_0 . If \mathcal{H}_1 is sufficiently vague, the significant result may be unlikely under the theory. And, in the fourth scenario, we found that in

general the strength of evidence for \mathcal{H}_1 rather than \mathcal{H}_0 depends on what the \mathcal{H}_1 is, a sensible state of affairs that a p -value cannot reflect.

While in the examples we have used $B > 3$ (or $< 1/3$) as a criterion for sufficient evidence to draw a conclusion, we have done so merely because that roughly matches the standard of evidence the psychology community has been using up to now. However, our aim has been to advocate using a genuine measure of evidence, which is different from advocating a particular degree of evidence as sufficient. A conjecture is that the current standard of evidence has arisen for psychological reasons, namely it is a point where researchers typically judge that evidence is just enough to be worth taking notice of. (Compare the equivalent two-sigma, i.e., $t = 2$, criterion in the particle physics community, a criterion which means “maybe there is something there” (see, e.g., Gibney, 2016). Five-sigma in that community is taken as warranting a conclusion, which would be closer to $B = 5 \times 10^4$.) Because $B = 3$ is typically around the borderline of what is worth taking note of, analytic flexibility could push conclusions around when $B = 3$ is used as a threshold (see, e.g., Dienes, 2016). Schönbrodt et al. (2015) recommend using $B = 6$ as a conventional threshold; Morey (2015) recommends negotiating the threshold for each particular case. However a threshold of evidence for reaching a decision by a journal or scientists is chosen, it is important that the threshold is seen as only a useful convention, while bearing in mind that what the Bayes factor actually shows is a continuous degree of evidence.

In this paper we have focused on examples that involve direct replications. The same principles apply for calculating Bayes factors in other situations; Dienes (2014, 2015) gives examples of specifying the model of \mathcal{H}_1 in ANOVA, regression and contingency table cases.

The role of Bayes factors in addressing problems with how research is conducted goes beyond the issues discussed here. For example, the role of Bayes factors in experiments with optional stopping is discussed by Rouder (2014) and Schönbrodt et al. (2015); the role of Bayes factors in addressing these and other issues involved in the “credibility crisis” in psychology (e.g. Open Science Collaboration, 2015) and other sciences is discussed by Dienes (2016) and the reproducibility project in particular by Etz and Vandekerckhove (2016); Guan and Vandekerckhove (2016) introduce a Bayesian method for mitigating publication bias; and Lee and Wagenmakers (2013) and Vanpaemel and Lee (2012) describe Bayesian methods for incorporating more theory into models in testable ways.

What is the way forward? We suggest a community learning process in which orthodox statistics are reported, but along with the orthodox statistics such as F s and the p s, B s are reported as well (see, e.g., Ziori & Dienes, 2015, for a paper illustrating this policy). Interpretation can be done with respect to the B s – and in many cases a p -aficionado may agree with the conclusion (e.g. as in Ziori & Dienes, 2015). On the one hand, distinctions would be drawn not available to the p -aficionado, and more informed decisions taken. On the other hand, a significant p -value at the 5% level indicates there is some way of specifying \mathcal{H}_1 such that $B > 3$ (Royall, 1997), which may be worth considering. In the process of implementing “a B for every p ,” we as a community would learn to see the relationship between significance testing and Bayes factors – and, crucially, come to debate the optimal Bayesian ways of addressing different research questions.

References

- Baguley, T. (2009). Standardized or simple effect size: what should be reported? *British Journal of Psychology*, 100, 603-617.
- Baguley, T., & Kaye, W. S. (2010). Review of understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical & Statistical Psychology*, 63, 695-698.
- Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? *Recalling moral behavior changes perception of brightness*, 23, 407-409.
- Berry, D. A. (1996). *Statistics: A Bayesian perspective*. Duxbury Press.
- Brandt, M. J., IJzerman, H., & Blanken, I. (2014). Does recalling moral behavior change the perception of brightness? *Social Psychology*, 45, 246-252.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365-376.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. palgrave macmillan. Website for associated online Bayes factor calculator.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781.
- Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (p. 199-220). Oxford: Oxford University Press.
- Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78-89.
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, 11, 2.
- Gibney, E. (2016). Morphing neutrinos provide clue to antimatter mystery. *Nature*, 536, 261-262.
- Gibson, C. E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility (shih, pittinsky, & ambady, 1999): Identity salience and shifts in quantitative performance. *Social Psychology*, 45, 194-198.
- Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. University of Minnesota Press.
- Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, 23, 74-86.
- Ioannides, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640-648.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *Chance*, 18(4), 40-47.
- Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

- Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? *A Direct Replication of Schnall, Benton, and Harvey* (, 2008, 209-215.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573-603.
- Kruschke, J. K., & Liddell, T. (this volume). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*.
- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, 9(9), e105825.
- Lee, M. D., & Wagenmakers, E.-j. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1), 187-192.
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15, 22-25.
- Lindley, D. V. (2004). That wretched prior. *Significance*, 1, 85-87.
- Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., et al. (2014). Replication of “experiencing physical warmth promotes interpersonal warmth” by williams and bargh (2008) social psychology. 45, 216-222.
- Moon, A., & Roeder, S. S. (2014). A secondary replication attempt of stereotype susceptibility (shih, pittinsky. & Ambady, 1999, 199-201.
- Morey, R. D. (2015). *On verbal categories for the interpretation of bayes factors*. (Downloaded 2 Sept 2016)
- Morey, R. D., Romeijn, J.-w., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6-18.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406-419.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45, 137-141.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943-951.
- Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson: London.
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge: London.
- Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, 21, 301-308.
- Rouder, J. N., Morey, R. D., Verhagen, J. A., Province, J. M., & Wagenmakers, E.-j. (2015). *The p > .05 rule and the hidden costs of the free lunch in inference*. Manuscript submitted for publication.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman and Hall.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and social psychology bulletin*, 34(8), 1096–1109.
- Schönbrodt, F. D., Wagenmakers, E.-j., Zehetleitner, M., & Perugini, M. (2015). *Sequential*

- hypothesis testing with Bayes factors: Efficiently testing mean differences.* Psychological Methods.
- Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80-83.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after p-hacking* (Unpublished manuscript). : Retrieved 9 Nov 2015.
- van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-j. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, 6, 1365.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491-498.
- Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, 55, 106-117.
- Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, 19, 1047-1056.
- Verhagen, A. J., & Wagenmakers, E.-j. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143, 1457-1475.
- Wagenmakers, E.-j., Lee, M. D., Rouder, J. N., & Morey, R. D. (2014, Nov). Another statistical paradox. *Manuscript submitted for publication*, 9, 2015.
- Wagenmakers, E.-j., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., et al. (2015). A power fallacy. *Behavior Research Methods*, 47(4), 913-917.
- Wetzel, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6, 291-298.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing, second edition*. Academic Press: London.
- Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322, 306-307.
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error cost us jobs, justice and lives*. The University of Michigan Press: Ann Arbor.
- Ziori, E., & Dienes, Z. (2015). Facial beauty affects implicit and explicit learning of men and women differently. *Frontiers in Psychology*, 6, 1124.

Appendix A: Details for the case studies

We discuss a Bayes factor (introduced in Dienes, 2016) that uses a t -distribution to model both \mathcal{H}_1 and the likelihood, and can use raw effect sizes. First, consider \mathcal{H}_1 . We may have an estimate of the effect size we are trying to pick up based on a previous study. Verhagen and Wagenmakers (2014) suggest using the posterior distribution of the standardized effect size from the original experiment as the model of \mathcal{H}_1 . In the Half-Normal method discussed in the text, an effect size E is used to inform \mathcal{H}_1 , but no use is made of any knowledge of the uncertainty in estimating E . This makes the procedure widely applicable as a default,

precisely because no knowledge is needed of the estimate of E . On the other hand, there will be situations where it makes sense to make use of knowledge of the uncertainty in estimating E .

A common situation is where E has been derived from observations coming from roughly normal distributions but where the variance is unknown and only estimated. Given only vague information about the possible variance, the resulting posterior distribution of the effect is t -distributed (Jeffreys, 1939). Thus, \mathcal{H}_1 can be modelled as a t -distribution having a mean the same as the mean difference, an SD equal to the standard error of that difference and with degrees of freedom equal to those in the original study. We can notate the B in this way: $B_{t(\text{mean difference, SE, df})}$.

The Dienes (2008) calculator, used for the Half-Normal method in this paper, assumed a normal likelihood. However, once again, if the variance of the data is only estimated, the likelihood is best treated not as normal but as t -distributed. Adapting a procedure introduced by Berry (1996), Dienes recommended adjusting the standard error according to the degrees of freedom, because the likelihood then approximates the t -distribution. But better would be to use the t -distribution in this situation. So here the likelihood is t -distributed, and so the full notation for B is: $B_{t(\text{mean difference, SE, df}), L=(\text{mean difference, SE, df})}$. In the first brackets are the parameters of the theory, i.e., of \mathcal{H}_1 ; thus for a replication, they refer to the first study, and in the code below they are notated **meanoftheory** (i.e., the raw mean difference for study 1), **sdtheory** (i.e., the SE of that difference from study 1) and **dftheory** (i.e., the degrees of freedom from study 1). The brackets after the L refer to the replication study and in the code below are notated **obtained** (i.e., the mean difference in the replication study), **sd** (the standard error of that difference) and **dfdata** (the degrees of freedom of the replication study). So we have $B_{t(\text{meanoftheory, sdtheory, dftheory}), L=(\text{obtained, sd, dfdata})}$. The R code is based on that originally provided by Baguley and Kaye (2010) and for the Dienes (2008) calculator.

The case studies reported in the main text were analysed both by modelling \mathcal{H}_1 with Half-Normal distributions and, where standard deviations were estimated from data, by a t -distribution for modelling \mathcal{H}_1 , as shown in Table 6.1. As it happened, the two types of Bayes factor produced similar degrees of evidence for their \mathcal{H}_1 s versus \mathcal{H}_0 . However, they do have different properties, discussed throughout the text, which we summarize here. First, it is typical for the t -method rather than the Half-Normal method, to give more evidence for \mathcal{H}_0 when the sample mean is close to 0 – because the Half-Normal method loads plausibility around 0, typically making the models harder to distinguish than with the t -method (see footnote 12). Second, the larger the estimated effect, the vaguer the theory modelled by the Half-Normal method. This may be reasonable when effects are just significant. However, for the t -method size of effect and vagueness of theory can be represented independently. This is useful when a large effect has been estimated with high precision, and we believe small effects are unlikely. Finally, the t -method involves taking the posterior distribution of the effect seriously for predicting the effect in a new study. This approach is most plausible for direct replications. In many other cases, the uncertainty in the estimate as an estimate for a new study would be broader than that given by the posterior distribution for the original study. The Half-Normal provides a simple default for such situations.

Table 6.1: The case studies with Bayes factors (B) based on either the Half-Normal or the t -distribution.

Gibson et al. (2014)

Raw effect:	12%	
Significance test:	$t(81) = 2.40, p = .02$	
Half-Normal \mathcal{H}_1 :	$H(0, 11)$	$\rightarrow B = 4.50$
t -distribution \mathcal{H}_1 :	$t(11, 5.4, 29), L = t(12, 5, 81)$	$\rightarrow B = 7.84$

Brandt et al. (2014)

Raw effect:	5.5 Watts	
Significance test:	$t(119) = 0.17, p = .87$	
Half-Normal \mathcal{H}_1 :	$H(0, 13.3)$	$\rightarrow B = 0.97$
t -distribution \mathcal{H}_1 :	$t(13.3, 4.93, 72), L = t(5.47, 32.2, 119)$	$\rightarrow B = 0.97$

Moon and Roeder (2014)

Raw effect:	-4%	
Significance test:	$t(99) = 1.15, p = .25$	
Half-Normal \mathcal{H}_1 :	$H(0, 5)$	$\rightarrow B = 0.31$
t -distribution \mathcal{H}_1 :	$t(5, 14, 30), L = t(-4, 3.48, 99)$	$\rightarrow B = 0.38$

Moon and Roeder (2014), counterfactually, with SE twice as large

Raw effect:	-4%	
Significance test:	$t(99) = 0.58, p = .57$	
Half-Normal \mathcal{H}_1 :	$H(0, 5)$	$\rightarrow B = 0.63$
t -distribution \mathcal{H}_1 :	$t(5, 14, 30), L = t(-4, 6.96, 99)$	$\rightarrow B = 0.44$

Johnson et al. (2014)

Raw effect:	0.15 scale points (scale 1 – 7)	
Significance test:	$t(124) = 0.63, p = .53$	
Half-Normal \mathcal{H}_1 :	$H(0, 1.11)$	$\rightarrow B = 0.37$
t -distribution \mathcal{H}_1 :	$t(1.11, 0.43, 41), L = t(0.15, 0.24, 124)$	$\rightarrow B = 0.09$

Appendix B: R code for the Bayes factor

```
Bf <- function( sd, obtained, dfdata,
                 meanoftheory, sdtheory,
                 dftheory, tail=2)
{
  area      <- 0
  normarea  <- 0
  theta     <- meanoftheory - 10 * sdtheory
  incr      <- sdtheory / 200

  for (A in -2000:2000) {
    theta <- theta + incr
    dist_theta <- dt((theta - meanoftheory) / sdtheory, df=dftheory)
    if(identical(tail, 1)) {
      if (theta <= 0) {
        dist_theta <- 0
      } else {
        dist_theta <- dist_theta * 2
      }
    }
    height   <- dist_theta * dt((obtained - theta) / sd, df=dfdata)
    area     <- area + height * incr
    normarea <- normarea + dist_theta * incr
  }

  LikelihoodTheory <- area / normarea
  Likelihoodnull   <- dt(obtained / sd, df=dfdata)
  BayesFactor       <- LikelihoodTheory / Likelihoodnull

  BayesFactor
}
```

How to become a Bayesian in eight easy steps: An annotated reading list

Alexander Etz, Quentin F. Gronau, Fabian Dablander, Peter A.

Edelsbrunner and Beth Baribault

Introduction

In recent decades, significant advances in computational software and hardware have allowed Bayesian statistics to rise to greater prominence in psychology (R. Van de Schoot, Winder, Ryan, Zondervan-Zwijnenburg, & Depaoli, *in press*). In the past few years, this rise has accelerated as a result of increasingly vocal criticism of *p*-values in particular (Nickerson, 2000; Wagenmakers, 2007), and classical statistics in general (Trafimow & Marks, 2015). When a formerly scarcely used statistical method rapidly becomes more common, editors and peer reviewers are expected to master it readily, and to adequately evaluate and judge manuscripts in which the method is applied. However, many researchers, reviewers, and editors in psychology are still unfamiliar with Bayesian methods.

We believe that this is at least partly due to the perception that a high level of difficulty is associated with proper use and interpretation of Bayesian statistics. Many seminal texts in Bayesian statistics are dense, mathematically demanding, and assume some background in mathematical statistics (e.g., Gelman et al., 2013). Even texts that are geared toward psychologists (e.g., Lee & Wagenmakers, 2014; Kruschke, 2015), while less mathematically difficult, require a radically different way of thinking than the classical statistical methods most researchers are familiar with. Furthermore, transitioning to a Bayesian framework requires a level of time commitment that is not feasible for many researchers. More approachable sources that survey the core tenets and reasons for using Bayesian methods exist, yet identifying these sources can prove difficult for researchers with little or no previous exposure to Bayesian statistics.

In this guide, we provide a small number of primary sources that editors, reviewers, and other interested researchers can study to gain a basic understanding of Bayesian statistics. Each of these sources was selected for their balance of accessibility with coverage of essential

Bayesian topics. By focusing on interpretation, rather than implementation, the guide is able to provide an introduction to core concepts, from Bayes' theorem through to Bayesian cognitive models, without getting mired in secondary details.

This guide is divided into two primary sections. The first, *Theoretical sources*, includes commentaries on three articles and one book chapter that explain the core tenets of Bayesian methods as well as their philosophical justification. The second, *Applied sources*, includes commentaries on four articles that cover the most commonly used methods in Bayesian data analysis at a primarily conceptual level. This section emphasizes issues of particular interest to reviewers, such as basic standards for conducting and reporting Bayesian analyses.

We suggest that for each source, readers first review our commentary, then consult the original source. The commentaries not only summarize the essential ideas discussed in each source, but also give a sense of how those ideas fit into the bigger picture of Bayesian statistics. This guide is part of a larger special issue in *Psychonomic Bulletin & Review* on the topic of Bayesian inference that contains articles which elaborate on many of the same points we discuss here, so we will periodically point to these as potential next steps for the interested reader. For those who would like to delve further into the theory and practice of Bayesian methods, the Appendix provides a number of supplemental sources that would be of interest to researchers and reviewers. To facilitate readers' selection of additional sources, each source is briefly described and has been given a rating by the authors that reflects its level of difficulty and general focus (i.e., theoretical versus applied; see Figure 7.2). It is important to note that our reading list covers sources published up to the time of this writing (August, 2016).

Overall, the guide is designed such that a researcher might be able to read all eight of the highlighted articles¹ and some supplemental readings within a week. After readers acquaint themselves with these sources, they should be well-equipped both to interpret existing research and to evaluate new research that relies on Bayesian methods.

Theoretical sources

In this section, we discuss the primary ideas underlying Bayesian inference in increasing levels of depth. Our first source introduces *Bayes' theorem* and demonstrates how Bayesian statistics are based on a different conceptualization of probability than classical, or *frequentist*, statistics (Lindley, 1993). These ideas are extended in our second source's discussion of Bayesian inference as a reallocation of credibility between possible states of nature (Kruschke, 2015). The third source demonstrates how the concepts established in the previous sources lead to many practical benefits for experimental psychology (Dienes, 2011). The section concludes with an in-depth review of Bayesian hypothesis testing using Bayes factors with an emphasis on this technique's theoretical benefits (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

¹Links to freely available versions of each article are provided in the *References* section.

Conceptual introduction: What is Bayesian inference?

Source: Lindley (1993) — The analysis of experimental data: The appreciation of tea and wine

Lindley leads with a story in which renowned statistician Ronald A. Fisher is having his colleague, Dr. Muriel Bristol, over for tea. When Fisher prepared the tea—as the story goes—Dr. Bristol protested that Fisher had made the tea all wrong. She claims that tea tastes better when milk is added first and infusion second,² rather than the other way around; she furthermore professes her ability to tell the difference. Fisher subsequently challenged Dr. Bristol to prove her ability to discern the two methods of preparation in a perceptual discrimination study. In Lindley’s telling of the story, which takes some liberties with the actual design of the experiment in order to emphasize a point, Dr. Bristol correctly identified 5 out of 6 cups where the tea was added either first or second. This result left Fisher faced with the question: Was his colleague merely guessing, or could she really tell the difference? Fisher then proceeded to develop his now classic approach in a sequence of steps, recognizing at various points that tests that seem intuitively appealing actually lead to absurdities, until he arrived at a method that consists of calculating the total probability of the observed result plus the probability of any more extreme results possible under the null hypothesis (i.e., the probability that she would correctly identify 5 *or* 6 cups by sheer guessing). This probability is the *p*-value. If it is less than .05, then Fisher would declare the result significant and reject the null hypothesis of guessing.

Lindley’s paper essentially continues Fisher’s work, showing that Fisher’s classic procedure is inadequate and itself leads to absurdities because it hinges upon the nonexistent ability to define what other unobserved results would count as “more extreme” than the actual observations. That is, if Fisher had set out to serve Dr. Bristol 6 cups (and only 6 cups) and she is correct 5 times, then we get a *p*-value of .109, which is not statistically significant. According to Fisher, in this case we should not reject the null hypothesis that Dr. Bristol is guessing. But had he set out to keep giving her additional cups until she was correct 5 times, which incidentally required 6 cups, we get a *p*-value of .031, which is statistically significant. According to Fisher, we should now reject the null hypothesis. Even though the data observed in both cases are exactly the same, we reach different conclusions because our definition of “more extreme” results (that did not occur) changes depending on which sampling plan we use. Absurdly, the *p*-value, and with it our conclusion about Dr. Bristol’s ability, depends on how we think about results that might have occurred but never actually did, and that in turn depends on how we planned the experiment (rather than only on how it turned out).

Lindley’s Bayesian solution to this problem considers only the probability of observations actually obtained, avoiding the problem of defining more extreme, unobserved results. The observations are used to assign a probability to each possible value of Dr. Bristol’s success rate. Lindley’s Bayesian approach to evaluating Dr. Bristol’s ability to discriminate between

²As a historical note: Distinguishing milk-first from infusion-first tea preparation was not a particular affectation of Dr. Bristol’s, but a cultural debate that has persisted for over three centuries (e.g., Orwell, 1946).

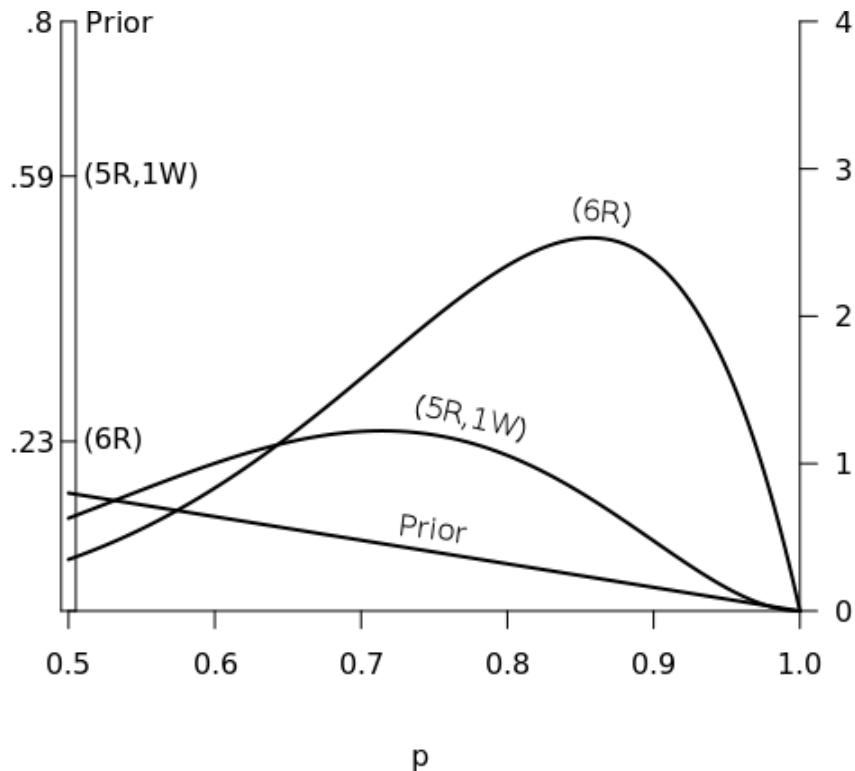


Figure 7.1: A reproduction of Figure 2 from Lindley (1993). The left bar indicates the probability that Dr. Bristol is guessing prior to the study (.8), if 5 right and 1 wrong judgments are observed (.59), and if 6 right and 0 wrong judgments are observed (.23). The lines represent Lindley's corresponding beliefs about Dr. Bristol's accuracy if she is not guessing.

the differently made teas starts by assigning a priori probabilities across the range of values of her success rate. If it is reasonable to consider that Dr. Bristol is simply guessing the outcome at random (i.e., her rate of success is .5), then one must assign an a priori probability to this null hypothesis (see our Figure 1, and note the separate amount of probability assigned to $p = .5$). The remaining probability is distributed among the range of other plausible values of Dr. Bristol's success rate (i.e., rates that do not assume that she is guessing at random)³. Then the observations are used to update these probabilities using *Bayes' rule* (this is derived in detail in Etz & Vandekerckhove, this volume). If the observations fit better with the null hypothesis (pure guessing), then the probability assigned to the null hypothesis will increase; if the data fit better with the alternative hypothesis, then the probability assigned to the alternative hypothesis will increase, and subsequently the probability attached to the null hypothesis will decrease (note the decreasing probability of the null hypothesis on the left axis of Figure 2). The factor by which the data shift the balance of the hypotheses' probabilities

³If the null hypothesis is not initially considered tenable, then we can proceed without assigning separate probability to it and instead focus on estimating the parameters of interest (e.g., the taster's accuracy in distinguishing wines, as in Lindley's second example; see Lindley's Figure 1, and notice that the amount of probability assigned to $p = .5$ is gone). Additionally, if a range of values of the parameter is considered impossible—such as rates that are below chance—then this range may be given zero prior probability.

is the *Bayes factor* (Kass & Raftery, 1995; see also Rouder et al., 2009, and Dienes, 2011, below).

A key takeaway from this paper is that Lindley's Bayesian approach depends only on the observed data, so the results are interpretable regardless of whether the sampling plan was rigid or flexible or even known at all. Another key point is that the Bayesian approach is inherently *comparative*: Hypotheses are tested against one another and never in isolation. Lindley further concludes that, since the posterior probability that the null is true will often be higher than the p -value, the latter metric will discount null hypotheses more easily in general.

Bayesian credibility assessments

Source: Kruschke (2015, Chapter 2) — Introduction: Credibility, models, and parameters

“How often have I said to you that when all other θ yield $P(x|\theta)$ of 0, whatever remains, however low its $P(\theta)$, must have $P(\theta|x) = 1$? ”

— Sherlock Holmes, paraphrased

In this book chapter, Kruschke explains the fundamental Bayesian principle of *reallocation of probability*, or “credibility,” across possible states of nature. Kruschke uses an example featuring Sherlock Holmes to demonstrate that the famous detective essentially used Bayesian reasoning to solve his cases. Suppose that Holmes has determined that there exist only four different possible causes (A, B, C, and D) of a committed crime which, for simplicity in the example, he holds to be equally credible at the outset. This translates to equal *prior* probabilities for each of the four possible causes (i.e., a prior probability of 1/4 for each). Now suppose that Holmes gathers evidence that allows him to rule out cause A with certainty. This development causes the probability assigned to A to drop to zero, and the probability that used to be assigned to cause A to be then redistributed across the other possible causes. Since the probabilities for the four alternatives need to sum to one, the probability for each of the other causes is now equal to 1/3 (Figure 2.1, p. 17). What Holmes has done is reallocate credibility across the different possible causes based on the evidence he has gathered. His new state of knowledge is that only one of the three remaining alternatives can be the cause of the crime and that they are all equally plausible. Holmes, being a man of great intellect, is eventually able to completely rule out two of the remaining three causes, leaving him with only one possible explanation—which has to be the cause of the crime (as it now must have probability equal to 1), no matter how improbable it might have seemed at the beginning of his investigation.

The reader might object that it is rather unrealistic to assume that data can be gathered that allow a researcher to completely rule out contending hypotheses. In real applications, psychological data are noisy, and outcomes are only probabilistically linked to the underlying causes. In terms of reallocation of credibility, this means that possible hypotheses can rarely be ruled out completely (i.e., reduced to zero probability), however, their credibility can be greatly diminished, leading to a substantial increase in the credibility of other possible

hypotheses. Although a hypothesis has not been eliminated, something has been learned: Namely, that one or more of the candidate hypotheses has had their probabilities reduced and are now less likely than the others.

In a statistical context, the possible hypotheses are parameter values in mathematical models that serve to describe the observed data in a useful way. For example, a scientist could assume that their observations are normally distributed and be interested in which range of values for the mean is most credible. Sherlock Holmes only considered a set of discrete possibilities, but in many cases it would be very restrictive to only allow a few alternatives (e.g., when estimating the mean of a normal distribution). In the Bayesian framework one can easily consider an infinite continuum of possibilities, across which credibility may still be reallocated. It is easy to extend this framework of reallocation of credibility to hypothesis testing situations where one parameter value is seen as “special” and receives a high amount of prior probability compared to the alternatives (as in Lindley’s tea example above).

Kruschke (2015) serves as a good first introduction to Bayesian thinking, as it requires only basic statistical knowledge (a natural follow-up is Kruschke & Liddell, this volume). In this chapter, Kruschke also provides a concise introduction to mathematical models and parameters, two core concepts which our other sources will build on. One final key takeaway from this chapter is the idea of sequential updating from prior to posterior (Figure 2.1, p. 17) as data are collected. As Dennis Lindley famously said: “Todays posterior is tomorrows prior” (Lindley, 1972, p. 2).

Implications of Bayesian statistics for experimental psychology

Source: Dienes (2011) — Bayesian versus orthodox statistics: Which side are you on?

Dienes explains several differences between the frequentist (which Dienes calls *orthodox* and we have called *classical*; we use these terms interchangeably) and Bayesian paradigm which have practical implications for how experimental psychologists conduct experiments, analyze data, and interpret results (a natural follow-up to the discussion in this section is available in Dienes & McLatchie, this volume). Throughout the paper, Dienes also discusses *subjective* (or context-dependent) Bayesian methods which allow for inclusion of relevant problem-specific knowledge in to the formation of one’s statistical model.

The probabilities of data given theory and of theory given data

When testing a theory, both the frequentist and Bayesian approaches use probability theory as the basis for inference, yet in each framework, the interpretation of probability is different. It is important to be aware of the implications of this difference in order to correctly interpret frequentist and Bayesian analyses. One major contrast is a result of the fact that frequentist statistics only allow for statements to be made about $P(\text{data} \mid \text{theory})^4$: Assuming the theory is correct, the probability of observing the obtained (or more extreme) data is evaluated. Dienes argues that often the probability of the data assuming a theory is correct is not the probability the researcher is interested in. What researchers typically want to know is

⁴The conditional probability (P) of data given (\mid) theory.

$P(\text{theory} \mid \text{data})$: Given that the data were those obtained, what is the probability that the theory is correct? At first glance, these two probabilities might appear similar, but Dienes illustrates their fundamental difference with the following example: The probability that a person is dead (i.e., *data*) given that a shark has bitten the person's head off (i.e., *theory*) is 1. However, given that a person is dead, the probability that a shark has bitten this person's head off is very close to zero (see Senn, 2013, for an intuitive explanation of this distinction). It is important to keep in mind that a *p*-value does *not* correspond to $P(\text{theory} \mid \text{data})$; in fact, statements about this probability are only possible if one is willing to attach prior probabilities (degrees of plausibility or credibility) to theories—which can only be done in the Bayesian paradigm.

In the following sections, Dienes explains how the Bayesian approach is more liberating than the frequentist approach with regard to the following concepts: *stopping rules*, *planned versus post hoc comparisons*, and *multiple testing*. For those new to the Bayesian paradigm, these proposals may seem counterintuitive at first, but Dienes provides clear and accessible explanations for each.

Stopping rules

In the classical statistical paradigm, it is necessary to specify in advance how the data will be collected. In practice, one usually has to specify how many participants will be collected; stopping data collection early or continuing after the pre-specified number of participants has been reached is not permitted. One reason why collecting additional participants is not permitted in the typical frequentist paradigm is that, given the null hypothesis is true, the *p*-value is not driven in a particular direction as more observations are gathered. In fact, in many cases the distribution of the *p*-value is uniform when the null hypothesis is true, meaning that every *p*-value is equally likely under the null. This implies that even if there is no effect, a researcher is guaranteed to obtain a statistically significant result if they simply continue to collect participants and stop when the *p*-value is sufficiently low. In contrast, the Bayes factor, the most common Bayesian method of hypothesis testing, will approach infinite support in favor of the null hypothesis as more observations are collected if the null hypothesis is true. Furthermore, since Bayesian inference obeys the *likelihood principle*, one is allowed to continue or stop collecting participants at any time while maintaining the validity of one's results (p. 276; see also Cornfield, 1966, Rouder, 2014, and Royall, 2004 in the appended *Further Reading* section).

Planned versus post hoc comparisons

In the classical hypothesis-testing approach, a distinction is made between planned and post hoc comparisons: It matters whether the hypothesis was formulated before or after data collection. In contrast, Dienes argues that adherence to the likelihood principle entails that a theory does not necessarily need to precede the data when a Bayesian approach is adopted; since this temporal information does not enter into the likelihood function for the data, the evidence for or against the theory will be the same no matter its temporal relation to the data.

Multiple testing

When conducting multiple tests in the classical approach, it is important to correct for the number of tests performed (see Gelman & Loken, 2014). Dienes points out that within the Bayesian approach, the number of hypotheses tested does not matter—it is not the number of tests that is important, but the evaluation of how accurately each hypothesis predicts the observed data. Nevertheless, it is crucial to consider all relevant evidence, including so-called “outliers,” because “cherry picking is wrong on all statistical approaches” (Dienes, 2011, p. 280).

Context-dependent Bayes factors

The last part of the article addresses how problem-specific knowledge may be incorporated in the calculation of the Bayes factor. As is also explained in our next highlighted source (Rouder et al., 2009), there are two main schools of Bayesian thought: default (or *objective*) Bayes and context-dependent (or *subjective*) Bayes. In contrast to the default Bayes factors for general application that are designed to have certain desirable mathematical properties (e.g., Jeffreys, 1961; Rouder et al., 2009; Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Ly, Verhagen, & Wagenmakers, 2016), Dienes provides an online calculator⁵ that enables one to obtain context-dependent Bayes factors that incorporate domain knowledge for several commonly used statistical tests. In contrast to the default Bayes factors, which are typically designed to use standardized effect sizes, the context-dependent Bayes factors specify prior distributions in terms of the raw effect size. Readers who are especially interested in prior elicitation should see the appendix of Dienes’ article for a short review of how to appropriately specify prior distributions that incorporate relevant theoretical information (and Dienes, 2014, for more details and worked examples).

Structure and motivation of Bayes factors

Source: Rouder et al. (2009) — Bayesian *t*-tests for accepting and rejecting the null hypothesis

In many cases, a scientist’s primary interest is in showing evidence for an *invariance*, rather than a difference. For example, researchers may want to conclude that experimental and control groups do not differ in performance on a task (e.g., Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014), that participants were performing at chance (Dienes & Overgaard, 2015), or that two variables are unrelated (Rouder & Morey, 2012). In classical statistics this is generally not possible as significance tests are asymmetric; they can only serve to reject the null hypothesis and never to affirm it. One benefit of Bayesian analysis is that inference is perfectly symmetric, meaning evidence can be obtained that favors the null hypothesis as well as the alternative hypothesis (see Gallistel, 2009, as listed in our *Further Reading* appendix). This is made possible by the use of *Bayes factors*.⁶ The

⁵<http://www.lifesci.sussex.ac.uk/home/Zoltan.Dienes/inference/Bayes.htm>

⁶Readers for whom Rouder and colleagues’ (2009) treatment is too technical could focus on Dienes’ conceptual ideas and motivations underlying the Bayes factor.

section covering the shortcomings of classical statistics (“Critiques of Inference by Significance Tests”) can safely be skipped, but readers particularly interested in the motivation of Bayesian inference are advised to read it.

What is a Bayes factor?

The Bayes factor is a representation of the relative predictive success of two or more models, and it is a fundamental measure of relative evidence. The way Bayesians quantify predictive success of a model is to calculate the probability of the data given that model—also called the *marginal likelihood* or sometimes simply the *evidence*. The ratio of two such probabilities is the Bayes factor. Rouder and colleagues (2009) denote the probability of the data given some model, represented by H_i , as $f(\text{data} | H_i)$.⁷ The Bayes factor for H_0 versus H_1 is simply the ratio of $f(\text{data} | H_0)$ and $f(\text{data} | H_1)$ written B_{01} (or BF_{01}), where the B (or BF) indicates a Bayes factor, and the subscript indicates which two models are being compared (see p. 228). If the result of a study is $B_{01} = 10$ then the data are ten times more probable under H_0 than under H_1 . Researchers should report the exact value of the Bayes factor since it is a continuous measure of evidence, but various benchmarks have been suggested to help researchers interpret Bayes factors, with values between 1 and 3, between 3 and 10, and greater than 10 generally taken to indicate inconclusive, weak, and strong evidence, respectively (see Jeffreys, 1961; Wagenmakers, 2007; Etz & Vandekerckhove, 2016), although different researchers may set different benchmarks. Care is need when interpreting Bayes factors against these benchmarks, as they are not meant to be bright lines against which we judge a study’s success (as opposed to how a statistical significance criterion is sometimes treated); the difference between a Bayes factor of, say, 8 and 12 is more a difference of degree than of category. Furthermore, Bayes factors near 1 indicate the data are uninformative, and should not be interpreted as even mild evidence for either of the hypotheses under consideration.

Readers who are less comfortable with reading mathematical notation may skip over most of the equations without too much loss of clarity. The takeaway is that to evaluate which model is better supported by the data, we need to find out which model has done the best job predicting the data we observe. To a Bayesian, the probability a model assigns to the observed data constitutes its predictive success (see Morey, Romeijn, & Rouder, 2016); a model that assigns a high probability to the data relative to another model is best supported by the data. The goal is then to find the probability a given model assigns the data, $f(\text{data} | H_i)$. Usually the null hypothesis specifies that the true parameter is a particular value of interest (e.g., 0), so we can easily find $f(\text{data} | H_0)$. However, we generally do not know the value of the parameter if the null model is false, so we do not know what probability it assigns the data. To represent our uncertainty with regard to the true value of the parameter if the null hypothesis is false, Bayesians specify a range of plausible values that the parameter might take under the alternative hypothesis. All of these parameter values are subsequently used in computing an average probability of the data given the alternative

⁷The probability (f) of the observed data given ($|$) hypothesis i (H_i), where i indicates one of the candidate hypotheses (e.g., 0, 1, A, etc.). The null hypothesis is usually denoted H_0 and the alternative hypothesis is usually denoted either H_1 or H_A .

hypothesis, $f(\text{data} \mid H_1)$ (for an intuitive illustration, see Gallistel, 2009 as listed in our *Further Reading* appendix). If the prior distribution gives substantial weight to parameter values that assign high probability to the data, then the average probability the alternative hypothesis assigns to the data will be relatively high—the model is effectively rewarded for its accurate predictions with a high value for $f(\text{data} \mid H_1)$.

The role of priors

The form of the prior can have important consequences on the resulting Bayes factor. As discussed in our third source (Dienes, 2011), there are two primary schools of Bayesian thought: default (objective) Bayes (Berger, 2006) and context-dependent (subjective) Bayes (Goldstein et al., 2006; Rouder, Morey, & Wagenmakers, 2016). The default Bayesian tries to specify prior distributions that convey little information while maintaining certain desirable properties. For example, one desirable property is that changing the scale of measurement should not change the way the information is represented in the prior, which is accomplished by using standardized effect sizes. Context-dependent prior distributions are often used because they more accurately encode our prior information about the effects under study, and can be represented with raw or standardized effect sizes, but they do not necessarily have the same desirable mathematical properties (although sometimes they can).

Choosing a prior distribution for the standardized effect size is relatively straightforward for the default Bayesian. One possibility is to use a normal distribution centered at 0 and with some standard deviation (i.e., spread) σ . If σ is too large, the Bayes factor will always favor the null model, so such a choice would be unwise (see also DeGroot, 1982; Robert, 2014). This happens because such a prior distribution assigns weight to very extreme values of the effect size, when in reality, the effect is most often reasonably small (e.g., almost all psychological effects are smaller than Cohen's $d = 2$). The model is penalized for low predictive success. Setting σ to 1 is reasonable and common—this is called the *unit information prior*. However, using a Cauchy distribution (which resembles a normal distribution but with less central mass and fatter tails) has some better properties than the unit information prior, and is now a common default prior on the alternative hypothesis, giving rise to what is now called the *default Bayes factor* (see Rouder & Morey, 2012 for more details; see also Wagenmakers, Love, et al., this volume and Wagenmakers, Marsman, et al., this volume). To use the Cauchy distribution, like the normal distribution, again one must specify a scaling factor. If it is too large, the same problem as before occurs where the null model will always be favored. Rouder and colleagues suggest a scale of 1, which implies that the effect size has a prior probability of 50% to be between $d = -1$ and $d = 1$. For some areas, such as social psychology, this is not reasonable, and the scale should be reduced. However, slight changes to the scale often do not make much difference in the qualitative conclusions one draws.

Readers are advised to pay close attention to the sections “Subjectivity in priors” and “Bayes factors with small effects.” The former explains how one can tune the scale of the default prior distribution to reflect more contextually relevant information while maintaining the desirable properties attached to prior distributions of this form, a practice that is a reasonable compromise between the default and context-dependent schools. The latter shows why the Bayes factor will often show evidence in favor of the null hypothesis if the observed

effect is small and the prior distribution is relatively diffuse.

Applied sources

At this point, the essential concepts of Bayesian probability, Bayes' theorem, and the Bayes factor have been discussed in depth. In the following four sources, these concepts are applied to real data analysis situations. Our first source provides a broad overview of the most common methods of model comparison, including the Bayes factor, with a heavy emphasis on its proper interpretation (Vandekerckhove, Matzke, & Wagenmakers, 2015). The next source begins by demonstrating Bayesian estimation techniques in the context of developmental research, then provides some guidelines for reporting Bayesian analyses (R. van de Schoot et al., 2014). Our final two sources discuss issues in Bayesian cognitive modeling, such as the selection of appropriate priors (Lee & Vanpaemel, this volume), and the use of cognitive models for theory testing (Lee, 2008).

Before moving on to our final four highlighted sources, it will be useful if readers consider some differences in perspective among practitioners of Bayesian statistics. The application of Bayesian methods is very much an active field of study, and as such, the literature contains a multitude of deep, important, and diverse viewpoints on how data analysis should be done, similar to the philosophical divides between Neyman–Pearson and Fisher concerning proper application of classical statistics (see Lehmann, 1993). The divide between subjective Bayesians, who elect to use priors informed by theory, and objective Bayesians, who instead prefer “uninformative” or default priors, has already been mentioned throughout the *Theoretical sources* section above.

A second division of note exists between Bayesians who see a place for hypothesis testing in science, and those who see statistical inference primarily as a problem of estimation. The former believe statistical models can stand as useful surrogates for theoretical positions, whose relative merits are subsequently compared using Bayes factors and other such “scoring” metrics (as reviewed in Vandekerckhove et al., 2015, discussed below; for additional examples, see Jeffreys, 1961 and Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). The latter would rather delve deeply into a single model or analysis and use point estimates and credible intervals of parameters as the basis for their theoretical conclusions (as demonstrated in Lee, 2008, discussed below; for additional examples, see Gelman & Shalizi, 2013 and McElreath, 2016).⁸

Novice Bayesians may feel surprised that such wide divisions exist, as statistics (of any persuasion) is often thought of as a set of prescriptive, immutable procedures that can be only right or wrong. We contend that debates such as these should be expected due to the wide variety of research questions—and diversity of contexts—to which Bayesian methods are applied. As such, we believe that the existence of these divisions speaks to the intellectual vibrancy of the field and its practitioners. We point out these differences here so that readers

⁸This divide in Bayesian statistics may be seen as a parallel to the recent discussions about use of classical statistics in psychology (e.g., Cumming, 2014), where a greater push has been made to adopt an estimation approach over null hypothesis significance testing (NHST). Discussions on the merits of hypothesis testing have been running through all of statistics for over a century, with no end in sight.

might use this context to guide their continued reading.

Bayesian model comparison methods

Source: Vandekerckhove et al. (2015) — Model comparison and the principle of parsimony

John von Neumann famously said: “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk” (as quoted in Mayer, Khairy, & Howard, 2010, p. 698), pointing to the natural tension between model parsimony and goodness of fit. The tension occurs because it is always possible to decrease the amount of error between a model’s predictions and the observed data by simply adding more parameters to the model. In the extreme case, any data set of N observations can be reproduced perfectly by a model with N parameters. Such practices, however, termed *overfitting*, result in poor generalization and greatly reduce the accuracy of out-of-sample predictions. Vandekerckhove and colleagues (2015) take this issue as a starting point to discuss various criteria for model selection. How do we select a model that both fits the data well and generalizes adequately to new data?

Putting the problem in perspective, the authors discuss research on recognition memory that relies on multinomial processing trees, which are simple, but powerful, cognitive models. Comparing these different models using only the likelihood term is ill-advised, because the model with the highest number of parameters will—all other things being equal—yield the best fit. As a first step to addressing this problem, Vandekerckhove et al. (2015) discuss the popular Akaike information criterion (AIC) and Bayesian information criterion (BIC).

Though derived from different philosophies (for an overview, see Aho, Derryberry, & Peterson, 2014), both AIC and BIC try to solve the trade-off between goodness-of-fit and parsimony by combining the likelihood with a penalty for model complexity. However, this penalty is solely a function of the number of parameters and thus neglects the functional form of the model, which can be informative in its own right. As an example, the authors mention Fechner’s law and Steven’s law. The former is described by a simple logarithmic function, which can only ever fit negatively accelerated data. Steven’s law, however, is described by an exponential function, which can account for both positively and negatively accelerated data. Additionally, both models feature just a single parameter, nullifying the benefit of the complexity penalty in each of the two aforementioned information criteria.

The Bayes factor yields a way out. It extends the simple likelihood ratio test by integrating the likelihood with respect to the prior distribution, thus taking the predictive success of the prior distribution into account (see also Gallistel, 2009, in the *Further Reading* appendix). Essentially, the Bayes factor is a likelihood ratio test averaged over all possible parameter values for the model, using the prior distributions as weights: It is the natural extension of the likelihood ratio test to a Bayesian framework. The net effect of this is to penalize complex models. While a complex model can predict a wider range of possible data points than a simple model can, each individual data point is less likely to be observed under the complex model. This is reflected in the prior distribution being more spread out in the complex model. By weighting the likelihood by the corresponding tiny prior probabilities, the Bayes factor in favor of the complex model decreases. In this way, the Bayes factor

instantiates an automatic Ockham’s Razor (see also Myung & Pitt, 1997, in the appended *Further Reading* section).

However, the Bayes factor can be difficult to compute because it often involves integration over very many dimensions at once. Vandekerckhove and colleagues (2015) advocate two methods to ease the computational burden: importance sampling and the Savage-Dickey density ratio (see also Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010 in our *Further reading* appendix); additional common computational methods include the Laplace approximation (Kass & Raftery, 1995), bridge sampling (Meng & Wong, 1996; Gronau et al., 2017), and the encompassing prior approach (Hoijtink, Klugkist, & Boelen, 2008). They also provide code to estimate parameters in multinomial processing tree models and to compute the Bayes factor to select among them. Overall, the chapter provides a good overview of different methods used to tackle the tension between goodness-of-fit and parsimony in a Bayesian framework. While it is more technical than the sources reviewed above, this article can greatly influence how one thinks about models and methods for selecting among them.

Bayesian estimation

Source: R. van de Schoot et al. (2014) — A gentle introduction to Bayesian analysis: Applications to developmental research

This source approaches practical issues related to parameter estimation in the context of developmental research. This setting offers a good basis for discussing the choice of priors and how those choices influence the posterior estimates for parameters of interest. This is a topic that matters to reviewers and editors alike: How does the choice of prior distributions for focal parameters influence the statistical results and theoretical conclusions that are obtained? The article discusses this issue on a basic and illustrative level.

At this point we feel it is important to note that the difference between hypothesis testing and estimation in the Bayesian framework is much greater than it is in the frequentist framework. In the frequentist framework there is often a one-to-one relationship between the null hypothesis falling outside the sample estimate’s 95% confidence interval and rejection of the null hypothesis with a significance test (e.g., when doing a *t*-test). This is not so in the Bayesian framework; one cannot test a null hypothesis by simply checking if the null value is inside or outside a credible interval. A detailed explanation of the reason for this deserves more space than we can afford to give it here, but in short: When testing hypotheses in the Bayesian framework one should calculate a model comparison metric. See Rouder and Vandekerckhove (this volume) for an intuitive introduction to (and synthesis of) the distinction between Bayesian estimation and testing.

Van de Schoot and colleagues (2014) begin by reviewing the main differences between frequentist and Bayesian approaches. Most of this part can be skipped by readers who are comfortable with basic terminology at that point. The only newly introduced term is *Markov chain Monte Carlo (MCMC)* methods, which refers to the practice of drawing samples from the posterior distribution instead of deriving the distribution analytically (which may not be feasible for many models; see also van Ravenzwaaij, Cassey, & Brown, this volume and Matzke, Boehm, & Vandekerckhove, this volume). After explaining this alternative approach

(p. 848), Bayesian estimation of focal parameters and the specification of prior distributions is discussed with the aid of two case examples.

The first example concerns estimation of an ordinary mean value and the variance of reading scores and serves to illustrate how different sources of information can be used to inform the specification of prior distributions. The authors discuss how expert domain knowledge (e.g., reading scores usually fall within a certain range), statistical considerations (reading scores are normally distributed), and evidence from previous studies (results obtained from samples from similar populations) may be jointly used to define adequate priors for the mean and variance model parameters. The authors perform a prior sensitivity analysis to show how using priors based on different considerations influence the obtained results. Thus, the authors examine and discuss how the posterior distributions of the mean and variance parameters are dependent on the prior distributions used.

The second example focuses on a data set from research on the longitudinal reciprocal associations between personality and relationships. The authors summarize a series of previous studies and discuss how results from these studies may or may not inform prior specifications for the latest obtained data set. Ultimately, strong theoretical considerations are needed to decide whether data sets that were gathered using slightly different age groups can be used to inform inferences about one another.

The authors fit a model with data across two time points and use it to discuss how convergence of the MCMC estimator can be supported and checked. They then evaluate overall model fit via a posterior predictive check. In this type of model check, data simulated from the specified model are compared to the observed data. If the model is making appropriate predictions, the simulated data and the observed data should appear similar. The article concludes with a brief outline of guidelines for reporting Bayesian analyses and results in a manuscript. Here, the authors emphasize the importance of the specification of prior distributions and of convergence checks (if MCMC sampling is used) and briefly outline how both might be reported. Finally, the authors discuss the use of default priors and various options for conducting Bayesian analyses with common software packages (such as Mplus and WinBUGS).

The examples in the article illustrate different considerations that should be taken into account for choosing prior specifications, the consequences they can have on the obtained results, and how to check whether and how the choice of priors influenced the resulting inferences.

Prior elicitation

Source: Lee and Vanpaemel (this volume) — Determining priors for cognitive models

Statistics does not operate in a vacuum, and often prior knowledge is available that can inform one's inferences. In contrast to classical statistics, Bayesian statistics allows one to formalize and use this prior knowledge for analysis. The paper by Lee and Vanpaemel (this volume) fills an important gap in the literature: What possibilities are there to formalize and uncover prior knowledge?

The authors start by noting a fundamental point: Cognitive modeling (as discussed in our

final source, Lee, 2008) is an extension of general purpose statistical modeling (e.g., linear regression). Cognitive models are designed to instantiate theory, and thus may need to use richer information and assumptions than general purpose models (see also Franke, 2016). A consequence of this is that the prior distribution, just like the likelihood, should be seen as an integral part of the model. As Jaynes (2003) put it: “If one fails to specify the prior information, a problem of inference is just as ill-posed as if one had failed to specify the data” (p. 373).

What information can we use to specify a prior distribution? Because the parameters in such a cognitive model usually have a direct psychological interpretation, theory may be used to constrain parameter values. For example, a parameter interpreted as a probability of correctly recalling a word must be between 0 and 1. To make this point clear, the authors discuss three cognitive models and show how the parameters instantiate relevant information about psychological processes. Lee and Vanpaemel also discuss cases in which all of the theoretical content is carried by the prior, while the likelihood does not make any strong assumptions. They also discuss the principle of *transformation invariance*, that is, prior distributions for parameters should be invariant to the scale they are measured on (e.g., measuring reaction time using seconds versus milliseconds).

Lee and Vanpaemel also discuss specific methods of prior specification. These include the maximum entropy principle, the prior predictive distribution, and hierarchical modeling. The prior predictive distribution is the model-implied distribution of the data, weighted with respect to the prior. Recently, iterated learning methods have been employed to uncover an implicit prior held by a group of participants. These methods can also be used to elicit information that is subsequently formalized as a prior distribution. (For a more in-depth discussion of hierarchical cognitive modeling, see Lee, 2008, discussed below.)

In sum, the paper gives an excellent overview of why and how one can specify prior distributions for cognitive models. Importantly, priors allow us to integrate domain-specific knowledge, and thus build stronger theories (Platt, 1964; Vanpaemel, 2010). For more information on specifying prior distributions for data-analytic statistical models rather than cognitive models see Rouder, Morey, Verhagen, Swagman, and Wagenmakers (In Press) and Rouder, Engelhardt, McCabe, and Morey (2016).

Bayesian cognitive modeling

Source: Lee (2008) — Three case studies in the Bayesian analysis of cognitive models

Our final source (Lee, 2008) further discusses cognitive modeling, a more tailored approach within Bayesian methods. Often in psychology, a researcher will not only expect to observe a particular effect, but will also propose a verbal theory of the cognitive process underlying the expected effect. Cognitive models are used to formalize and test such verbal theories in a precise, quantitative way. For instance, in a cognitive model, psychological constructs, such as attention and bias, are expressed as model parameters. The proposed psychological process is expressed as dependencies among parameters and observed data (the “structure” of the model).

In peer-reviewed work, Bayesian cognitive models are often presented in visual form as a

graphical model. Model parameters are designated by nodes, where the shape, shading, and style of border of each node reflect various parameter characteristics. Dependencies among parameters are depicted as arrows connecting the nodes. Lee gives an exceptionally clear and concise description of how to read graphical models in his discussion of multidimensional scaling (Lee, 2008, p. 2).

After a model is constructed, the observed data are used to update the priors and generate a set of posterior distributions. Because cognitive models are typically complex, posterior distributions are almost always obtained through sampling methods (i.e., MCMC; see van Ravenzwaaij et al., this volume), rather than through direct, often intractable, analytic calculations.

Lee demonstrates the construction and use of cognitive models through three case studies. Specifically, he shows how three popular process models may be implemented in a Bayesian framework. In each case, he begins by explaining the theoretical basis of each model, then demonstrates how the verbal theory may be translated into a full set of prior distributions and likelihoods. Finally, Lee discusses how results from each model may be interpreted and used for inference.

Each case example showcases a unique advantage of implementing cognitive models in a Bayesian framework (see also Bartlema, Voorspoels, Rutten, Tuerlinckx, & Vanpaemel, this volume). For example, in his discussion of signal detection theory, Lee highlights how Bayesian methods are able to account for individual differences easily (see also Rouder & Lu, 2005, in the *Further reading* appendix). Throughout, Lee emphasizes that Bayesian cognitive models are useful because they allow the researcher to reach new theoretical conclusions that would be difficult to obtain with non-Bayesian methods. Overall, this source not only provides an approachable introduction to Bayesian cognitive models, but also provides an excellent example of good reporting practices for research that employs Bayesian cognitive models.

Conclusion

By focusing on interpretation, rather than implementation, we have sought to provide a more accessible introduction to the core concepts and principles of Bayesian analysis than may be found in introductions with a more applied focus. Ideally, readers who have read through all eight of our highlighted sources, and perhaps some of the supplementary reading, may now feel comfortable with the fundamental ideas in Bayesian data analysis, from basic principles (Kruschke, 2015; Lindley, 1993) to prior distribution selection (Lee & Vanpaemel, this volume), and with the interpretation of a variety of analyses, including Bayesian analogs of classical statistical tests (e.g., *t*-tests; Rouder et al., 2009), estimation in a Bayesian framework (R. van de Schoot et al., 2014), Bayes factors and other methods for hypothesis testing (Dienes, 2011; Vandekerckhove et al., 2015), and Bayesian cognitive models (Lee, 2008).

Reviewers and editors unfamiliar with Bayesian methods may initially feel hesitant to evaluate empirical articles in which such methods are applied (Wagenmakers, Love, et al., this volume). Ideally, the present article should help ameliorate this apprehension by offering

an accessible introduction to Bayesian methods that is focused on interpretation rather than application. Thus, we hope to help minimize the amount of reviewer reticence caused by authors' choice of statistical framework.

Our overview was not aimed at comparing the advantages and disadvantages of Bayesian and classical methods. However, some conceptual conveniences and analytic strategies that are only possible or valid in the Bayesian framework will have become evident. For example, Bayesian methods allow for the easy implementation of hierarchical models for complex data structures (Lee, 2008), they allow multiple comparisons and flexible sampling rules during data collection without correction of inferential statistics (Dienes, 2011; see also Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015, as listed in our *Further reading* appendix, and also Schönbrodt & Wagenmakers, this volume), and they allow inferences that many researchers in psychology are interested in but are not able to answer with classical statistics such as providing support for a null hypothesis (for a discussion, see Wagenmakers, 2007). Thus, the inclusion of more research that uses Bayesian methods in the psychological literature should be to the benefit of the entire field (Etz & Vandekerckhove, 2016). In this article, we have provided an overview of sources that should allow a novice to understand how Bayesian statistics allow for these benefits, even without prior knowledge of Bayesian methods.

Acknowledgments

The authors would like to thank Jeff Rouder, E.-J. Wagenmakers, and Joachim Vandekerckhove for their helpful comments. AE and BB were supported by grant #1534472 from NSF's Methods, Measurements, and Statistics panel. AE was further supported by the National Science Foundation Graduate Research Fellowship Program (#DGE1321846).

References

- Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the world-views of aic and bic. *Ecology*, 95(3), 631–636.
- Bartlema, A., Voorspoels, W., Rutten, F., Tuerlinckx, F., & Vanpaemel, W. (this volume). Sensitivity to the prototype in children with high-functioning autism spectrum disorder: An example of Bayesian cognitive psychometrics. *Psychonomic Bulletin and Review*.
- Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian analysis*, 1(3), 385–402.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76(2), 159–165.
- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317–335.
- Cornfield, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, 20, 18–23.
- Cumming, G. (2014). The new statistics why and how. *Psychological Science*, 25(1), 7–29.

- DeGroot, M. H. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association*, 336–339.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3), 274–290.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5.
- Dienes, Z., & McLatchie, N. (this volume). Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review*.
- Dienes, Z., & Overgaard, M. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. *Behavioural methods in consciousness research*, 199–220.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychology research. *Psychological Review*, 70(3), 193–242.
- Etz, A., & Vandekerckhove, J. (2016). *PLOS ONE*, 11, e0149794.
- Etz, A., & Vandekerckhove, J. (this volume). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin and Review*.
- Etz, A., & Wagenmakers, E.-J. (in press). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*.
- Franke, M. (2016). Task types, link functions & probabilistic modeling in experimental pragmatics. In F. Salfner & U. Sauerland (Eds.), *Preproceedings of 'trends in experimental pragmatics'* (pp. 56–63).
- Gallistel, C. (2009). The importance of proving the null. *Psychological review*, 116(2), 439.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Vol. 3). Chapman & Hall/CRC.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606.
- Goldstein, M., et al. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3), 403–420.
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *arXiv preprint arXiv:1703.05984*.
- Hoijsink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. Springer Science & Business Media.
- Jaynes, E. T. (1986). Bayesian methods: General background. In J. H. Justice (Ed.), *Maximum entropy and bayesian methods in applied statistics* (pp. 1–25). Cambridge University Press.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.
- Jeffreys, H. (1936). Xxviii. on some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 22(146), 337–359.

- Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
- Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). Guilford New York, NY.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- Kruschke, J. K., & Liddell, T. (this volume). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, 15(1), 1–15.
- Lee, M. D., & Vanpaemel, W. (this volume). Determining priors for cognitive models. *Psychonomic Bulletin & Review*.
- Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lehmann, E. (1993). The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424), 1242–1249.
- Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia (PA): SIAM.
- Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15(1), 22–25.
- Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49(3), 293–337.
- Lindley, D. V. (2006). *Understanding uncertainty*. John Wiley & Sons.
- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., et al. (2015). JASP (version 0.7.1.12). *Computer Software*.
- Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016). Harold Jeffreys default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Matzke, D., Boehm, U., & Vandekerckhove, J. (this volume). Bayesian inference for psychology, Part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin and Review*.
- Mayer, J., Khairy, K., & Howard, J. (2010). Drawing an elephant with four complex parameters. *American Journal of Physics*, 78(6), 648–649.
- McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Vol. 122). CRC Press.
- Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.
- Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occams razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1), 79–95.
- Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2), 241.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

- Orwell, G. (1946). A nice cup of tea. *Evening Standard, January*.
- Platt, J. R. (1964). Strong inference. *Science, 146*(3642), 347–353.
- Ravenzwaaij, D. van, Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General, 143*(5), 1794–1805.
- Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science, 81*(2), 216–232.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 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., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47*(6), 877–903.
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374.
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*, 520–547.
- Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (In Press). Bayesian analysis of factorial designs. *Psychological Methods*.
- Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra, 2*(1).
- 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 and Review, 16*(2), 225–237.
- Rouder, J. N., & Vandekerckhove, J. (this volume). Bayesian inference for psychology, Part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin and Review*.
- Royall, R. (1997). *Statistical evidence: A likelihood paradigm* (Vol. 77). CRC press.
- Royall, R. (2004). The likelihood paradigm for statistical inference. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical and empirical considerations* (pp. 119–152). The University of Chicago Press.
- Schönbrodt, F. D., & Wagenmakers, E.-J. (this volume). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*.
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.
- Schoot, R. van de, Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860.
- Schoot, R. Van de, Winder, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (in

- press). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*.
- Senn, S. (2013). Invalid inversion. *Significance*, 10(2), 40–42.
- 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*(3).
- Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis*. Sebtel Press.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2.
- van Ravenzwaaij, D., Cassey, P., & Brown, S. (this volume). A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin and Review*.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eides (Eds.), *Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–317). Oxford University Press.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.
- Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4), 14–57.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, 14(5), 779–804.
- 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(3), 158–189.
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (this volume). Bayesian inference for psychology, Part II: Example applications with JASP. *Psychonomic Bulletin and Review*.
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (this volume). Bayesian inference for psychology, Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*.
- Wagenmakers, E.-J., Morey, R. D., & Lee, M. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3).
- Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior research methods*, 1–14.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 *t*-tests. *Perspectives on Psychological Science*, 6(3), 291–298.
- Winkler, R. L. (2003). *An introduction to Bayesian inference and decision* (2 ed.). Holt, Rinehart and Winston New York.

Appendix

Further reading

In this Appendix, we provide a concise overview of 32 additional articles and books that provide further discussion of various theoretical and applied topics in Bayesian inference. For example, the list includes articles that editors and reviewers might consult as a reference while reviewing manuscripts that apply advanced Bayesian methods such as structural equation models (Kaplan & Depaoli, 2012), hierarchical models (Rouder & Lu, 2005), linear mixed models (Sorensen, Hohenstein, & Vasishth, 2016), and design (i.e., power) analyses (Schönbrodt et al., 2015). The list also includes books that may serve as accessible introductory texts (e.g., Dienes, 2008) or as more advanced textbooks (e.g., Gelman et al., 2013). To aid in readers' selection of sources, we have summarized the associated focus and difficulty ratings for each source in Figure 7.2.

Recommended articles

9. **Cornfield (1966)** — Sequential Trials, Sequential Analysis, and the Likelihood Principle. *Theoretical focus (3), moderate difficulty (5).*

A short exposition of the difference between Bayesian and classical inference in sequential sampling problems.

10. **Lindley (2000)** — The Philosophy of Statistics. *Theoretical focus (1), moderate difficulty (5).*

Dennis Lindley, a foundational Bayesian, outlines his philosophy of statistics, receives commentary, and responds. An illuminating paper with equally illuminating commentaries.

11. **Jaynes (1986)** — Bayesian Methods: General Background. *Theoretical focus (2), low difficulty (2).*

A brief history of Bayesian inference. The reader can stop after finishing the section titled, "Is our logic open or closed," because the further sections are somewhat dated and not very relevant to psychologists.

12. **Edwards, Lindman, and Savage (1963)** — Bayesian Statistical Inference for Psychological Research. *Theoretical focus (2), high difficulty (9).*

The article that first introduced Bayesian inference to psychologists. A challenging but insightful and rewarding paper. Much of the more technical mathematical notation can be skipped with minimal loss of understanding.

13. **Rouder, Morey, and Wagenmakers (2016)** — The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Theoretical focus (2), low difficulty (3)*

All forms of statistical analysis, both Bayesian and frequentist, require some subjective input (see also Berger & Berry, 1988). In this article, the authors emphasize that subjectivity is in fact desirable, and one of the benefits of the Bayesian approach is that the inclusion of subjective elements is transparent and therefore open to discussion.

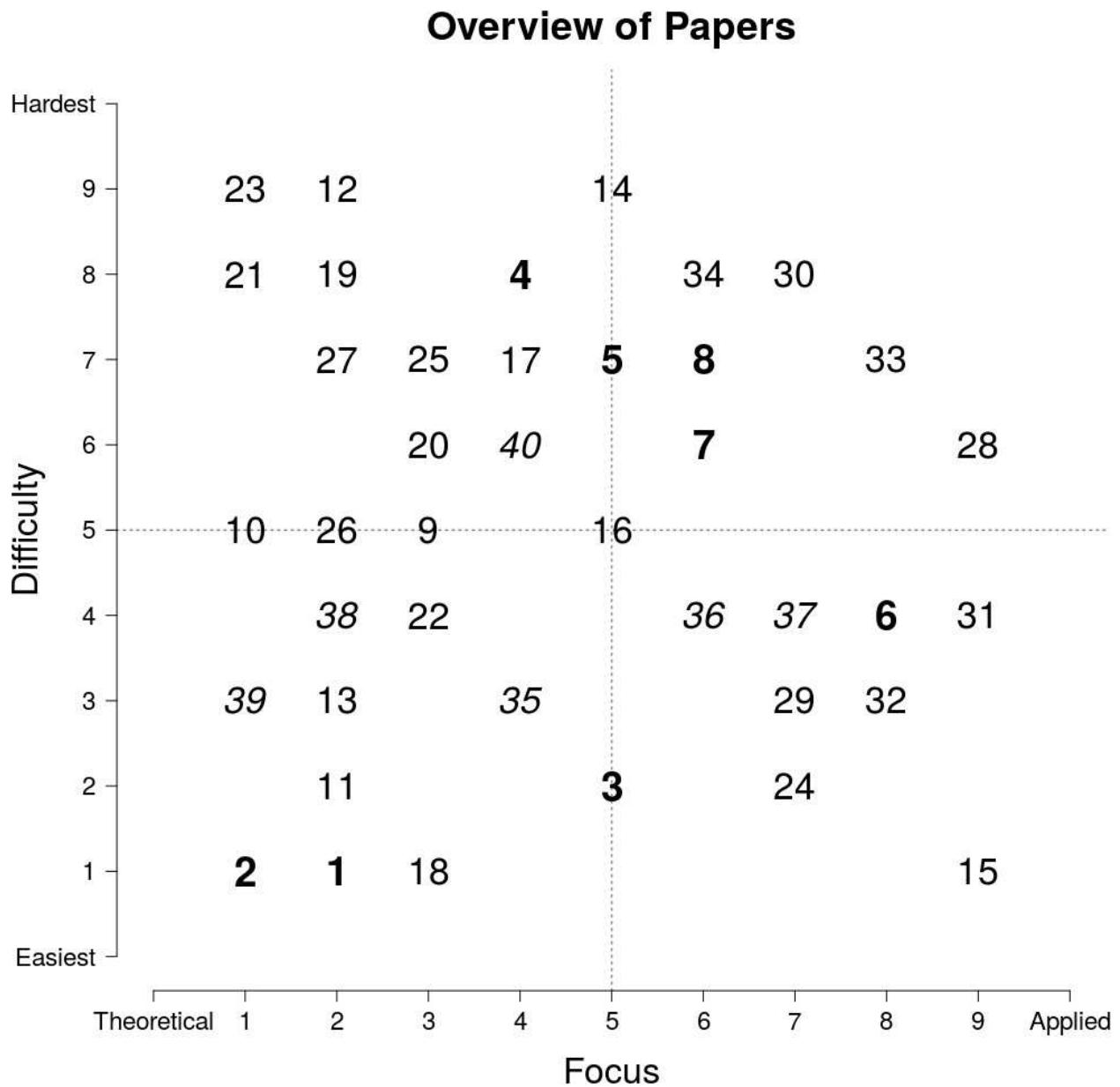


Figure 7.2: An overview of focus and difficulty ratings for all sources included in the present paper. Sources discussed at length in the *Theoretical sources* and *Applied sources* sections are presented in bold text. Sources listed in the appended *Further reading* appendix are presented in light text. Source numbers representing books are italicized.

14. **Myung and Pitt (1997)** — Applying Occam’s Razor in Cognitive Modeling: A Bayesian Approach. *Balanced focus (5), high difficulty (9).*

This paper brought Bayesian methods to greater prominence in modern psychology, discussing the allure of Bayesian model comparison for non-nested models and providing worked examples. As the authors provide a great discussion of the principle of parsimony, thus this paper serves as a good follow-up to our fifth highlighted source

(Vandekerckhove et al., 2015).

15. **Wagenmakers, Morey, and Lee (2016)** — Bayesian Benefits for the Pragmatic Researcher. *Applied focus (9), low difficulty (1).*

Provides pragmatic arguments for the use of Bayesian inference with two examples featuring fictional characters Eric Cartman and Adam Sandler. This paper is clear, witty, and persuasive.

16. **Rouder (2014)** — Optional Stopping: No Problem for Bayesians. *Balanced focus (5), moderate difficulty (5).*

Provides a simple illustration of why Bayesian inference is valid in the case of optional stopping. A natural follow-up to our third highlighted source (Dienes, 2011).

17. **Verhagen and Wagenmakers (2014)** — Bayesian Tests to Quantify the Result of a Replication Attempt. *Balanced focus (4), high difficulty (7).*

Outlines so-called “replication Bayes factors,” which use the original study’s estimated posterior distribution as a prior distribution for the replication study’s Bayes factor. Given the current discussion of how to estimate replicability (Open Science Collaboration, 2015), this work is more relevant than ever. (See also Wagenmakers, Verhagen, and Ly (2015) for a natural follow-up.)

18. **Gigerenzer (2004)** — Mindless Statistics. *Theoretical focus (3), low difficulty (1).*

This paper constructs an enlightening and witty overview on the history and psychology of statistical thinking. It contextualizes the need for Bayesian inference.

19. **Ly et al. (2016)** — Harold Jeffreys’s Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology. *Theoretical focus (2), high difficulty (8).*

A concise summary of the life, work, and thinking of Harold Jeffreys, inventor of the Bayes factor (see also Etz & Wagenmakers, in press). The second part of the paper explains the computations in detail for *t*-tests and correlations. The first part is essential in grasping the motivation behind the Bayes factor.

20. **Robert (2014)** — On the Jeffreys–Lindley Paradox. *Theoretical focus (3), moderate difficulty (6).*

Robert discusses the implications of the Jeffreys–Lindley paradox, so-called because Bayesians and frequentist hypothesis tests can come to diametric conclusions from the same data—even with infinitely large samples. The paper further outlines the need for caution when using *improper priors*, and why they present difficulties for Bayesian hypothesis tests. (For more on this topic see DeGroot, 1982).

21. **Jeffreys (1936)** — On Some Criticisms of the Theory of Probability. *Theoretical focus (1), high difficulty (8).*

An early defense of probability theory's role in scientific inference by one of the founders of Bayesian inference as we know it today. The paper's notation is somewhat outdated and makes for rather slow reading, but Jeffreys's writing is insightful nonetheless.

22. **Rouder, Morey, Verhagen, et al. (2016)** — Is There a Free Lunch in Inference? *Theoretical focus (3), moderate difficulty (4).*

A treatise on why making detailed assumptions about alternatives to the null hypothesis is requisite for a satisfactory method of statistical inference. A good reference for why Bayesians cannot do hypothesis testing by simply checking if a null value lies inside or outside of a credible interval, and instead must calculate a Bayes factor to evaluate the plausibility of a null model.

23. **Berger and Delampady (1987)** — Testing Precise Hypotheses. *Theoretical focus (1), high difficulty (9).*

Explores the different conclusions to be drawn from hypothesis tests in the classical versus Bayesian frameworks. This is a resource for readers with more advanced statistical training.

24. **Wetzels et al. (2011)** — Statistical Evidence in Experimental Psychology: An Empirical Comparison using 855 *t*-tests. *Applied focus (7), low difficulty (2).*

Using 855 *t*-tests from the literature, the authors quantify how inference based on *p* values, effect sizes, and Bayes factors differ. An illuminating reference to understand the practical differences between various methods of inference.

25. **Vanpaemel (2010)** — Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor. *Theoretical focus (3), high difficulty (7).*

The authors defend Bayes factors against the common criticism that the inference is sensitive to specification of the prior. They assert that this sensitivity is valuable and desirable.

26. **Royall (2004)** — The Likelihood Paradigm for Statistical Inference. *Theoretical focus (2), moderate difficulty (5).*

An accessible introduction to the Likelihood principle, and its relevance to inference. Contrasts are made among different accounts of statistical evidence. A more complete account is given in Royall (1997).

27. **Gelman and Shalizi (2013)** — Philosophy and the Practice of Bayesian Statistics. *Theoretical focus (2), high difficulty (7).*

This is the centerpiece of an excellent special issue on the philosophy of Bayesian inference. We recommend that discussion groups consider reading the entire special issue (*British Journal of Mathematical and Statistical Psychology*, February, 2013), as it promises intriguing and fundamental discussions about the nature of inference.

28. **Wagenmakers et al. (2010)** — Bayesian Hypothesis Testing for Psychologists: A Tutorial on the Savage-Dickey Ratio. *Applied focus (9), moderate difficulty (6).*

Bayes factors are notoriously hard to calculate for many types of models. This article introduces a useful computational trick known as the “Savage-Dickey Density Ratio,” an alternative conception of the Bayes factor that makes many computations more convenient. The Savage-Dickey ratio is a powerful visualization of the Bayes factor, and is the primary graphical output of the Bayesian statistics software JASP (Love et al., 2015).

29. **Gallistel (2009)** — The Importance of Proving the Null. *Applied focus (7), low difficulty (3).*

The importance of null hypotheses is explored through three thoroughly worked examples. This paper provides valuable guidance for how one should approach a situation in which it is theoretically desirable to accumulate evidence for a null hypothesis.

30. **Rouder and Lu (2005)** — An Introduction to Bayesian Hierarchical Models with an Application in the Theory of Signal Detection. *Applied focus (7), high difficulty (8).*

This is a good introduction to hierarchical Bayesian inference for the more mathematically inclined readers. It demonstrates the flexibility of hierarchical Bayesian inference applied to signal detection theory, while also introducing augmented Gibbs sampling.

31. **Sorensen et al. (2016)** — Bayesian Linear Mixed Models Using Stan: A Tutorial for Psychologists. *Applied focus (9), moderate difficulty (4).*

Using the software Stan, the authors give an accessible and clear introduction to hierarchical linear modeling. Because both the paper and code are hosted on github, this article serves as a good example of open, reproducible research in a Bayesian framework.

32. **Schönbrodt et al. (2015)** — Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Applied focus (8), low difficulty (3).*

For Bayesians, power analysis is often an afterthought because sequential sampling is encouraged, flexible, and convenient. This paper provides Bayes factor simulations that give researchers an idea of how many participants they might need to collect to achieve moderate levels of evidence from their studies.

33. **Kaplan and Depaoli (2012)** — Bayesian Structural Equation Modeling. *Applied focus (8), high difficulty (7).*

One of few available practical sources on Bayesian structural equation modeling. The article focuses on the Mplus software but also stands as a general source.

34. **Rouder et al. (In Press)** — Bayesian Analysis of Factorial Designs. *Balanced focus (6), high difficulty (8).*

Includes examples of how to set up Bayesian ANOVA models, which are some of the more challenging Bayesian analyses to perform and report, as intuitive hierarchical

models. In the appendix, how to use the BayesFactor R package and JASP software for ANOVA is demonstrated. The relatively high difficulty rating is due to the large amount of statistical notation.

Recommended books

35. **Winkler (2003)** — Introduction to Bayesian Inference and Decision. *Balanced focus (4), low difficulty (3).*

As the title suggests, this is an accessible textbook that introduces the basic concepts and theory underlying the Bayesian framework for both inference and decision-making. The required math background is elementary algebra (i.e., no calculus is required).

36. **McElreath (2016)** — Statistical Rethinking: A Bayesian Course with Examples in R and Stan. *Balanced focus (6), moderate difficulty (4).*

Not your traditional applied introductory statistics textbook. McElreath focuses on education through simulation, with handy R code embedded throughout the text to give readers a hands-on experience.

37. **Lee and Wagenmakers (2014)** — Bayesian Cognitive Modeling: A Practical Course. *Applied focus (7), moderate difficulty (4).*

A textbook on Bayesian cognitive modeling methods that is in a similar vein to our eighth highlighted source (Lee, 2008). It includes friendly introductions to core principles of implementation and many case examples with accompanying MATLAB and R code.

38. **Lindley (2006)** — Understanding Uncertainty. *Theoretical focus (2), moderate difficulty (4).*

An introduction to thinking about uncertainty and how it influences everyday life and science. Lindley proposes that all types of uncertainty can be represented by probabilities. A largely non-technical text, but a clear and concise introduction to the general Bayesian perspective on decision making under uncertainty.

39. **Dienes (2008)** — Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. *Theoretical focus (1), low difficulty (3).*

A book that covers a mix of philosophy of science, psychology, and Bayesian inference. It is a very accessible introduction to Bayesian statistics, and it very clearly contrasts the different goals of Bayesian and classical inference.

40. **Stone (2013)** — Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. *Balanced focus (4), moderate difficulty (6).*

In this short and clear introductory text, Stone explains Bayesian inference using accessible examples and writes for readers with little mathematical background. Accompanying Python and MATLAB code is provided on the author's website.

III

Advanced topics

Determining informative priors for cognitive models

Michael D. Lee and Wolf Vanpaemel

Introduction

One way to think of cognitive modeling is as a natural extension of data analysis. Both involve developing, testing, and using formal models as accounts of brain and behavioral data. The key difference is the interpretation of the model likelihood and parameters. Data analysis typically relies on a standard set of statistical models, especially Generalized Linear Models (GLMs) that form the foundations of regression and the analysis of variance. In these models, parameters have generic interpretations, like locations and scales. Cognitive models, in contrast, aim to afford more substantive interpretations. It is natural to interpret the parameters in cognitive models as psychological variables like memory capacities, attention weights, or learning rates.

For both data-analytic and cognitive models, the likelihood is the function that gives the probability of observed data for a given set of parameter values. For data-analytic models, these likelihoods typically follow from GLMs. Cognitive models often use likelihoods designed to formalize assumptions about psychological processes, such as the encoding of a stimulus in memory, or the termination of search in decision making. Even when a cognitive model uses a likelihood function consistent with GLMs—for example, modeling choice probabilities as weighted linear combinations of stimulus attributes—it is natural to interpret the likelihood as corresponding to cognitive processes, because of the psychological interpretability of the parameters.

Their more elaborate interpretation means that cognitive models aim to formalize and use richer information and assumptions than data-analytic models do. In the standard frequentist approach, assumptions can only be used to specify the likelihood, and, less commonly, the bounds of the parameter space. The Bayesian approach offers the additional possibility of expressing assumptions in the prior distribution over the parameters. These prior distributions are representations of the relative probability that a parameter—or more generally, sets of parameters—have specific values, and thus formalize what is known and unknown

about psychological variables.

Conceived in this way, priors are clearly an advantage of the Bayesian approach. They provide a way of formalizing available information and making theoretical assumptions, enabling the evaluation of the assumptions by empirical evidence, and applying what is learned to make more complete model-based inferences and predictions. Priors are often, however, maligned by those resistant to Bayesian methods (e.g., Edwards, 1991; Trafimow, 2005). Even those who advocate Bayesian methods in cognitive modeling sometimes regard the need to specify a prior as a cost that must be borne to reap the benefits of complete and coherent inference. This lack of interest in the prior often results in what Gill (2014) terms “Bayesians of convenience”, who use priors they label vague, flat, non-committal, weakly informative, default, diffuse, or something else found nearby in a thesaurus.

We believe failing to give sufficient attention to specifying priors is unfortunate, and potentially limits what cognitive modeling can achieve. Our view is that priors should be informative, which means that they should capture the relevant theoretical, logical, and empirical information about the psychological variables they represent (Dienes, 2014; Vanpaemel & Lee, 2012). Only when modelers genuinely have no information about their parameters should informative priors be vague. In the usual and desirable situation in which something is known about parameters, assuming a vague prior loses useful information. The problem is put most emphatically by Jeff Gill (personal communication, August 2015):

“Prior information is all over the place in the social sciences. I really don’t want to read a paper by authors who didn’t know *anything* about their topic before they started.”

Modelers do not strive to make likelihoods vague, but aim to make them consistent with theory, empirical regularities, and other relevant information. Since, in the Bayesian approach, priors and likelihoods combine to form the predictive distribution over data that is the model, priors should also aim to be informative. It seems ironic to make the effort of developing a likelihood that is as informative as possible, only to dilute the predictions of the model by choosing a prior of convenience that ignores relevant theory, data, and logic. A worked example from psychophysics, showing how the unthinking assumption of vague priors can undo the theoretical content of a likelihood, is provided by (Lee, in press, see especially Figures 9 and 11).

There are probably two reasons for the routine use of vague priors, and the lack of effort in specifying informative priors. One involves discomfort with the fact that the choice of different informative priors will affect inference. These sorts of concerns about subjectivity are easy to dismiss. One reaction is to point out that it would be non-sensical if modeling assumptions like priors did not affect inference. A more constructive way to address the concern is to point out that developing likelihoods is just as challenging as developing priors, and inference is also sensitive to choices about likelihoods. Proposing models is a creative scientific act that, in a Bayesian approach, extends to include both priors and likelihoods. The sort of attitudes and practices modelers have in developing, justifying, and testing likelihoods should naturally carry over to priors. Leamer (1983, p.37) insightfully highlights that both the likelihood and the prior are assumptions, and that a perceived difference in their subjectivity simply reflects the frequency of their use:

"The difference between a fact and an opinion for purposes of decision making and inference is that when I use opinions, I get uncomfortable. I am not too uncomfortable with the opinion that error terms are normally distributed because most econometricians make use of that assumption. This observation has deluded me into thinking that the opinion that error terms are normal may be a fact, when I know deep inside that normal distributions are actually used only for convenience. In contrast, I am *quite* uncomfortable using a prior distribution, mostly I suspect because hardly anyone uses them. If convenient prior distributions were used as often as convenient sampling distributions, I suspect that I could be as easily deluded into thinking that prior distributions are facts as I have been into thinking that sampling distributions are facts."

The second probable reason for the reliance on vague priors involves a lack of established methods for determining informative priors. Against this concern, the goal of this paper is to discuss how informative priors can be developed for cognitive models so that they are reasonable, useful, and capture as much information as possible. We identify several sources of information that can help to specify priors for cognitive models, and then discuss some of the methods by which this information can be incorporated into formal priors within a model. Finally, we identify a number of benefits arising from including informative priors in cognitive models. We mostly rely on published examples of the use of priors in cognitive modeling, but also point to under-used sources and methods that we believe provide important future directions for the field.

Three illustrative cognitive models

To help make some general and abstract ideas clear, we draw repeatedly upon three illustrative cognitive models, involving memory, categorization, and decision making. In this section, we describe these models in some detail.

Exponential decay model of memory retention

A simple and standard model of memory retention assumes that the probability of recalling an item decays exponentially with time (Rubin & Wenzel, 1996). One way to formalize this model is to assume that the probability of recalling the i th item at time t_i if it was last studied at time τ_i , is $p_i = \phi \exp\{-\psi(t_i - \tau_i)\}$. Figure 8.1 illustrates this model, showing the study times for three items, and the retention curves assumed by the model.

The ϕ parameter has the psychological interpretation of the initial probability of recall, that is, $\phi = p_i$ when $t_i = \tau_i$, while the ψ parameter controls the rate at which recall probabilities change over time. The parameter space is restricted to $\psi > 0$, so that the model formalizes the assumption of decay (e.g., Wickens, 1998). The usual assumption is that the τ_i time intervals are known from the experimental design, based on explicit study presentations, or that all $\tau_i = 0$ corresponding to the end of the study period. We consider a richer model in which the τ_i rehearsal times are treated as parameters, representing the last

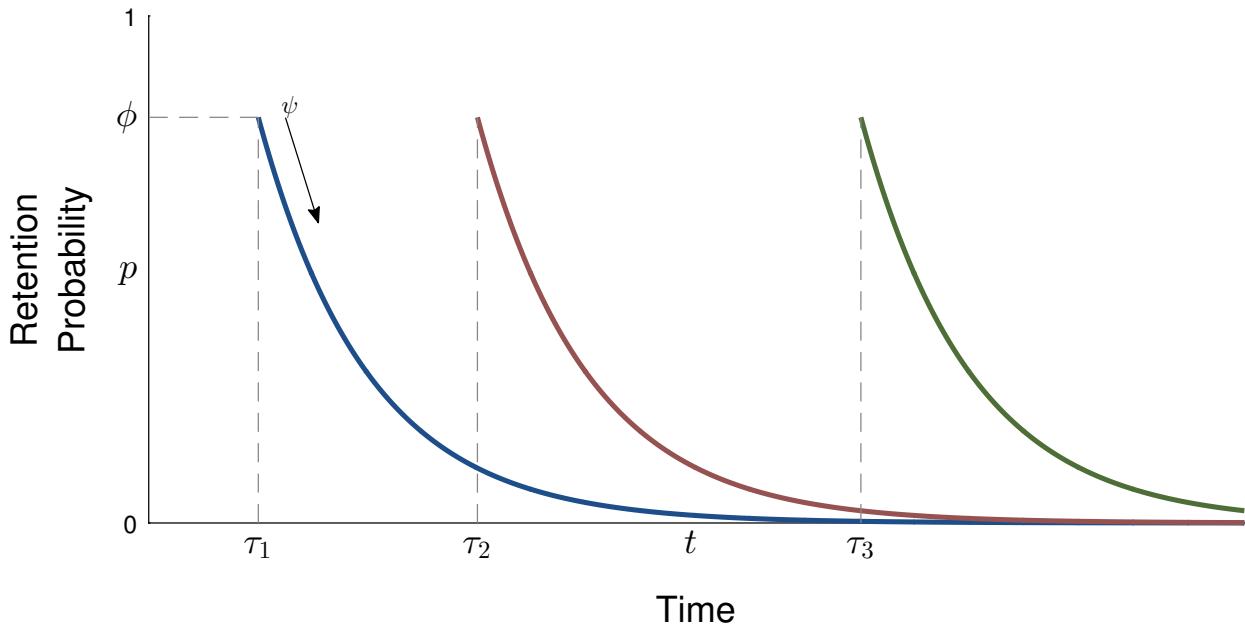


Figure 8.1: An exponential decay model of memory retention. The x -axis corresponds to time t , and the y -axis corresponds to the probability p that an item will be recalled at a specified time. Retention curves for three items are shown. Each curve starts at the time the item was last rehearsed, corresponding to the parameters τ_1 , τ_2 , and τ_3 . The initial probability of recall at this time of last rehearsal is given by the parameter ϕ . The rate of decrease in the probability of recall as time progresses depends on a decay parameter ψ .

unobserved mental rehearsal of the item. This extension is made possible by the flexibility of Bayesian methods, and raises interesting questions about determining appropriate priors for the τ_i latent rehearsal parameters.

Generalized Context Model of categorization

The Generalized Context Model (GCM: Nosofsky, 1986) is a seminal model of categorization. It assumes that categorization behavior is based on comparing the attention-weighted similarity of a presented stimulus to known exemplars of the possible alternative categories. A visual representation of the core assumptions of the model is provided in Figure 8.2. This figure shows, in an attention-weighted psychological space, the generalization gradients of similarity for a new stimulus “?” into two categories represented by circle and square exemplars.

Formally, in the version of the GCM that we consider, the i th stimulus is represented by the coordinate location \mathbf{x}_i , so that the attention-weighted distance between the i th and j th stimuli is $d_{ij} = \sum_k \omega_k |x_{ik} - x_{jk}|$, where ω_k is the attention given to the k th dimension. Accordingly, a dimension receiving more attention will be more influential in determining distances than the ones receiving less attention. The similarity between these stimuli is then $s_{ij} = \exp(-\lambda d_{ij})$, with λ controlling the generalization gradient between stimuli. The

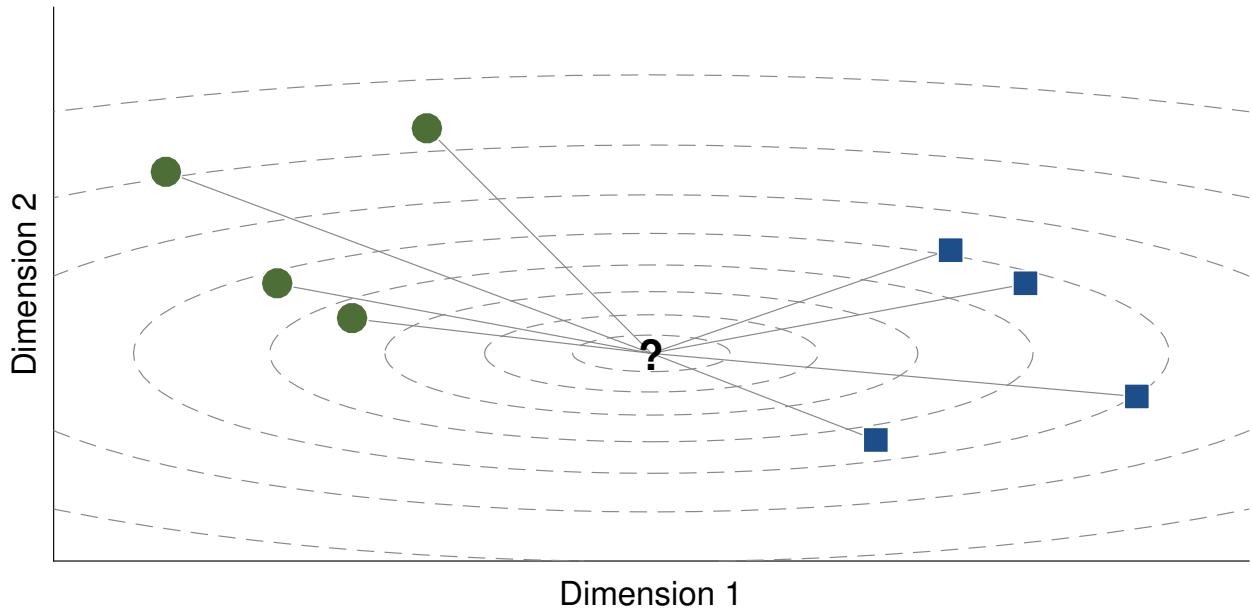


Figure 8.2: The Generalized Context Model of categorization. Eight stimuli are shown in an attention-weighted two-dimensional representation. Four stimuli in one category are represented by circles, and four stimuli in an alternative category are represented by squares. More attention is given to the first stimulus dimension than to the second stimulus dimension, which “stretches” the space to emphasize differences between the stimuli on the first dimension. Generalization gradients from the stimulus to be categorized, marked by “?”, to the known stimuli are shown by ellipses. These gradients produce measures of similarity between the stimuli, based on their distance in the space, and the steepness of the generalization gradient. The total similarity between the stimulus to be categorized and the known exemplars determines, together with response determinism and category bias, the categorization response probabilities.

similarity of the i th stimulus to category A is the sum of the similarities to all the stimuli in the category: $s_{iA} = \sum_{j \in A} s_{ij}$. Finally, the probability of a category response placing the i th stimulus in category A is $p_{iA} = \beta_A s_{iA}^\gamma / \sum_C \beta_C s_{iC}^\gamma$, where the index C is across all possible categories, β_C is a response bias to category C, and γ controls the extent to which responding is deterministic or probabilistic, with higher values corresponding to more determinism.

Wiener diffusion model of decision making

Sequential sampling models of decision making assume that evidence is gathered from a stimulus over time until enough has been gathered to make a choice (Luce, 1986). The Wiener diffusion model (Ratcliff & McKoon, 2008) model is a simple, but widely used, sequential sampling model for two-choice decisions. It assumes evidence takes the form of samples from a Gaussian distribution with mean ν . Total evidence starts at θ and is summed until it reaches a lower bound of zero or an upper bound of α . The decision made corresponds to the boundary reached, and the response time is proportional to the number of samples, with

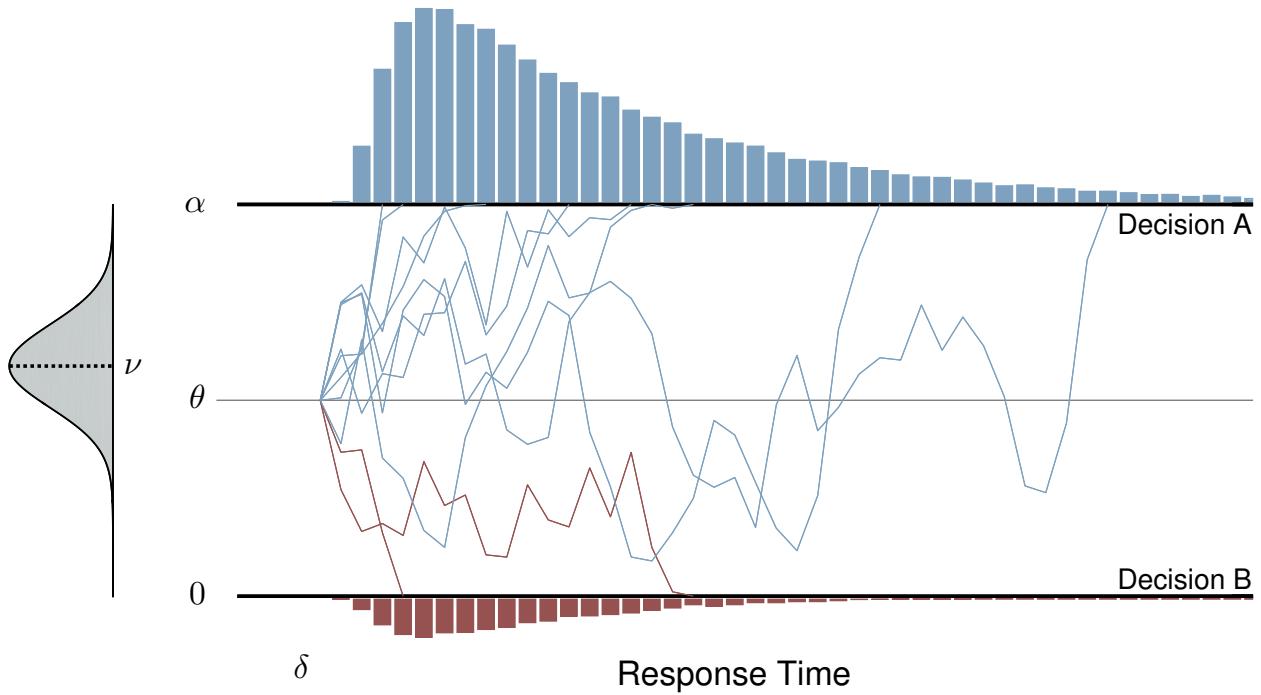


Figure 8.3: The Wiener diffusion model of decision making. A two-choice decision about a stimulus is made by sampling repeatedly from an evidence distribution for the stimulus, represented by a Gaussian distribution with mean ν . The samples are combined to form an evidence path, and a number of illustrative sample paths are shown. These paths start from an initial evidence value θ , and continue until they reach an upper bound of α or a lower bound of 0. The decision made corresponds to which boundary is reached. The response time is proportional to the number of samples collected, plus a constant δ representing the additional time needed to encode the stimuli and execute the response behavior. The decision and response time behavior is shown by the histograms above and below the decision boundaries. The histogram at each boundary is proportional to the response time distribution for that decision, and the area under each distribution represents the overall probability of that decision.

the inclusion of an additive offset δ .

The decision model is shown Figure 8.3. The stimulus provides evidence favoring decision A, because the mean ν of the Gaussian characterizing the evidence is greater than zero. The decision and response times are shown by the histograms at each boundary. The shape of the histogram represents the response time distribution for each type of decision, and the area under each distribution represents the probability of each decision. It is clear that decision A is more likely, and both response time distributions have a characteristic non-monotonic shape with a long-tailed positive skew.

The ν parameter, usually called the drift rate, corresponds to the informativeness of the stimulus. Larger absolute values of ν correspond to stimuli that provide stronger evidence in favor of one or other of the decisions. Smaller absolute values of ν correspond to less informative stimuli, with $\nu = 0$ representing a stimulus that provides no overall information

about which decision to make.

Figure 8.3 also shows a number of sample paths of evidence accumulation. All of the paths begin at the starting point θ , which is half-way between the boundaries at $\theta = \alpha/2$. Other starting points would favor one or other decision. The starting point parameter θ can theoretically be conceived as a bias in favor of one of the decisions. Such a bias could arise, psychologically, from prior evidence in favor of a decision, or as a way of incorporating utilities for correct and incorrect decisions of each type.

The α parameter, usually called boundary separation, corresponds to the caution used to make a decision, as manipulated, for example, by speed or accuracy instructions. Larger values of α lead to slower and more accurate decisions, while smaller values lead to faster but more error-prone decisions.

Finally, the offset δ corresponds to the component of the response time not accounted for by the sequential sampling process, such as the time taken to encode the stimulus and produce motor movements for a response. It is shown in Figure 8.3 as an offset at the beginning of the evidence sampling process, but could also be conceived as having two components, with an encoding part at the beginning, and a responding part at the end.

Sources for determining informative priors

In this section, we identify several sources of information that can be used in determining priors, and explain how these relate to the meaningful parameters of the three illustrative cognitive models.

Psychological and other scientific theory

The most important source of information for specifying priors in cognitive models is psychological theory. In cognitive modeling, likelihood functions largely reflect theoretical assumptions about cognitive processes. The exponential decay memory retention model commits to the way in which information is lost over time, assuming, in part, that the rate of this loss is greatest immediately after information is acquired. The categorization model commits to assumptions of exemplar representation, selective attention, and similarity comparisons in categorization. The decision model commits to the sequential sampling of information from a stimulus until a threshold level of evidence is reached. These assumptions are the cornerstones on which the likelihood functions of the models are founded. Analogously, theoretical assumptions about psychological variables should be the cornerstones on which priors are determined (Vanpaemel, 2009, 2010). Ideally, psychological theories should make assumptions about not just psychological processes, but also about the psychological variables that control those processes, leading to theory-informed priors.

One possibility is that theoretical assumptions dictate that some parameter values are impossible, consistent with the non-Bayesian restriction of the parameter space. In the memory retention model, the theoretical assumption that the probability of recall decreases over time constrains the memory retention parameter $\psi > 0$. In the categorization model, the

theoretical assumption that generalization gradients decrease as stimuli become less similar, constrains the parameter $\lambda \geq 0$ (Nosofsky, 1986; Shepard, 1987).

Other sorts of theorizing can provide more elaborate information about possible combinations of values for a set of parameters. Theories of attention, for example, often assume it is a capacity-limited resource. In the categorization model, this constraint is usually implemented as $\sum_k \omega_k = 1$, so that the values of the attention parameters collectively meet a capacity bound. In effect, the theoretical assumption still dictates that some parameter values are impossible, but now the constraint applies jointly to a set of parameters.

As theories become more general and complete they can provide richer information. Theory can provide information beyond which values are possible, and indicate which values are probable. The optimal-attention hypothesis (Nosofsky, 1986) assumes that people distribute their attention near optimally in learning a category structure for a set of stimuli. This assumption implies that values of the ω_k parameters that maximally separate the stimuli in each category from each other are expected. For example, in Figure 8.2, the stimuli in the two different categories vary more along the first than the second dimension. The optimal-attention hypothesis thus assumes that attention will be given to the first dimension to a level ω_1 somewhere near 1 that maximally distinguishes the two categories.

The optimality principle underlying the optimal-attention hypothesis could be extended to other cognitive models and phenomena. The principle that the most likely values of a parameter are those that maximize some aspect of behavioral performance is a generally applicable one. Optimality could be a fundamental source for setting priors in cognitive process models, but is currently under-used. Embedding the optimality principle within cognitive process models through priors would bring these models in closer contact with the successful rational models of cognition, where optimal behavior is a core theoretical assumption (e.g., Anderson, 1992; Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

A different example of using theory to develop a prior is provided by Rouder, Morey, Speckman, and Pratte (2007), who propose a mass-at-chance model for performance in subliminal priming tasks. Their theoretical expectations are that some people will perform at chance, but others will use a threshold-based detection process to perform above chance. Rouder et al. (2007, see especially their Figure 3) consider different theoretical possibilities about the distribution of detection probabilities for people performing above chance. One possibility is that all detection probabilities are equally likely, so that it is constrained between $\frac{1}{2}$ and 1. Another possibility is that they are only slightly above chance, so that, for example, few people are expected to have a detection probability higher than (say) 70%. A third possibility is that people who are not at chance all have perfect accuracy, so that there are only two possible detection probabilities, $\frac{1}{2}$ and 1. Rouder et al. (2007) consider only the first two options to be reasonable, and express this theoretical assumption by constraining a variance parameter to be smaller than 1. In this way, Rouder et al. (2007) establish a direct link between substantive theoretical assumptions about the nature of people's performance on the task and an exact range constraint on a variance parameter.

In some modeling situations, the likelihood can carry little theoretical content, and the theoretically most-relevant information is about the parameters. One example is provided by Lee (2016), in a Bayesian implementation of a model originally developed by Hilbig and

Moshagen (2014), for inferring which of a number of decision strategies people used in a cue-based decision-making task. The likelihood function is made up of simple binomial distributions, corresponding to how often an alternative is chosen for the trials within each decision type. Because different strategies predict different choice patterns, all of the important theoretical content is reflected in constraints on the choice parameters within the binomial distributions. For example, the new strategy introduced by Hilbig and Moshagen (2014) assumes an ordering for the probability of choice of different types of questions, and this information is represented by order constraints on the parameters corresponding to these probabilities in a joint prior. A similar earlier example in which the prior is theoretically more important than the likelihood is provided by Myung, Karabatsos, and Iverson (2005), who use order constraints on the parameters representing probabilities, to formalize several decision-making axioms such as the monotonicity of joint receipt axiom and the stochastic transitivity axiom.

Finally, we note that sciences other than psychology can and should provide relevant theoretical information. Physics, for example, provides the strong constraint—unless the controversial assumption of the existence of extra-sensory perception is made—that an item in a memory task cannot be rehearsed before it has been presented. This means, in the memory model, that each τ_i rehearsal parameter is constrained not to come before the actual time t_i the item was presented, so that $\tau_i \geq t_i$. Another example of the potential relevance of multiple other scientific fields to determine priors is provided by the offset parameter δ in the decision model. Neurobiological and chemical processes, such as the time taken for physical stimulus information to transmit through the brain regions responsible for low-level visual processing, should constrain the component of this parameter that corresponds to the time needed to encode stimuli. Physiological theories specifying, for example, distributions of the speeds of sequences of motor movements, should constrain the component of the parameter that corresponds to the time taken to produce an overt response. Thus, a theoretically meaningful prior for δ in the decision model could potentially be determined almost entirely by theories from scientific fields outside cognitive psychology.

Logic and invariances

The meaning of parameters can have logical implications for their prior distribution. Logic can dictate, for example, that some values of a parameter are impossible (Taagepera, 2007). Probabilities are logically constrained to be between 0 and 1, and variances and other scale parameters are constrained to be positive. In the memory and decision models, the probability parameters ϕ , β , and θ are both logically constrained to be between 0 and 1.

The nature of a modeling problem can also provide logical constraints. The decision model has no meaning unless the starting point θ is between 0 and the boundary α , and has the same substantive interpretation under the transformation $(\alpha, \theta) \rightarrow (-\alpha, -\theta)$ that “flips” the boundary and starting point below zero. This invariance leads to the constraints $\alpha, \theta > 0$ and $0 < \theta < \alpha$ to make the model meaningful.

In general, superficial changes to a modeling problem that leave the basic problem unchanged should not affect inference, and priors must be consistent with this. In our memory and decision models, for example, inferences should not depend on whether time is measured

in seconds of milliseconds, and the way priors over (ϕ, ψ, τ) and $(\alpha, \theta, \nu, \delta)$ are determined should lead to the same results regardless of the unit of measurement. This is a specific example of the general principle of transformation invariance, which requires that priors lead to the same result under transformations of a problem that change its surface form, but leave the fundamental problem itself unchanged (Lee & Wagenmakers, 2005). In the time scale example, the transformation is scalar multiplication of a measurement scale. In general, the transformation can involve much more elaborate and abstract manipulations of the inference problem being posed, as in Jaynes' (2003, Ch. 12) discussion of a famous problem in statistics known as Bertrand's paradox. The problem involves the probability of randomly thrown sticks intersecting a circle and is notorious for having different reasonable priors lead to different inferences. By considering logical rotation, translation, and dilation invariances for the circle, inherent in the statement of the problem, it is possible to determine an appropriate and unique prior. Motivated by these sorts of examples, we think that transformation invariance is a potentially important principle for determining priors. It is difficult, however, to find examples in cognitive modeling, and we believe more effort should be devoted to exploring the possibilities of this approach.

Previous data and modeling

Cognitive psychology has a long history as an empirical science, and has accumulated a wealth of behavioral data. Empirical regularities for basic cognitive phenomena are often well established. These regularities provide an accessible and substantial source of information for constructing priors. For example, response time distributions typically have a positive skew (e.g., Luce, 1986) and people often probability match in categorization, which means their probability of choosing each alternative is given by the relative evidence for that alternative (Shanks, Tunney, & McCarthy, 2002). This last observation is a good example of how empirical regularities can help determine a prior, and is applicable to the γ parameter in the categorization model. Different values of this parameter correspond to different assumptions about how people convert evidence for response alternatives into a single choice response. When $\gamma = 1$, decisions are made by probability matching. As γ increases above one, decision making becomes progressively more deterministic in choosing the alternative with the most evidence. As γ decreases below one, the evidence plays a lesser role in guiding the choice until, when $\gamma = 0$, choices are made at random. Thus, previous empirical findings that provide evidence as to whether people respond deterministically, probability match, and so on, can naturally provide useful information for determining a data-informed prior over the γ parameter (e.g., Lee, Abramyan, & Shankle, 2016).

Cognitive psychology is also a model-based science, and so there are many reported applications of models to data. These efforts provide inferences about parameters that can inform the development of priors. For each of the memory, categorization, and decision models, there are many published relevant applications to data, including inferred parameter values (e.g., Nosofsky, 1991; Ratcliff & Smith, 2004; Rubin & Wenzel, 1996). The approach of relying on previous parameter inferences to determine priors for related models is becoming more frequent in cognitive modeling. Some recent examples include Gu et al. (2016) in psychophysics, Gershman (2016) in reinforcement learning models, Vincent (2016) in the

context of temporal discounting, Wiehler, Bromberg, and Peters (2015) for different clinical sub-populations in the context of gambling, and Donkin, Tran, and Le Pelley (2015) in the context of a visual working memory model. In an interesting application of the latter model, Kary, Taylor, and Donkin (2015) used vague priors for key parameters, and used the data from the first half of their participants to derive the posterior distributions. These posteriors were subsequently used as a basis for priors in the analysis of the data from the remaining half of the participants.

Elicitation

There is a reasonably well-developed literature on methods designed to elicit priors from people (e.g., Albert et al., 2012; Garthwaite, Kadane, & O'Hagan, 2005; Kadane & Wolfson, 1998; O'Hagan et al., 2006). These methods are used quite extensively in modeling in some empirical sciences, but do not seem to be used routinely in cognitive modeling. Elicitation methods are designed to collect judgments from people—often with a focus on experts—that allow inferences about a probability distribution over unknown quantities. The most common general approach involves asking for estimates of properties of the required distribution. This can be as simple as asking for a minimum and maximum possible value, or the bounds on (say) an 80% credible interval for an observed quantity.

These elicitation methods can ask directly about latent parameters of interest, or about predicted observable quantities implied by values of those parameters. Obviously, when elicitation focuses on quantities related to the parameters, rather than the parameters themselves, a model is needed to relate people's judgments to the desired probability distributions. For example, in a signal detection theory setting, it is possible to elicit distributions for discriminability and bias parameters directly, or infer them from elicited hit and false-alarm rates based on a standard model. The logical end-point of asking about quantities implied by parameters is to ask about idealized data (Winkler, 1967). This is a potentially very useful approach, because often experts can express their understanding most precisely and accurately in terms of data. Kruschke (2013) provides a good example of this approach for data-analytic models in psychology, and it is clear it generalizes naturally to cognitive models.

Another approach to constructing elicitation-based priors used in applied settings require a series of judgments between discrete options, from which a probability distribution representing uncertainty can be derived (e.g., Welsh, Begg, Bratvold, & Lee, 2004). Along these lines, one potentially useful recent development is the elicitation procedure known as iterated learning (Kalish, Griffiths, & Lewandowsky, 2007; Lewandowsky, Griffiths, & Kalish, 2009). This clever procedure requires a sequence of people to do a task, such as learning a category structure, or the functional relationship between variables. Each person's task depends on the answers provided by the previous person, in a way that successively amplifies the assumed common prior information, or inductive bias, people bring to the task. Applying this procedure to categorization, Canini, Griffiths, Vanpaemel, and Kalish (2014) found that, when learning categories, people have a strong bias for a linear category boundary on a single dimension, provided that such a dimension can be identified. Translating this observation to the ω_k parameters in the categorization model implies that, in absence of any

other information about category structures, these parameters are expected to be close to 0 or 1. It is a worthwhile topic for future research to find ways of formally translating this sort of information into a prior for a cognitive model.

Methods for determining informative priors

The sources of information identified in the previous section are only pre-cursors to the complete formalization of a prior distribution. Knowing, for example, that some values of a parameter are theoretically impossible does not determine what distribution should be placed on the possible values. In this section, we identify some methods for taking relevant information, and using it to construct a formal prior distribution.

Constraint satisfaction

If available information, whether by theoretical assumption, out of logical necessity, or from some other source, constrains parameter values, these constraints can be used as bounds. To determine the prior distribution within these bounds, the maximum-entropy principle provides a powerful and general approach (Jaynes, 2003, Ch. 11; Robert, 2007). Conceptually, the idea of maximum entropy is to specify a prior distribution that satisfies the constraints, but is otherwise as uninformative as possible. In other words, the idea is for the prior to capture the available information, but no more. Common applications of this approach in cognitive modeling include setting uniform priors between 0 and 1 on probabilities, setting a form of inverse-gamma prior on variances (see Gelman, 2006, for discussion), and enforcing order constraints between parameters (e.g., Hoijtink, Klugkist, & Boelen, 2008; Lee, 2016).

A good example of applying the maximum-entropy principle to order constraints involves the τ_i rehearsal parameters in the memory model, if they are subject to the constraint that an item cannot be rehearsed before it has been presented. Figure 8.4 shows the resultant joint prior on (τ_1, τ_2, τ_3) if the three study items are presented at times t_1 , t_2 , and t_3 . Only rehearsal parameter combinations that are in the shaded cube have prior density. The uniformity of the prior in this region follows from the maximum-entropy principle, which ensures that it satisfies the known constraints about when the items could be rehearsed, but otherwise carries as little information as possible.

More general applications of the maximum-entropy principle are rare in the cognitive modeling literature. Vanpaemel and Lee (2012) present an example that is conceptually close, relating to setting the prior on the attention-weight parameter ω in the categorization model. The prior is assumed to be a beta distribution, and the optimal-attention hypothesis is used to set the mode of the prior to the value that best separates the stimuli from the different categories. The optimal-attention hypothesis, however, is not precise enough to determine an exact shape for the prior, but the precision of the beta distribution could have been determined in a more principled way by maximum-entropy methods. This would have improved on the heuristic approach actually used by Vanpaemel and Lee (2012) to set the precision. We think maximum-entropy methods are under-used, and that they are an approach cognitive modeling should adopt and develop, especially given the availability of

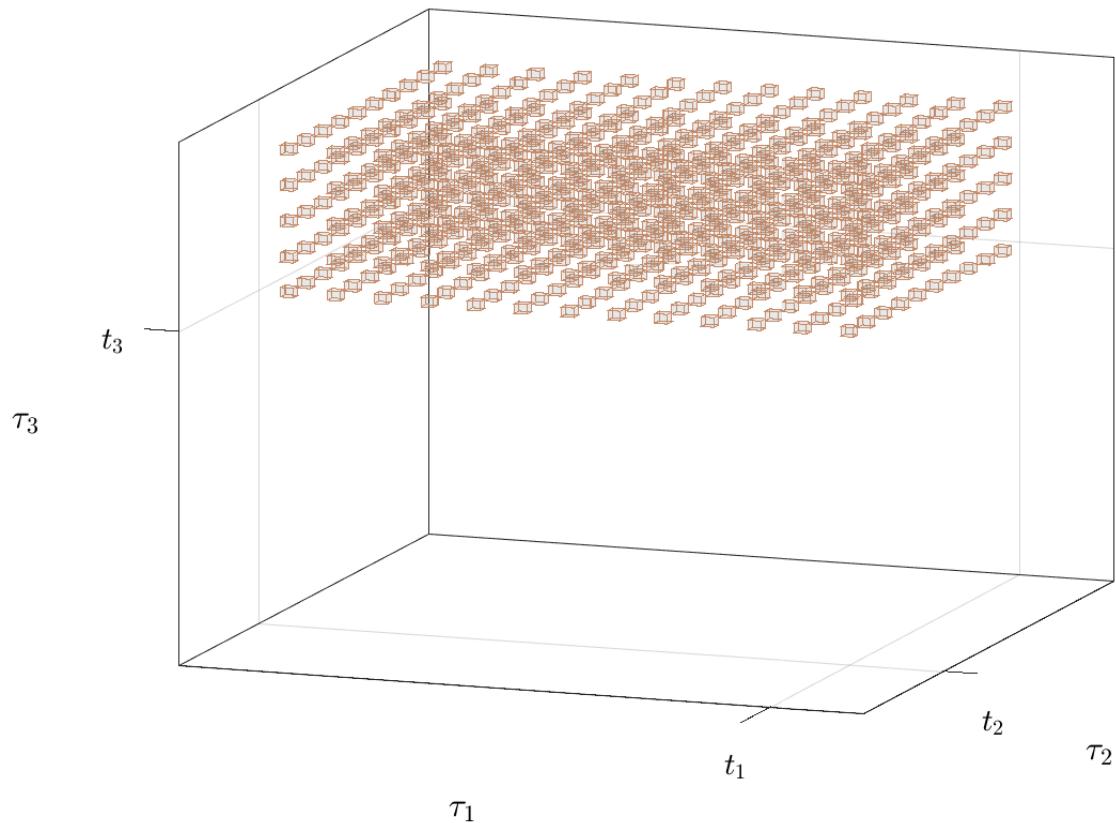


Figure 8.4: A prior specified by constraint satisfaction for the memory retention model. The three axes correspond to the last rehearsal times of three studied items, represented by the model parameters τ_1 , τ_2 , and τ_3 . The case considered involves these items having been first presented at known times t_1 , t_2 , and t_3 . The shaded region corresponds to the set of all possible rehearsal times (τ_1, τ_2, τ_3) that satisfy the logical constraint that an item can only be rehearsed after it is presented, so that $\tau_1 \geq t_1$, $\tau_2 \geq t_2$, and $\tau_3 \geq t_3$. The uniform distribution of prior probability within this constraint satisfaction region follows from the maximum-entropy principle.

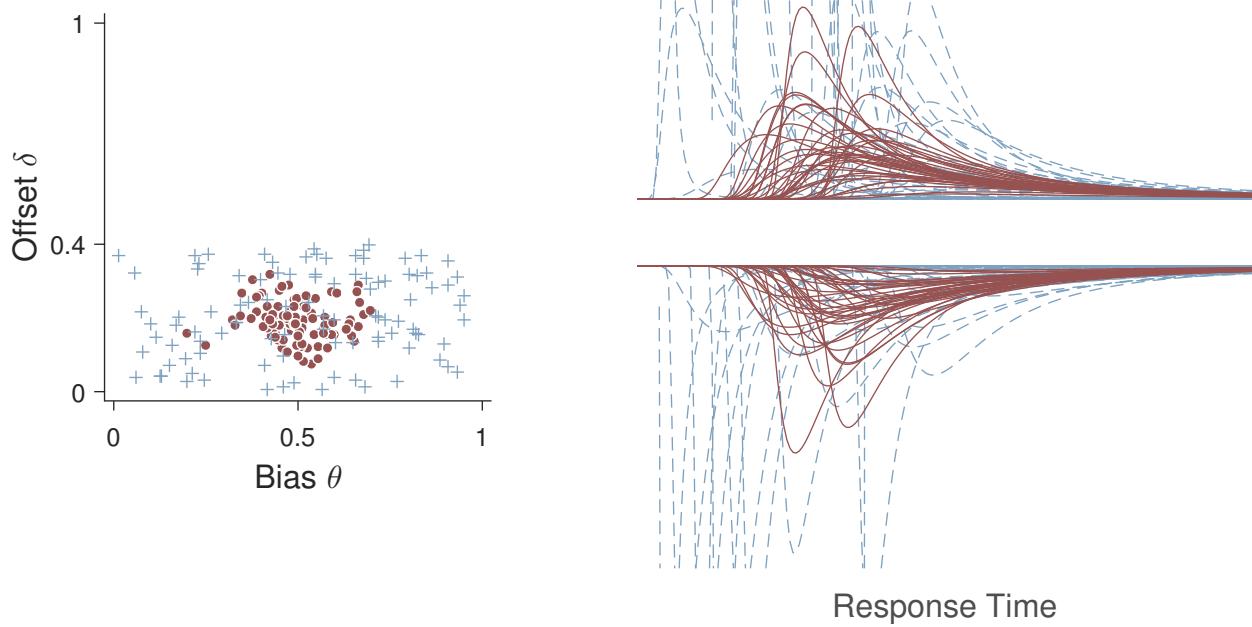


Figure 8.5: Developing a prior distribution using prior prediction for the decision model. The left panel shows the joint parameter space for the bias θ and offset δ parameters. The right panel shows the joint decision and response time distributions generated by the model. Two specific prior distributions are considered, represented by circles and crosses in the parameter space, with corresponding solid and broken lines in the model space. The prior represented by the circles makes stronger assumptions about both bias and offset, and predicts a more reasonable set of response time distribution than the vaguer prior represented by the crosses.

general statistical results that relate known constraints to maximum-entropy distributions (e.g., Lisman & Van Zuylen, 1972).

Prior prediction

By specifying a likelihood and a prior, it is possible to calculate the prior predictive distribution, which is a prediction about the relative probability of all possible data sets, based solely on modeling assumptions. If information is available about possible or plausible data patterns, most likely based on previously established empirical regularities or on elicitation, then one approach is to develop a prior distribution that leads to prior predictive distributions consistent with this information. A very similar approach is Parameter Space Partitioning (PSP: Pitt, Kim, Navarro, & Myung, 2006), which divides the entire parameter space into mutually exclusive regions that correspond to different qualitative data patterns a model can generate. Priors can then be determined by favoring those regions of the parameter space that generate data patterns consistent with expectations, and down-weighting or excluding regions corresponding to less plausible or implausible data patterns.

A closely-related approach involves considering the predictions over psychologically mean-

ingful components of a model that are implied by priors over their parameters. If information is available about the plausible form of these parts of models, most likely based on theory, it makes sense to define parameter priors that produce reasonable prior distributions for them. Figure 8.5 shows an example of this second approach using the decision model. Each combination of the starting point θ and offset δ parameters, which lie in the two-dimensional parameter space on the left, corresponds to a single joint decision and response time distribution for the two choices, shown on the right. Two different joint prior distributions over the parameters are considered. The first prior distribution, shown by circles in parameter space, has a truncated Gaussian prior for θ with a mean of 0.5 and a standard deviation of 0.1 in the valid range $0 < \theta < 1$, and a truncated Gaussian prior for δ with a mean of 0.2 and a standard deviation of 0.05 in the valid range $\delta > 0$. The second prior, shown by the crosses, simply uses uniform priors on reasonable ranges for the parameters: $0 < \theta < 1$, and $0 < \delta < 0.4$.

The consequences of these different assumptions are clear from the corresponding distributions shown in the model space, which shows response time distributions generated by the decision models corresponding to both priors, for the same assumptions about boundary separation and the distribution of drift rates. The predictions of the decision model with the first prior distribution, shown by solid lines, cover the sorts of possibilities that might be expected, in terms of their qualitative position and shape. The predictions for the second prior distribution, shown by broken lines, however, are much less reasonable. Many of the predicted response time distributions start too soon, and are too peaked. These weaknesses can be traced directly to the vague priors allowing starting points too close to the boundaries, and permitting very fast non-decision times. This analysis suggests that the sorts of assumptions about the starting point and offset made in forming the first prior may be good ones for the decision model. In this way, the relationship between prior distributions and psychologically interpretable components of the model provides a natural way to apply relevant knowledge in developing priors.

Using prior prediction to determine prior distributions in cognitive modeling is a general and relatively easy approach. Theorists often have clear expectations about model components like retention functions, generalization gradients, or the shapes of response time distributions, as well as about the data patterns that will be observed in specific experiments, which can be examined in prior predictive distributions. While it is currently hard to find cognitive modeling examples of priors being developed by the examination of prior predictions (see Lee, 2015; Lee & Danileiko, 2014, for exceptions), we expect this state of affairs will change quickly. One reason for this optimism is that prior predictions are slowly starting to appear in the cognitive modeling literature, with goals that are closely related to setting priors. For example, Kary et al. (2015) and Turner, Dennis, and Van Zandt (2013) examine the prior predictions of memory models, as a sanity check before application. In addition, prior predictive distributions have been used for assessing model complexity (Vanpaemel, 2009), for evaluating model falsifiability, and for testing a model against empirical data (Vanpaemel, submitted).

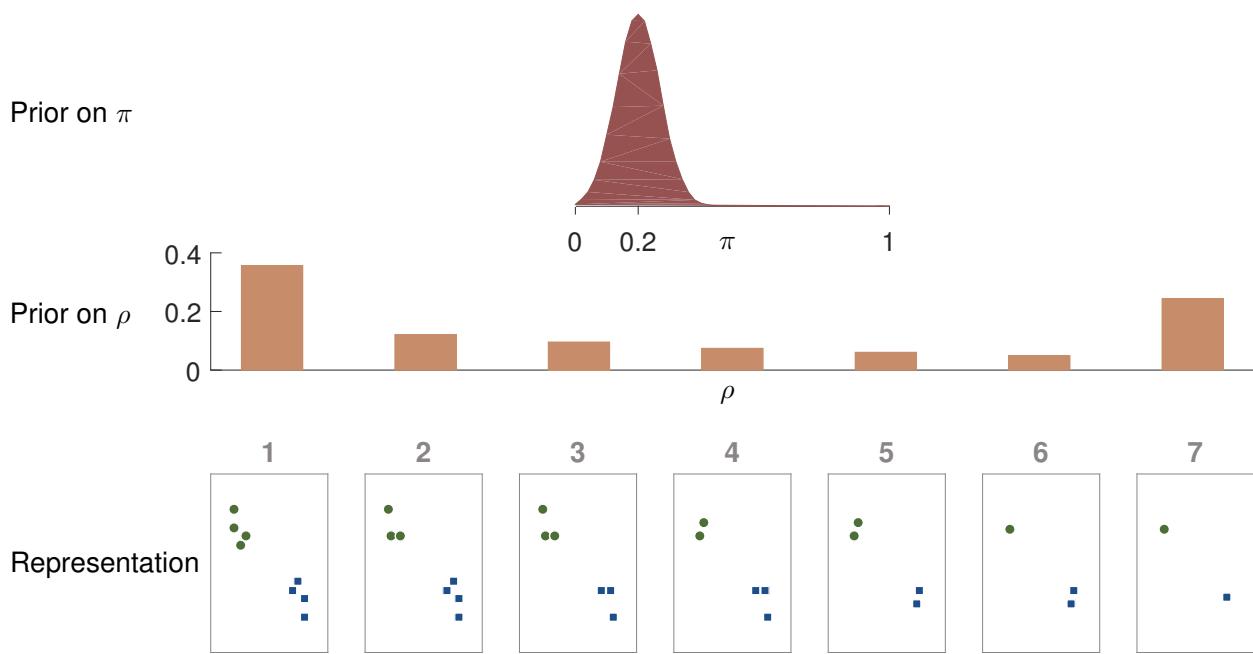


Figure 8.6: A hierarchical approach to determining a prior distribution for the representation index parameter ρ in an expanded version of the categorization model. The top panel shows an assumed prior distribution over a parameter π that corresponds to the probability of merging a pair of stimuli in an exemplar representation. The bottom panels show a selection of 7 possible representations generated by this merging process, for a categorization problem with four stimuli in each of two categories, distinguished as circles and squares. The full exemplar representation is shown on the left, the prototype representation is shown on the right, and some of the representations with intermediate levels of abstraction are shown between. The bar graph in the middle panel shows the prior probability on the representational index parameter ρ implied by the merging process and the prior distribution on π .

Hierarchical extension

An especially important method for developing priors in cognitive modeling involves extending the cognitive model itself. The basic idea is to extend the model so that priors on parameters are determined as the outcome of other parts of an extended model. This involves incorporating additional theoretical assumptions into the model, and is naturally achieved by hierarchical or multi-level model structures (Lee, 2011; Vanpaemel, 2011). None of the illustrative memory, categorization, or decision models, as we presented them, have this property, which is representative of the field as a whole. The parameters in these models represent psychological variables that initiate a data generating process, and so priors must be placed explicitly on these parameters. The key insight of the hierarchical approach is that these psychological variables do not exist in isolation in a complete cognitive system, but can be conceived as the outcomes of other cognitive processes. Including those other processes within a more complete model thus naturally defines a prior for the original parameters.

An example of this approach is provided by Lee and Vanpaemel (2008), who focus on

the Varying Abstraction Model (VAM: Vanpaemel & Storms, 2008). This model expands the categorization model by allowing for different sorts of category representations, ranging from an exemplar representation in which every stimulus in each category is represented, to a prototype representation in which each category is represented by a single summary point. Some of these possibilities are shown in the 7 bottom panels in Figure 8.6, for a case in which there are two categories with four stimuli each. The representation on the far left is the exemplar representation, as assumed by the original categorization model, while the representation on the far right is the prototype representation. The intermediate representations show different levels of abstraction, as the detail of exemplar representation gives way to summary representations of the categories. The inference about which representation is used is controlled by a discrete parameter ρ , which simply indexes the representations. In the example in Figure 8.6, ρ is a number between 1 and 7, and requires a prior distribution that gives the prior probabilities to each of these 7 possibilities.

Lee and Vanpaemel (2008) introduce a hierarchical extension of the VAM that is shown by the remainder of Figure 8.6. A new cognitive process is included in the model, which generates the different possible representations. This process begins with the exemplar representation, but can successively merge pairs of stimuli. At each stage, the probability of a merge is given by a new model parameter π . At each stage in the merging process, two stimuli are merged with probability π , otherwise the merging process stops and the current representation is used. Thus, there is probability $1 - \pi$ that the full exemplar representation is used, probability $\pi(1 - \pi)$ that a representation with a single merge is used, and so on. Having formalized this merging process as a model of representational abstraction, a prior over the parameter π automatically corresponds to a prior over the indexing parameter ρ . Figure 8.6 shows a Gaussian prior over π with a mean near the merge probability 0.2, and the bar graph shows the implied prior this places on ρ for the 7 different representations. Ideally, the sources and methods discussed earlier should be used to set the top-level prior on π , but its impact even with the current less formal approach is clear. More prior mass is placed on the exemplar and prototype representations, while allowing some prior probability for the intermediate representations. This prior on ρ is non-obvious, and seems unlikely to have been proposed in the original non-hierarchical VAM. In the hierarchical approach in Figure 8.6, it arises through psychological theorizing about how different representations might be generated by merging stimuli, and related prior assumptions about the probability of each merge.

The hierarchical approach to determining priors is broadly applicable, because it is a natural extension of theory- and model-building. It is naturally also applied, for example, in both the memory and decision models. In the memory model, a theory of rehearsal should automatically generate a prior for the τ parameters. For example, one prominent idea is that rehearsal processes are similar to free recall processes themselves (e.g., Rundus, 1971; Tan & Ward, 2008). Making this assumption, it should be possible to make predictions about whether and when presented items will be rehearsed—in the same way it is possible to make predictions about observed recalled behavior itself—and thus generate a prior for the latent rehearsal τ parameters. In the decision model, the boundary separation parameter α could be modeled as coming from control processes that respond to task demands, such as speed or accuracy instructions, as well as the accuracy of previous decisions. There are some cognitive

models of these control processes, involving, for example, theories of reinforcement learning (Simen, Cohen, & Holmes, 2006), or self-regulation (Lee, Newell, & Vandekerckhove, 2015; Vickers, 1979), that could augment the decision model to generate the decision bound, and thus effectively place a prior on its possible values.

Benefits of informative priors

Capturing theoretical, logical, or empirical information in priors offers significant benefits for cognitive modeling. For example, the additional information priors provide can solve basic statistical issues, related to model identifiability. These occur regularly in cognitive models that use latent mixtures, which is sometimes done to model qualitative or discrete individual differences. Latent-mixture models involve a set of model components that mix to produce data, and are notorious for being statistically unidentifiable, in the sense that the likelihood of data is the same under permutation of the mixture components (Marin, Mengersen, & Robert, 2011). The use of priors that give each component a different meaning—by, for example, asserting that one sub-group of people has a higher value on a parameter than the other sub-group—makes the model interpretable, and makes it easier to analyze (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014).

Theory-informed priors can address modeling problems relating not only to statistical ambiguity, but also those relating to theoretical ambiguity. The starting point parameter θ in the decision model provides a good example. It has sensible psychological interpretations as a bias capturing information about base-rate of correct decisions on previous trials, or as an adjustment capturing utility information about payoffs for different sorts of correct or incorrect decisions. In practice, these different psychological interpretations will typically correspond to different priors on θ and, in this sense, specifying a prior encourages a modeler to disambiguate the model theoretically.

Informative priors often make a model simpler, by constraining and focusing its predictions. The γ parameter in the categorization model provides an intriguing example of this. Sometimes the γ parameter is not included in the categorization model, on the grounds that its inclusion increases the complexity of the model (J. D. Smith & Minda, 2002; see also Vanpaemel, 2016). It turns out, however, that including γ with a prior that emphasizes the possibility of near-deterministic responding, by giving significant prior probability to γ values much greater than 1, can result in a simpler model. This is because the range of predictions becomes more constrained as deterministic responding is given higher prior probability. This example shows that equating model complexity with counts of parameters can be mis-leading, and that the omission of a parameter does not necessarily represent theoretical neutrality or agnosticism. The omission of the γ parameter corresponds to a strong assumption that people always probability match, which turns out to make the model flexible and imprecise in its predictions. Thus, in this case, a prior on the γ parameter that captures additional psychological theory, by allowing for both probability matching and more deterministic responding, reduces the model's complexity.

Constraining predictions in this sort of way has the important scientific benefit that it increases what Popper (1959) terms the “empirischer Gehalt” or empirical content of a model

(see also Glöckner & Betsch, 2011; Vanpaemel & Lee, 2012). Empirical content corresponds to the amount of information a model conveys, and is directly related to falsifiability and testability. As a model that makes sharper predictions is more likely to rule out plausible outcomes, it runs a higher risk of being falsified by empirical observation, and thus gains more support from confirmation of its predictions (Lakatos, 1978; Roberts & Pashler, 2000; Vanpaemel, submitted).

Perhaps most importantly, using priors to place additional substantive content in a model makes the model a better formalization of the theory on which it is based. As noted by Vanpaemel and Lee (2012), the categorization model is a good example of this. Most of the theoretical assumptions on which the model is explicitly founded—involving exemplar representation, selective attention, and so on—are formalized in the likelihood of the model. The theoretical assumption that is conspicuously absent is the optimal-attention hypothesis. The difference is that most of the assumptions are about psychological processes, and so are naturally formalized in the likelihood function. The optimal-attention assumption, however, relates to a psychological variable, and so is most naturally formalized in the prior.

A similar story recently played out in the literature dealing with sequential sampling models very much like the decision model. In a critique of these sorts of decision models, Jones and Dzhafarov (2014a) allowed the drift-rate parameter ν to have little variability over trials. P. L. Smith, Ratcliff, and McKoon (2014) argued that this allowance was contrary to guiding theory, pointing out that it implied a deterministic growth process, which conflicts with the diffusion process assumptions on which the model is founded (Ratcliff & Smith, 2004). Jones and Dzhafarov (2014b) rejoindered that there is nothing in the standard model-fitting approach used by Ratcliff and Smith (2004) and others that precludes inferring parameters corresponding to the reduced deterministic growth model. From a Bayesian perspective, the problem is that theoretically available information about the variability of the distribution affecting the drift rate was not formalized in the traditional non-Bayesian modeling setting used by Ratcliff and Smith (2004). Because the theory makes assumptions about the plausible values of a parameter, rather than a process, it is naturally incorporated in the prior, which requires a Bayesian approach.

Discussion

A cognitive model that provides only a likelihood is not specific or complete enough to make detailed quantitative predictions. The Bayesian requirement of specifying the prior distribution over parameters produces models that do make predictions, consistent with the basic goals of modeling in the empirical sciences (Feynman, 1994, Chapter 7). We have argued that not giving sufficient attention to the construction of a prior that reflects all the available information corresponds to “leaving money on the table” (Weiss, 2014).

Using priors for cognitive modeling, however, comes with additional responsibilities. One consequence of using priors for cognitive modeling is the need to conduct additional sensitivity analyses. As our survey of information sources and methods makes clear, there is no automatic procedure for determining a prior. A combination of creative theorizing, logical analysis, and knowledge of previous data and modeling results is required. Different conclu-

sions can be reached using the same data for different choices of priors, just as they would if different likelihoods were used. This means a sensitivity analysis is appropriate when the available information and methods do not allow the complete determination of a prior distribution, and there is consequently some subjectivity or arbitrariness in the specification of the prior.

There is nothing inherent to the prior that makes it uniquely subject to some degree of arbitrariness. It is often the case that the likelihoods in models are defined with some arbitrariness, and it is good practice to undertake sensitivity analyses for likelihoods. Rubin and Wenzel (1996) consider a large number of theoretically plausible likelihoods for modeling memory retention, including many variants of exponential, logarithmic, and hyperbolic curves. A number of different forms of the GCM have been considered, including especially different response rules for transforming category similarity to choice probabilities (e.g., Nosofsky, 1986, 1992). Ratcliff (2013) reports a sensitivity analysis for some theoretically unconstrained aspects of the likelihood of a diffusion model of decision making. The same approach and logic applies to the part of cognitive modeling that involves choosing priors. Sensitivity analyses highlight whether and where arbitrariness in model specification is important—in the sense that it affects the inferences that address the current research questions—and so guides where clarifying theoretical development and empirical work is needed.

A standard concern in the application of Bayesian methods to cognitive modeling is that model selection measures like Bayes factors are highly sensitive to priors, but parameter inference based on posterior distributions and their summaries are far less sensitive. Part of the reason for the greater sensitivity of the Bayes factor probably stems from the fundamentally different inferential question it solves, and its formalization in optimizing zero-one loss. But it is also possible some of the perceived relative insensitivity of parameter inference to priors stems from the use of vague priors. It seems likely that informative priors will make inferences more sensitive to their exact specification. As a simple intuitive example, an informative prior that expresses an order constraint will dramatically affect inference about a parameter if the unconstrained inference has significant density around the values where the constraint is placed. In general, the heightened sensitivity of parameter inference to priors that capture all of the available information makes conceptual sense. These priors will generally make stronger theoretical commitments and more precise predictions about data, and Bayesian inferences will automatically represent the compromise between the information in the prior and in the data.

In this paper, we have identified sources of information that can be used to develop informative priors for cognitive models, have surveyed a set of methods that can be used for this development, and have highlighted the benefits of capturing the available information in the prior. The sources and methods we have discussed are not routinely used in cognitive modeling, and we certainly do not claim they are complete, nor that they constitute a general capability for all modeling challenges. In addition, the use of informative priors in cognitive modeling is not yet extensive or mature enough to provide a tutorial on best practice in the field. We hope, however, to have provided a useful starting point for determining informative priors, so that models can be developed that provide a more complete account of human cognition, are higher in empirical content, and make more precise, testable, falsifiable, and

useful predictions.

Acknowledgments

We thank Richard Morey for helpful discussions, and for drawing our attention to the Jeff Gill quotation. We also thank John Kruschke, Mike Kalish, and two anonymous reviewers for very helpful comments on earlier versions of this paper. The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (OT/11/032 and CREA/11/005).

References

- Albert, I., Donnet, S., Guienneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7, 503–532.
- Anderson, J. R. (1992). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14, 471–517.
- Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, 59, 132–150.
- Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, 21, 785–793.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781.
- Donkin, C., Tran, S. C., & Le Pelley, M. (2015). Location-based errors in change detection: A challenge for the slots model of visual working memory. *Memory & Cognition*, 43, 421–431.
- Edwards, A. F. W. (1991). Bayesian reasoning in science. *Nature*, 352, 386–387.
- Feynman, R. (1994). *The character of physical law*. Modern Press.
- Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680–701.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515–534.
- Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1–6.
- Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (Vol. 20). CRC press.

- Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, 6, 711–721.
- Gu, H., Kim, W., Hou, F., Lesmes, L. A., Pitt, M. A., Lu, Z.-L., et al. (2016). A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function. *Journal of Vision*, 16, 15–17.
- Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, 21, 1431–1443.
- Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.
- Jones, M., & Dzhafarov, E. N. (2014a). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, 121, 1–32.
- Jones, M., & Dzhafarov, E. N. (2014b). Analyzability, ad hoc restrictions, and excessive flexibility of evidence-accumulation models: Reply to two critical commentaries. *Psychological Review*, 121, 689–695.
- Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 3–19.
- Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14, 288–294.
- Kary, A., Taylor, R., & Donkin, C. (2015). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, 72, 210–219.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, 142, 573–603.
- Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.
- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73, 31–43.
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55, 1–7.
- Lee, M. D. (2015). Evidence for and against a simple interpretation of the less-is-more effect. *Judgment and Decision Making*, 10, 18–33.
- Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, 48, 29–41.
- Lee, M. D. (in press). Bayesian methods in cognitive modeling. In *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Fourth ed.).
- Lee, M. D., Abramyan, M., & Shankle, W. R. (2016). New methods, measures, and models for analyzing memory impairment using triadic comparisons. *Behavior Research Methods*, 48, 1492–1507.
- Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9, 259–273.

- Lee, M. D., Newell, B. R., & Vandekerckhove, J. (2015). Modeling the adaptation of the termination of search in human decision making. *Decision*, 1, 223–251.
- Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, 32, 1403–1424.
- Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, 112, 662–668.
- Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, 33, 969–998.
- Lisman, J., & Van Zuylen, M. (1972). Note on the generation of most probable frequency distributions. *Statistica Neerlandica*, 26, 19–23.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.
- Marin, J. M., Mengerson, K., & Robert, C. P. (2011). Bayesian modelling and inference on mixtures of distributions. In D. Dey & C. R. Rao (Eds.), *Essential Bayesian models. Handbook of statistics: Bayesian thinking – modeling and computation* 25. Elsevier.
- Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, 49, 205–225.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 3–27.
- Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honour of William K. Estes* vol. 1. Hillsdale, NJ: Lawrence Erlbaum.
- O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Wiley.
- Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113, 57–83.
- Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.
- Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, 120, 281–292.
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20, 873–922.
- Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111, 333–367.
- Robert, C. P. (2007). *The Bayesian choice*. New York, NY: Springer-Verlag.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358–367.
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, 14, 597–605.

- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734–760.
- Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*, 63–77.
- Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233–250.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
- Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks, 19*, 1013–1026.
- Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 800–811.
- Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review, 121*, 679–688.
- Taagepera, R. (2007). Predictive versus postdictive models. *European Political Science, 6*, 114–123.
- Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*, 535–542.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*, 1279–1285.
- Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers (2005). *Psychological Review, 112*, 669–674.
- Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review, 120*, 667–678.
- Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1919–1927). Red Hook, NY: Curran Associates Inc.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*, 491–498.
- Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology, 55*, 106–117.
- Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology, 72*, 183–190.
- Vanpaemel, W. (submitted). Complexity, data prior and the persuasiveness of a good fit: Comment on Veksler, Myers and Gluck (2015). *Manuscript submitted for publication*.
- Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the Generalized Context Model. *Psychonomic Bulletin & Review, 19*, 1047–1056.
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*, 732–749.
- Vickers, D. (1979). *Decision Processes in Visual Perception*. New York, NY: Academic Press.

- Vincent, B. (2016). Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior Research Methods*, 48, 1608–1620.
- Weiss, R. (2014). *Kathryn Chaloner 1954–2014*. (<https://faculty.biostat.ucla.edu/robweiss/node/169>)
- Welsh, M., Begg, S., Bratvold, R., & Lee, M. (2004). Problems with the elicitation of uncertainty. In *SPE Annual Technical Conference and Exhibition*. Richardson, TX: Society for Petroleum Engineers.
- Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, 105, 379–386.
- Wiehler, A., Bromberg, U., & Peters, J. (2015). The role of prospection in steep temporal reward discounting in gambling addiction. *Frontiers in Psychiatry*, 6, 112.
- Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 62, 776–800.

Introduction to Markov Chain Monte–Carlo Sampling

Don van Ravenzwaaij, Pete Cassey, and Scott D. Brown

Introduction

Over the course of the twenty-first century, the use of Markov chain Monte–Carlo sampling, or *MCMC*, has grown dramatically. But, what exactly is MCMC? And why is its popularity growing so rapidly? There are many other tutorial articles that address these questions, and provide excellent introductions to MCMC. The aim of this article is not to replicate these, but to provide a more basic introduction that should be accessible for even very beginning researchers. Readers interested in more detail, or a more advanced coverage of the topic, are referred to recent books on the topic, with a focus on cognitive science, by Lee and Wagenmakers (2013) and Kruschke (2014), or a more technical exposition by Gilks, Richardson, and Spiegelhalter (1996).

MCMC is a computer–driven sampling method (Gamerman & Lopes, 2006; Gilks et al., 1996). It allows one to characterize a distribution without knowing all of the distribution’s mathematical properties by randomly sampling values out of the distribution. A particular strength of MCMC is that it can be used to draw samples from distributions even when all that is known about the distribution is how to calculate the density for different samples. The name MCMC combines two properties: *Monte–Carlo* and *Markov chain*.¹ Monte–Carlo is the practice of estimating the properties of a distribution by examining random samples from the distribution. For example, instead of finding the mean of a normal distribution by directly calculating it from the distribution’s equations, a Monte–Carlo approach would be to draw a large number of random samples from a normal distribution, and calculate the sample mean of those. The benefit of the Monte–Carlo approach is clear: calculating the mean of a large sample of numbers can be much easier than calculating the mean directly from the normal distribution’s equations. This benefit is most pronounced when random samples are easy to draw, and when the distribution’s equations are hard to work with in

¹For these and other definitions, please see the glossary at the end of the paper.

other ways. The Markov chain property of MCMC is the idea that the random samples are generated by a special sequential process. Each random sample is used as a stepping stone to generate the next random sample (hence the *chain*). A special property of the chain is that, while each new sample depends on the one before it, new samples do *not* depend on any samples before the previous one (this is the “Markov” property).

MCMC is particularly useful in Bayesian inference because of the focus on posterior distributions which are often difficult to work with via analytic examination. In these cases, MCMC allows the user to approximate aspects of posterior distributions that cannot be directly calculated (e.g., random samples from the posterior, posterior means, etc.). Bayesian inference uses the information provided by observed data about a (set of) parameter(s), formally the *likelihood*, to update a *prior* state of beliefs about a (set of) parameter(s) to become a *posterior* state of beliefs about a (set of) parameter(s). Formally, Bayes’ rule is defined as

$$p(\mu|D) \propto p(D|\mu) \cdot p(\mu) \quad (9.1)$$

where μ indicates a (set of) parameter(s) of interest and D indicates the data, $p(\mu|D)$ indicates the posterior or the probability of μ given the data, $p(D|\mu)$ indicates the likelihood or the probability of the data given μ , and $p(\mu)$ indicates the prior or the a-priori probability of μ . The symbol \propto means “is proportional to”.

More information on this process can be found in Lee and Wagenmakers (2013), in Kruschke (2014), or elsewhere in this special issue. The important point for this exposition is that the way the data are used to update the prior belief is by examining the likelihood of the data given a certain (set of) value(s) of the parameter(s) of interest. Ideally, one would like to assess this likelihood for every single combination of parameter values. When an analytical expression for this likelihood is available, it can be combined with the prior to derive the posterior analytically. Often times in practice, one does not have access to such an analytical expression. In Bayesian inference, this problem is most often solved via MCMC: drawing a sequence of samples from the posterior, and examining their mean, range, and so on.

Bayesian inference has benefited greatly from the power of MCMC. Even in just in the domain of psychology, MCMC has been applied in a vast range of research paradigms, including Bayesian model comparison (Scheibehenne, Rieskamp, & Wagenmakers, 2013), memory retention (Shiffrin, Lee, Kim, & Wagenmakers, 2008), signal detection theory (Lee, 2008), extrasensory perception (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), multinomial processing trees (Matzke, Dolan, Batchelder, & Wagenmakers, 2015), risk taking (van Ravenzwaaij, Dutilh, & Wagenmakers, 2011), heuristic decision making (van Ravenzwaaij, Moore, Lee, & Newell, 2014) and primate decision making (Cassey, Heathcote, & Brown, 2014).

While MCMC may sound complex when described abstractly, its practical implementation can be very simple. The next section provides a simple example to demonstrate the straightforward nature of MCMC.

Example: In-Class Test

Suppose a lecturer is interested in learning the mean of test scores in a student population. Even though the mean test score is unknown, the lecturer knows that the scores are normally distributed with a standard deviation of 15. So far, the lecturer has observed a test score of a single student: 100. One can use MCMC to draw samples from the *target distribution*, in this case the posterior, which represents the probability of each possible value of the population mean given this single observation. This is an over-simplified example as there is an analytical expression for the posterior ($N(100, 15)$), but its purpose is to illustrate MCMC.

To draw samples from the distribution of test scores, MCMC starts with an initial guess: just one value that might be plausibly drawn from the distribution. Suppose this initial guess is 110. MCMC is then used to produce a chain of new samples from this initial guess. Each new sample is produced by two simple steps: first, a *proposal* for the new sample is created by adding a small random perturbation to the most recent sample; second, this new proposal is either accepted as the new sample, or rejected (in which case the old sample retained). There are many ways of adding random noise to create proposals, and also different approaches to the process of accepting and rejecting. The following illustrates MCMC with a very simple approach called the *Metropolis algorithm* (Smith & Roberts, 1993):

1. Begin with a plausible *starting value*; 110 in this example.
2. Generate a new proposal by taking the last sample (110) and adding some random noise. This random noise is generated from a *proposal distribution*, which should be symmetric and centered on zero. This example will use a proposal distribution that is normal with zero mean and standard deviation of 5. This means the new proposal is 110 (the last sample) plus a random sample from $N(0, 5)$. Suppose this results in a proposal of 108.
3. Compare the height of the posterior at the value of the new proposal against the height of the posterior at the most recent sample. Since the target distribution is normal with mean 100 (the value of the single observation) and standard deviation 15, this means comparing $N(100|108, 15)$ against $N(100|110, 15)$. Here, $N(\mu|x, \sigma)$ indicates the normal distribution for the posterior: the probability of value μ given the data x and standard deviation σ . These two probabilities tell us how plausible the proposal and the most recent sample are given the target distribution.
4. If the new proposal has a higher posterior value than the most recent sample, then accept the new proposal.
5. If the new proposal has a lower posterior value than the most recent sample, then randomly choose to accept or reject the new proposal, with a probability equal to the height of both posterior values. For example, if the posterior at the new proposal value is one-fifth as high as the posterior of the most recent sample, then accept the new proposal with 20% probability.

6. If the new proposal is accepted, it becomes the next sample in the MCMC chain, otherwise the next sample in the MCMC chain is just a copy of the most recent sample.
7. This completes one *iteration* of MCMC. The next iteration is completed by returning to step 2.
8. Stop when there are enough samples (e.g., 500). Deciding when one has enough samples is a separate issue, which will be discussed later in this section.

This very simple MCMC sampling problem only takes a few lines of coding in the statistical freeware program R, available online at cran.r-project.org. Code to do this may be found in Appendix A. The results of running this sampler once are shown in the left column of Figure 9.1. These samples can be used for Monte–Carlo purposes. For instance, the mean of the student population test scores can be estimated by calculating the sample mean of the 500 samples.

The top-left panel of Figure 9.1 shows the evolution of the 500 iterations; this is the Markov chain. The sampled values are centered near the sample mean of 100, but also contain values that are less common. The bottom-left panel shows the density of the sampled values. Again, the values center around the sample mean with a standard deviation that comes very close to the true population standard deviation of 15 (in fact, the sample standard deviation for this Markov chain is 16.96). Thus, the MCMC method has captured the essence of the true population distribution with only a relatively small number of random samples.

Limitations

The MCMC algorithm provides a powerful tool to draw samples from a distribution, when all one knows about the distribution is how to calculate its likelihood. For instance, one can calculate how much more likely a test score of 100 is to have occurred given a mean population score of 100 than given a mean population score of 150. The method will “work” (i.e., the sampling distribution will truly be the target distribution) as long as certain conditions are met. Firstly, the likelihood values calculated in steps 4 and 5 to accept or reject the new proposal must accurately reflect the density of the proposal in the target distribution. When MCMC is applied to Bayesian inference, this means that the values calculated must be posterior likelihoods, or at least be proportional to the posterior likelihood (i.e., the ratio of the likelihoods calculated relative to one another must be correct). Secondly, the proposal distribution should be symmetric (or, if an asymmetric distribution is used, a modified accept/reject step is required, known as the “Metropolis–Hastings” algorithm). Thirdly, since the initial guess might be very wrong, the first part of the Markov chain should be ignored; these early samples cannot be guaranteed to be drawn from the target distribution. The process of ignoring the initial part of the Markov chain is discussed in more detail later in this section.

The example MCMC algorithm above drew proposals from a normal distribution with zero mean and standard deviation 5. In theory, any symmetric distribution would have worked just as well, but in practice the choice of proposal distribution can greatly influence the performance of the sampler. This can be visualised by replacing the standard deviation

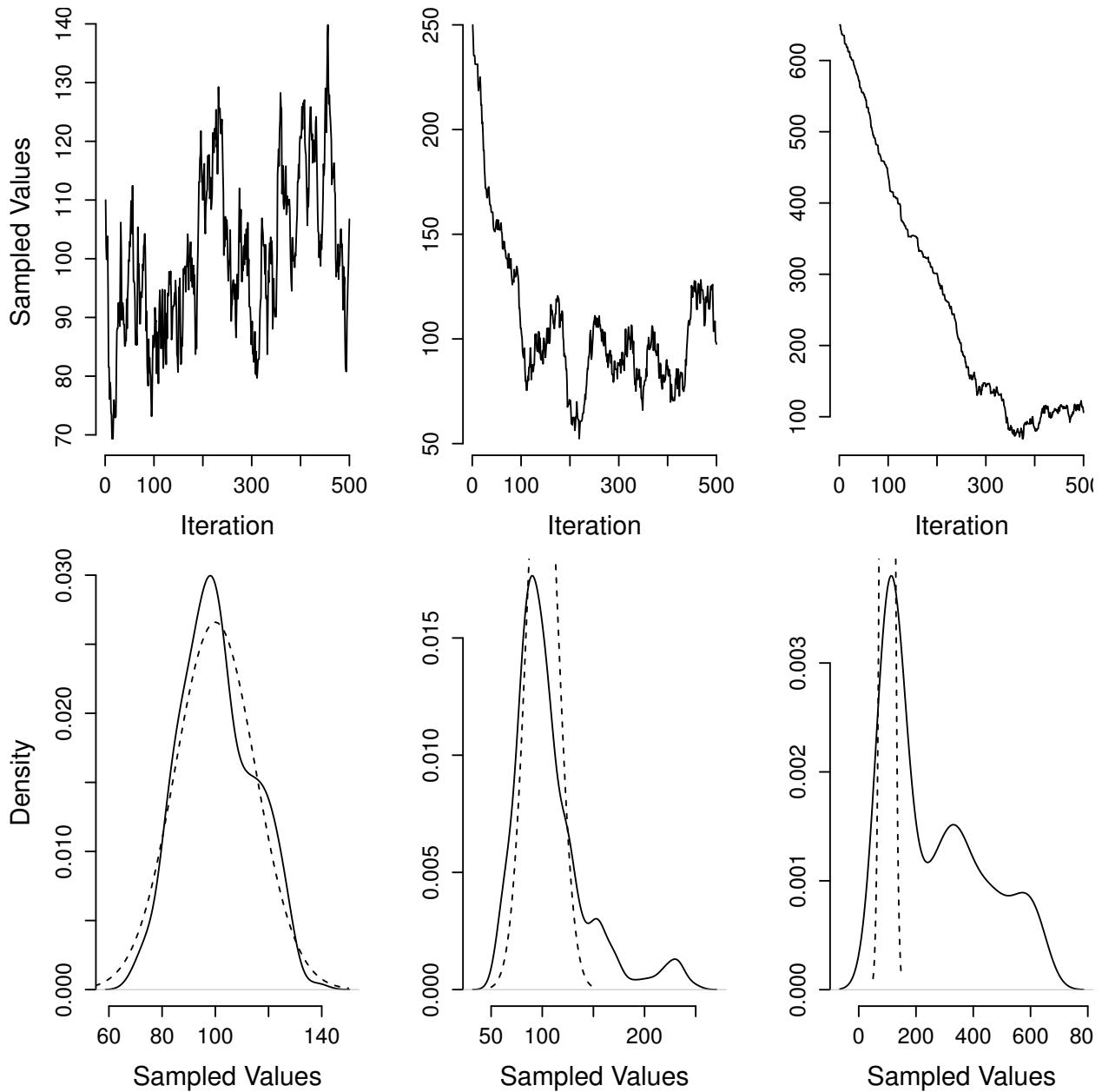


Figure 9.1: A simple example of MCMC. Left column: A sampling chain starting from a good starting value, the mode of the true distribution. Middle column: A sampling chain starting from a starting value in the tails of the true distribution. Right column: A sampling chain starting from a value far from the true distribution. Top row: Markov chain. Bottom row: sample density. The analytical (true) distribution is indicated by the dashed line.

for the proposal distribution in the above example with a very large value, such as 50. Then many of the proposals would be well outside the target distribution (e.g., negative test score proposals!) leading to a high *rejection rate*. On the other hand, with a very small standard deviation, such as 1, the sampler could take many iterations to converge from the starting

value to the target distribution. One also runs the risk of getting stuck in *local maxima*: areas where the likelihood is higher for a certain value than for its close neighbors, but lower than for neighbors that are further away.

The width of the proposal distribution is sometimes called a *tuning parameter* of this MCMC algorithm. The fact that the practical performance of the sampler can depend on the value of the tuning parameter is a limitation of the standard Metropolis–Hastings sampling algorithm, although there are many augmented methods that remedy the problem. For example, “auto-tuning” algorithms that adapt the width of the proposal distribution to the nature of the data and distribution (see Roberts & Rosenthal, 2009, for an overview).

The third condition, the fact that initial samples should be ignored as they might be very wrong, deals with a problem known as *convergence* and *burn-in*. For example, suppose the initial guess was one that was very unlikely to come from the target distribution, such as a test score of 250, or even 650. Markov chains starting from these values are shown in the middle and right columns of Figure 9.1. Examining the top–middle panel of Figure 9.1 shows that the Markov chain initially goes quickly down towards the true posterior. After only 80 iterations, the chain is then centered on the true population mean. Examining the top–right panel of Figure 9.1, which has an even more extreme starting point, demonstrates that the number of iterations needed to get to the true population mean — about 300 — is much larger than for better starting points. These two examples make it clear that the first few iterations in any Markov chain cannot safely be assumed to be drawn from the target distribution. For instance, including the first 80 iterations in the top–middle panel or those first 300 iterations in the top–right panel leads to an incorrect reflection of the population distribution, which is shown in the bottom–middle and –right panels of Figure 9.1.

One way to alleviate this problem is to use better starting points. Starting values that are closer to the mode of the posterior distribution will ensure faster burn–in and fewer problems with convergence. It can be difficult in practice to find starting points near the posterior mode, but maximum–likelihood estimation (or other approximations to that) can be useful in identifying good candidates. Another approach is to use multiple chains; to run the sampling many times with different starting values (e.g. with starting values sampled from the prior distribution). Differences between the distributions of samples from different chains can indicate problems with burn–in and convergence. Another element of the solution is to remove the early samples: those samples from the non–stationary parts of the chain. When examining again the chains in the top row of Figure 9.1, it can be seen that the chain in the top–left has come to some sort of an equilibrium (the chain is said to have “converged”). The chains in the top–middle and –right panel also converge, but only after about 80 and 300 iterations, respectively. The important issue here is that all the samples prior to convergence are *not* samples from the target distribution and must be discarded.

Deciding on the point at which a chain converges can be difficult, and is sometimes a source of confusion for new users of MCMC. The important aspect of burn–in to grasp is the post–hoc nature of the decision, that is, decisions about burn–in must be made after sampling, and after observing the chains. It is a good idea to be conservative: discarding extra samples is safe, as the remaining samples are most likely to be from the converged parts of the chain. The only constraint on this conservatism is to have enough samples after burn–in to ensure an adequate approximation of the distribution. Those users desiring a

more automated or objective method for assessing burn-in might investigate the \hat{R} statistic (Gelman & Rubin, 1992).

MCMC Applied to a Cognitive Model

We are often interested in estimating the parameters of cognitive models from behavioral data. As stated in the introduction, MCMC methods provide an excellent approach for parameter estimation in a Bayesian framework: see Lee and Wagenmakers (2013) for more detail. Examples of such cognitive models include response time models (Brown & Heathcote, 2008; Ratcliff, 1978; Vandekerckhove, Tuerlinckx, & Lee, 2011), memory models (Hemmer & Steyvers, 2009; Shiffrin & Steyvers, 1997; Vickers & Lee, 1997) and models based on signal detection theory (SDT: Green & Swets, 1966). Models based on SDT have had a seminal history in cognitive science, perhaps in part due to their intuitive psychological appeal and computational simplicity. The computational simplicity of SDT makes it a good candidate for estimating parameters via MCMC.

Suppose a memory researcher obtains data in the form of hits and false alarms from a simple visual detection experiment. Applying the SDT framework would allow the researcher to understand the data from a process, rather than descriptive (e.g. ANOVA) perspective. That is, estimating the parameters of the SDT model allows the researcher to gain an insight into how people make decisions under uncertainty. SDT assumes that when making a decision under uncertainty one needs to decide whether a certain pattern is more likely to be “signal” (e.g. a sign post on a foggy night) or merely “noise” (e.g. just fog). The parameters of SDT provide a theoretical understanding of how people distinguish between just noise and meaningful patterns within noise: sensitivity, or d' , gives a measure of the ability of the individual to distinguish between the noise and the pattern; criterion, or C , gives a measure of an individual’s bias, at what level of noise are they willing to call noise a meaningful pattern.

One way to estimate SDT parameters from data would be to use Bayesian inference and examine the posterior distribution over those parameters. Since the SDT model has two parameters (d' and C), the posterior distribution is bivariate; that is, the posterior distribution is defined over all different combinations of d' and C values. MCMC allows one to draw samples from this bivariate posterior distribution, as long as one can calculate the density for any given sample. This density is given by Equation 9.1: the likelihood of the hits and false alarms, given the SDT parameters, multiplied by the prior of those SDT parameters. With this calculation in hand, the process of MCMC sampling from the posterior distribution over d' and C is relatively simple, requiring only minor changes from the algorithm in the in-class test example above. The first change to note is that the sampling chain is multivariate; each sample in the Markov chain contains two values: one for d' and one for C .

The other important change is that the target distribution is a posterior distribution over the parameters. This allows the researcher to answer inferential questions, such as whether d' is reliably greater than zero, or whether C is reliably different from an unbiased value. To make the target distribution a posterior distribution over the parameters, the likelihood ratio in Step 3 above must be calculated using Equation 9.1. A simple working example of

such an MCMC sampler for an SDT model may be found in Appendix B.

An important aspect of the SDT example that has not come up before is that the model parameters are correlated. In other words, the relative likelihood of parameter values of d' will differ for different parameter values of C . While correlated model parameters are, in theory, no problem for MCMC, in practice they can cause great difficulty. Correlations between parameters can lead to extremely slow convergence of sampling chains, and sometimes to non-convergence (at least, in a practical amount of sampling time). There are more sophisticated sampling approaches that allow MCMC to deal efficiently with such correlations. A simple approach is *blocking*. Blocking allows the separation of sampling between certain sets of parameters. For example, imagine the detection experiment above included a difficulty manipulation where the quality of the visual stimulus is high in some conditions and low in others. There will almost surely be strong correlations between the two SDT parameters within different conditions: within each condition, high values of d' will tend to be sampled along with high values of C and vice versa for low values. Problems from these correlations can be reduced by blocking: that is, separating the propose-accept-reject step for the parameters from the two difficulty conditions (see e.g., Roberts & Sahu, 1997).

Sampling Beyond Basic Metropolis–Hastings

The Metropolis–Hastings algorithm is very simple, and powerful enough for many problems. However, when parameters are very strongly correlated, it can be beneficial to use a more complex approach to MCMC.

Gibbs Sampling

Given a multivariate distribution, like the SDT example above, Gibbs sampling (Smith & Roberts, 1993) breaks down the problem by drawing samples for each parameter directly from that parameter's *conditional distribution*, or the probability distribution of a parameter *given* a specific value of another parameter. An example of this type of MCMC is called Gibbs sampling, which is illustrated in the next paragraph using the SDT example from the previous section. More typically Gibbs sampling is combined with the Metropolis approach, and this combination is often referred to as "Metropolis within Gibbs". The key is that for a multivariate density, each parameter is treated separately: the propose/accept/reject steps are taken parameter by parameter. This algorithm shows how Metropolis within Gibbs might be employed for the SDT example:

1. Choose starting values for both d' and C , suppose these values are 1 and 0.5, respectively.
2. Generate a new proposal for d' , analogous to the second step in Metropolis–Hastings sampling described above. Suppose the proposal is 1.2.
3. Accept the new proposal if it is more plausible to have come out of the population distribution than the present value of d' , *given the present C value*. So, given the C

value of 0.5, accept the proposal of $d' = 1.2$ if that is a more likely value of d' than 1 for that specific C value. Accept the new value with a probability equal to the ratio of the likelihood of the new d' , 1.2, and the present d' , 1, given a C of 0.5. Suppose the new proposal (d' of 1.2) is accepted.

4. Generate a new proposal for C . For this a second proposal distribution is needed. This example will use a second proposal distribution that is normal with zero mean and standard deviation of 0.1. Suppose the new proposal for C is 0.6.
5. Accept the new proposal if it is more plausible to have come out of the population distribution than the C value, *given the present d' value*. So, given the d' value of 1.2, accept the proposal of $C = 0.6$ if that is a more likely value of C than 0.5 for that specific value of d' . Accept the new value with a probability equal to the ratio of the likelihood of the new C , 0.6, and the present C , 0.5, given a d' of 1.2. Suppose in this case that the proposal for C (0.6) is rejected. Then the sample for C stays at 0.5.
6. This completes one iteration of Metropolis within Gibbs sampling. Return to step 2 to begin the next iteration.

R-code for this example can be found in Appendix C. The results of running this sampler are shown in Figure 9.2. The left and middle columns show the d' and C variables respectively. Importantly, the right column shows samples out of the joint posterior, which is a bivariate distribution. It can be seen from this that the parameters are correlated. Such a correlation is typical with the parameters of cognitive models. This can cause a problem for Metropolis–Hastings sampling, because the correlated target distribution is very poorly matched by the proposal distribution, which does not include any correlation between parameters; sampling proposals from an uncorrelated joint distribution ignores the fact that the probability distribution of each parameter differs depending on the values of the other parameters. Metropolis within Gibbs sampling can alleviate this problem because it removes the need to consider multivariate proposals, and instead applies the accept/reject step to each parameter separately.

Differential Evolution

The previous section showed how Gibbs sampling is better able to capture correlated distributions of parameters by sampling from conditional distributions. This process, while accurate in the long run, can be slow. The reason is illustrated in the left panel of Figure 9.3.

Figure 9.3 shows a bivariate density very similar to the posterior distribution from the SDT example above. Suppose, during sampling, that the current MCMC sample is the value indicated by θ_t in Figure 9.3. The MCMC approaches discussed so far all use an uncorrelated proposal distribution, as represented by the circle around θ_t . This circle illustrates the fact that high and low values of the parameter on the x-axis are equally likely for any different value of the parameter on the y-axis. A problem arises because this uncorrelated proposal distribution does not match the correlated target distribution. In the target distribution, high values of the x-axis parameter tend to co-occur with high values of the y-axis parameter,

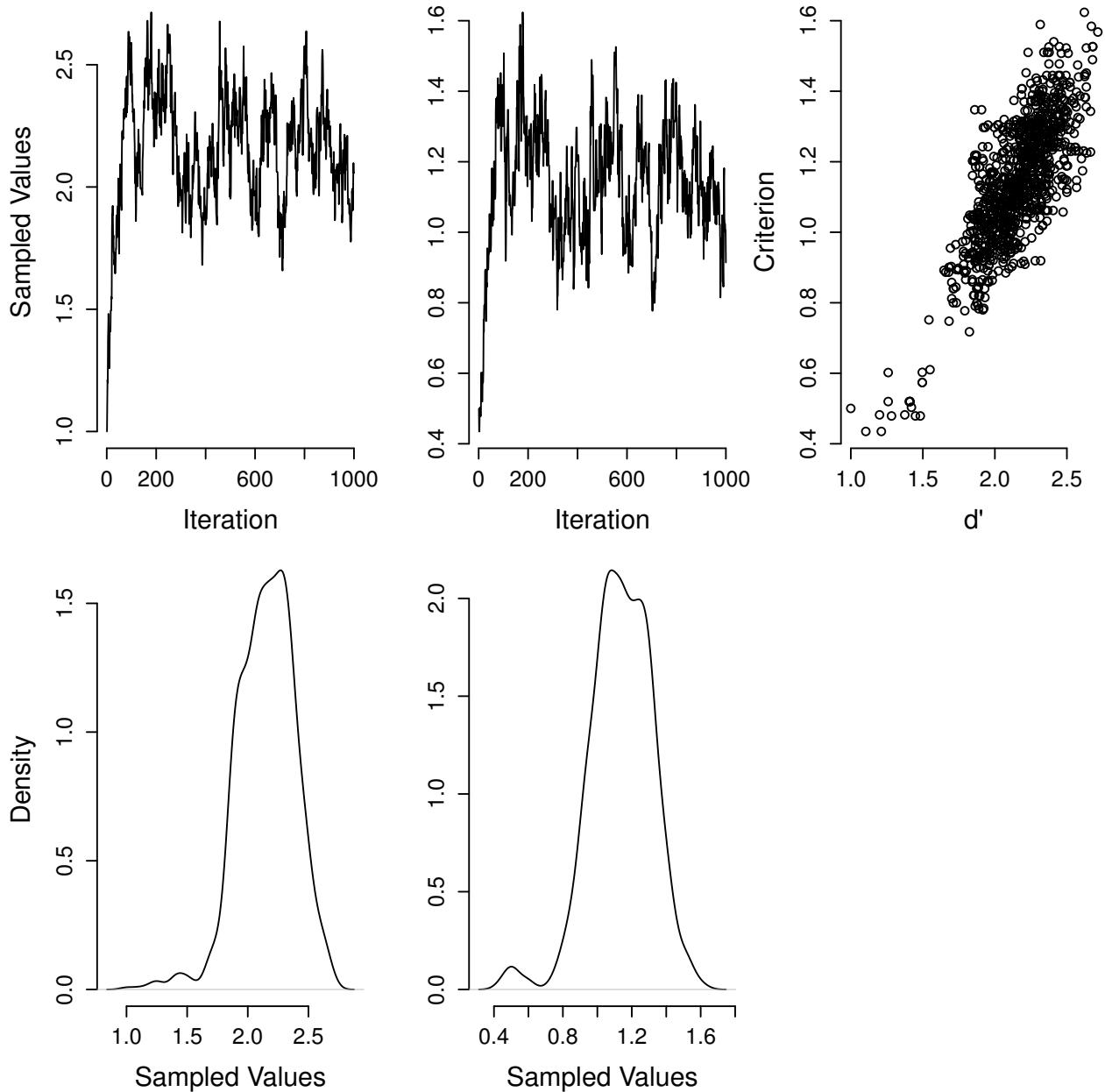


Figure 9.2: An example of Metropolis within Gibbs sampling. Left column: Markov chain and sample density of d' . Middle column: Markov chain and sample density of C . Right column: The joint samples, which are clearly correlated.

and vice versa. High values of the y-axis parameter almost never occur with low values of the x-axis parameter.

The mismatch between the target and proposal distributions means that almost half of all potential proposal values fall outside of the posterior distribution and are therefore sure to be rejected. This is illustrated by the white area in the circle, in which proposals have high values on the y-axis but low values on the x-axis. In higher dimensional problems (with

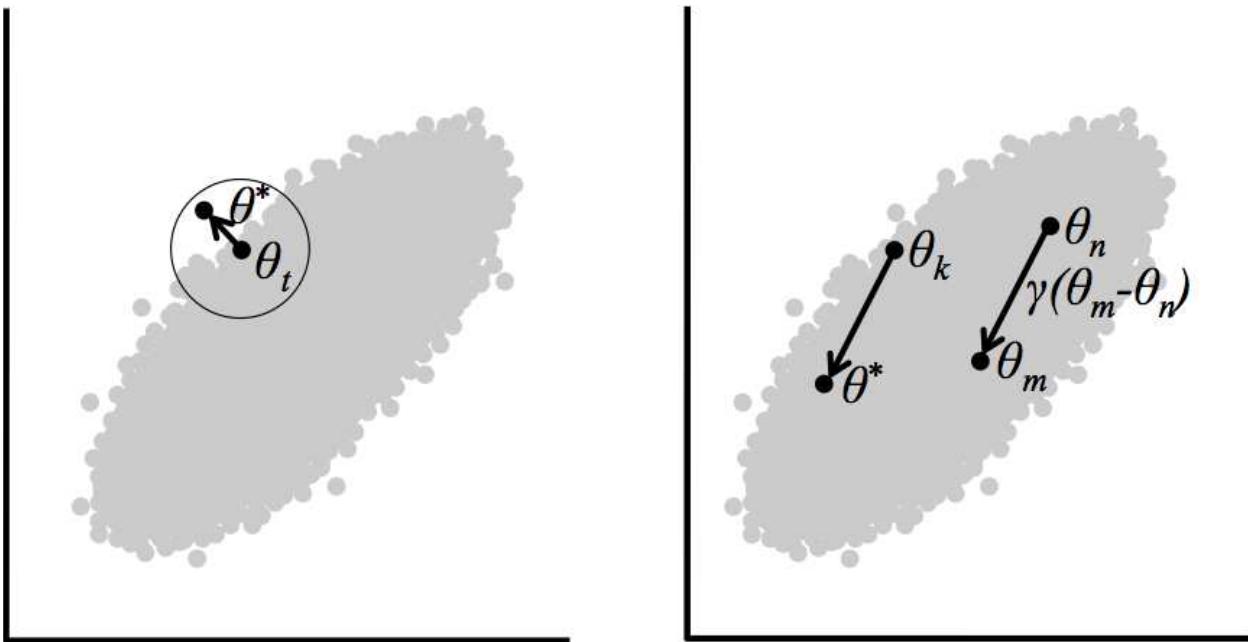


Figure 9.3: Left panel: MCMC sampling using a conventional symmetrical proposal distribution. Right panel: MCMC sampling using the crossover method in Differential Evolution. See text for details.

more parameters) this problem becomes much worse, with proposals almost certain to be rejected in all cases. This means that sampling can take a long time, and sometimes too long to wait for.

One approach to the problem is to improve proposals and have them respect the parameter correlation. There are many ways to do this, but a simple approach is called “differential evolution” or DE. This approach is one of many MCMC algorithms that use multiple chains: instead of starting with a single guess and generating a single chain of samples from that guess, DE starts with a set of many initial guesses, and generates one chain of samples from each initial guess. These multiple chains allow the proposals in one chain to be informed by the correlations between samples from the other chains, addressing the problem shown in Figure 9.3. A key element of the DE algorithm is that the chains are not independent – they interact with each other during sampling, and this helps address the problems caused by parameter correlations.

To illustrate the process of DE–MCMC, suppose there are multiple chains: $\theta_1, \theta_2, \dots$. The DE–MCMC algorithm works just like the simple Metropolis–Hastings algorithm from above, except that proposals are generated by information borrowed from the other chains (see the right panel of Figure 9.3):

1. To generate a proposal for the new value of chain θ_k , first choose two other chains at random. Suppose these are chains n and m . Find the distance between the current samples for those two chains, i.e.: $\theta_m - \theta_n$.

2. Multiply the distance between chains m and n by a value γ . Create the new proposal by adding this multiplied distance to the current sample. So, the proposal so far is: $\theta_k + \gamma(\theta_m - \theta_n)$. The value γ is a tuning parameter of the DE algorithm.
3. Add a very small amount of random noise to the resulting proposal, to avoid problems with identical samples (“degeneracy”). This leads to the new proposal value, θ^* .

Because DE uses the difference between other chains to generate new proposal values, it naturally takes into account parameter correlations in the joint distribution. To get an intuition of why this is so, consider the right panel of Figure 9.3. Due to the correlation in the distribution, samples from different chains will tend to be oriented along this axis. For example, very few pairs of samples will have one pair with a higher x-value but lower y-value than the other sample (i.e. the white area in the circle of the left panel of Figure 9.3). Generating proposal values by taking this into account therefore leads to fewer proposal values that are sampled from areas outside of the true underlying distribution, and therefore leads to lower rejection rates and greater efficiency. More information on MCMC using DE can be found in ter Braak (2006).

Like all MCMC methods, the DE algorithm has “tuning parameters” that need to be adjusted to make the algorithm sample efficiently. While the Metropolis-Hastings algorithm described earlier has separate tuning parameters for all model parameters (e.g. a proposal distribution width for the d' parameter, and another width for the C parameter), the DE algorithm has the advantage of needing just two tuning parameters in total: the γ parameter, and the size of the “very small amount of random noise”. These parameters have easily-chosen default values (see, e.g., Turner, Sederberg, Brown, & Steyvers, 2013). The default values work well for a very wide variety of problems, which makes the DE–MCMC approach almost “auto-tuning” (ter Braak, 2006). Typically, the random noise is sampled from a uniform distribution that is centered on zero and which is very narrow, in comparison to the size of the parameters. For example, for the SDT example, where the d' and C parameters are in the region of 0.5–1, the random noise might be sampled from a uniform distribution with minimum -0.001 and maximum +0.001. The γ parameter should be selected differently depending on the number of parameters in the model to be estimated, but a good guess is $2.38/\sqrt{2K}$, where K is the number of parameters in the model.

An example of cognitive models that deal with correlated parameters in practice is the class of response time modeling of decision making (e.g. Brown & Heathcote, 2008; Ratcliff, 1978; Usher & McClelland, 2001). As such, they are the kind of models that benefit from estimation of parameters via DE–MCMC. This particular type of MCMC is not trivial and as such a fully worked example of DE–MCMC for estimating response time model parameters is beyond the scope of this tutorial. The interested reader may find an application of DE–MCMC to estimating parameters for the Linear Ballistic Accumulator model of response times in Turner et al. (2013).

Summary

This tutorial provided an introduction to beginning researchers interested in MCMC sampling methods and their application, with specific references to Bayesian inference in cognitive science. Three MCMC sampling procedures were outlined: Metropolis(–Hastings), Gibbs, and Differential Evolution.² Each method differs in its complexity and the types of situations in which it is most appropriate. In addition, some tips to get the most out of your MCMC sampling routine (regardless of which kind ends up being used) were mentioned, such as using multiple chains, assessing burn-in, and using tuning parameters. Different scenarios were described in which MCMC sampling is an excellent tool for sampling from interesting distributions. The examples focussed on Bayesian inference, because MCMC is a powerful way to conduct inference on cognitive models, and to learn about the posterior distributions over their parameters. The goal of this paper was to demystify MCMC sampling and provide simple examples that encourage new users to adopt MCMC methods in their own research.

References

- Brown, S., & Heathcote, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, 57, 153–178.
- Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and behavior in decision-making. *PLoS Computational Biology*, 10, e1003700.
- Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Top Cogn Sci*, 1, 189–202.
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Elsevier Science.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1–15.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, 80, 205–235.

²For a visualization of Metropolis–Hastings and Gibbs sampling, see <http://twiecki.github.io/blog/2014/01/02/visualizing-mcmc/>.

- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18, 349–367.
- Roberts, G. O., & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B*, 59, 291–317.
- Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, 120, 39–64.
- Shiffrin, R. M., Lee, M. D., Kim, W. J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248–1284.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.
- Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 55, 3–23.
- ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16, 239–249.
- Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, 18, 368–384.
- Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, 108, 550–592.
- van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, 55, 94–105.
- van Ravenzwaaij, D., Moore, C. P., Lee, M. D., & Newell, B. R. (2014). A hierarchical Bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive Science*, 38, 1384–1405.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16, 44–62.
- Vickers, D., & Lee, M. D. (1997). Towards a dynamic connectionist model of memory. *Behavioral and Brain Sciences*, 20, 40–41.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 627–633.

Appendix A: Metropolis R-Code

Code for a Metropolis sampler, based on the in-class test example in the main text. In R, all text after the `#` symbol is a comment for the user and will be ignored when executing the code. The first two lines create a vector to hold the samples, and sets the first sample to 110. The loop repeats the process of generating a proposal value, and determining whether to accept the proposal value, or keep the present value.

```

samples      = numeric(500)                      # 500 samples.
samples[1]   = 110                                # The initial guess
for (i in 2:500) {
  p = samples[i-1] + rnorm (1, 0, 5)    # Proposal value
  if ((dnorm(p, 100, 15) / dnorm(samples[i-1], 100, 15)) > runif(1)) {
    samples[i] = p                            # Accept proposal
  } else {
    samples[i] = samples[i-1]                # Reject proposal
  }
}

```

Appendix B: SDT R-Code

Code for a Metropolis sampler for estimating the parameters of an SDT model. Given a specified number of trials with a target either present or absent, and given (fake) behavioral data of hits and false alarms, the code below evaluates the joint likelihood of SDT parameters, d' and C . New proposals for both parameters are sampled and evaluated simultaneously.

```

N.present = 100      # Number of trials on which a signal was present.
N.absent = 100       # Number of trials on which no signal was present.
N.hits = 85          # Correct responses to "present" trials.
N.falsealarms = 12  # Incorrect responses to "absent" trials.

posterior.density = function (parameters, h, fa, Np, Na) {
  # Function to calculate posterior density. "parameters"
  # is a 2-vector, with elements "d'" for d-prime and "C"
  # for criterion. "h" and "fa" are counts of hits and false
  # alarms. "Np" and "Na" are the number of trials with
  # target and no-target.

  # The model-predicted probability of a false alarm.
  prob.fa = pnorm (-parameters["C"])
  # The model-predicted probability of a hit.
  prob.h = pnorm (parameters["d'"] - parameters["C"])

```

```

# The log-likelihood of observing "fa" false alarms.
loglike.fa = dbinom (x = fa, size = Na, prob = prob.fa, log = TRUE)
# The log-likelihood of observing "h" hits.
loglike.h = dbinom (x = h, size = Np, prob = prob.h, log = TRUE)

# The prior log-likelihood of the parameters under a
# very simple prior of N(0,4) for both parameters.
loglike.prior = dnorm (parameters, mean = 0, sd = 4, log = TRUE)

# Return the posterior density: exp(sum).
exp(loglike.fa + loglike.h + sum (loglike.prior))
}

# Number of samples.
nmc = 500

# Create a vector to hold the samples.
samples = array(dim=c(2, nmc), dimnames=list(c("C", "d'"), NULL))

# Initial guess
samples[,1] = c(0.5, 1.0)

# Sample!
for (i in 2:nmc) {
  proposal = samples[,i-1] + rnorm(n=2, mean=0, sd=0.1)
  new.likelihood = posterior.density(parameters=proposal,
    h=N.hits, fa=N.falsealarms, Np=N.present, Na=N.absent)
  old.likelihood = posterior.density(parameters=samples[,i-1],
    h=N.hits, fa=N.falsealarms, Np=N.present, Na=N.absent)
  likelihood.ratio = new.likelihood / old.likelihood
  if (runif(1) < likelihood.ratio) {
    samples[,i] = proposal
  } else {
    samples[,i] = samples[, i - 1]
  }
}

```

Appendix C: Metropolis Within Gibbs Sampler R-Code

Code for a Metropolis within Gibbs sampler for estimating the parameters of an SDT model. The following code calculates the likelihood of the current d' and C parameter values (the

“posterior.density” function was omitted, but is identical to the one defined in Appendix B). The key difference between the Metropolis sampler in the previous section and the Metropolis within Gibbs sampler in this section is that the proposal and evaluation occurs separately for each parameter, instead of simultaneously for both parameters. The loop over the number of parameters, “for (j in rownames(samples))”, allows for parameter d' to have a new value proposed and its likelihood evaluated while parameter C is held at its last accepted value and vice versa.

```
# Number of samples.
nmc = 1000
# Number of parameters; d prime and criterion
n.pars = 2
# Create a vector to hold the samples
samples = array (dim = c(n.pars, nmc), dimnames = list( c("d'", "C"), NULL))

# Initial guess
samples[,1] = c(1, 0.5)

# Sample!
for (i in 2:nmc) {
  samples[,i] = samples[,i - 1]
  for (j in rownames(samples)) {
    proposal = samples[,i]
    proposal[j] = proposal[j] + rnorm (n=1 , mean=0, sd=0.1)
    new.likelihood = posterior.density (parameters=proposal, h=N.hits,
                                         fa=N.falsealarms, Np=N.present, Na=N.absent)
    old.likelihood = posterior.density (parameters=samples[,i], h=N.hits,
                                         fa=N.falsealarms, Np=N.present, Na=N.absent)
    likelihood.ratio = new.likelihood / old.likelihood
    if (runif(1) < likelihood.ratio) {
      samples[,i] = proposal
    }
  }
}
```

Glossary

Accepting A proposal value that is evaluated as more likely than the previously accepted value, or that is less likely but is accepted due to random chance. This value then becomes the value used in the next iteration.

Blocking Sampling only a subset of parameters at a time, while keeping the remaining parameters at their last accepted value.

Burn–In Early samples which are discarded, because the chain has not converged. Decisions about burn–in occur after the sampling routine is complete. Deciding on an appropriate burn–in is essential before performing any inference.

Chain One sequence of sampled values.

Conditional Distribution The probability distribution of a certain parameter given a specific value of another parameter. Conditional distributions are relevant when parameters are correlated, because the value of one parameter influences the probability distribution of the other.

Convergence The property of a chain of samples in which the distribution does not depend on the position within the chain. Informally, this can be seen in later parts of a sampling chain, when the samples are meandering around a stationary point (i.e., they are no longer coherently drifting in an upward or downward direction, but have moved to an equilibrium). Only after convergence is the sampler guaranteed to be sampling from the target distribution.

Differential Evolution A method for generating proposals in MCMC sampling. See section “Differential Evolution” for a more elaborate description.

Gibbs Sampling A parameter-by-parameter approach to MCMC sampling. See section “Gibbs Sampling” for a more elaborate description and an example.

Iteration One cycle or step of MCMC sampling, regardless of routine.

Local maxima Parameter values that have higher likelihood than their close neighbors, but lower likelihood than neighbors that are further away. This can cause the sampler to get “stuck”, and result in a poorly estimated target distribution.

Markov chain Name for a sequential process in which the current state depends in a certain way only on its direct predecessor.

MCMC Combining the properties of Markov chains and Monte–Carlo. See their respective entries.

Metropolis algorithm A kind of MCMC sampling. See section “In–Class Test” for a more elaborate description and an example.

Monte–Carlo The principle of estimating properties of a distribution by examining random samples from the distribution.

Posterior Used in Bayesian inference to quantify a researcher’s updated state of belief about some hypotheses (such as parameter values) after observing data.

Prior Used in Bayesian inference to quantify a researcher’s state of belief about some hypotheses (such as parameter values) before having observed any data. Typically represented as a probability distribution over different states of belief.

Proposal A proposed value of the parameter you are sampling. Can be accepted (used in the next iteration) or rejected (the old sample will be retained).

Proposal Distribution A distribution for randomly generating new candidate samples, to be accepted or rejected.

Rejecting A proposal might be discarded if it is evaluated as less likely than the present sample. The present sample will be used on subsequent iterations until a more likely value is sampled.

Rejection Rate The proportion of times proposals are discarded over the course of the sampling process.

Starting Value The initial “guess” for the value of the parameter(s) of interest. This is the starting point for the MCMC sampling routine.

Target Distribution The distribution one samples from in an attempt to estimate its properties. Very often this is a posterior distribution in Bayesian inference.

Tuning Parameter Parameters which influence the behavior of the MCMC sampler, but are not parameters of the model. For example, the standard deviation of a proposal distribution. Use caution when choosing this parameter as it can substantially impact the performance of the sampler by changing the rejection rate.

10

Bayesian latent variable models for the analysis of experimental psychology data

Edgar C. Merkle and Ting Wang

Traditional applications of factor analysis and related latent variable models include psychometric scale development, analysis of observational data, and possibly data reduction (though the related, but distinct, principal components analysis is more relevant here). Because experimental psychologists do not typically embark on such applications, they may regard latent variable models as being irrelevant to their statistical toolboxes. Furthermore, applications of latent variable models to experimental data are relatively uncommon in the literature (though see Bagozzi & Yi, 1989; Donaldson, 2003; Miyake, Friedman, Emerson, Witzki, & Howerter, 2000; Russell, Kahn, Spoth, & Altmaier, 1998; Wicherts, Dolan, & Hessen, 2005). In this paper, we illustrate that Bayesian latent variable models can be advantageous for the analysis of multivariate data collected in many laboratory experiments. The models flexibly allow us to handle data from multiple trials, conditions, or scenarios, without the need for aggregation across trials or subjects. This allows us to pool data across multiple trials in order to draw general conclusions about the research question of interest, while preserving variability inherent in the raw data. We focus on Bayesian versions of the models because they are flexible and advantageous for model extension, comparison, and interpretation.

Bayesian approaches to estimating latent variable models have been considered for at least forty years, with increased attention to the subject following the computational advances in Markov chain Monte Carlo (MCMC). Early treatments include Press (1972), Martin and McDonald (1975), Bartholomew (1981), and Lee (1981), while later treatments that incorporate MCMC include Scheines, Hoijtink, and Boomsma (1999), Lee (2007), Song and Lee (2012a), and B. Muthén and Asparouhov (2012). Some authors of the recent papers refer to a need for fully automated software that fits latent variable models via MCMC, with **Mplus** (L. K. Muthén & Muthén, 1998–2012) providing SEM functionality and **BUGS** (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012), **JAGS** (Plummer, 2003), and **Stan** (Stan Development Team, 2014) providing general functionality. This automation has the reward of easily estimating complex models coupled with the risk of estimating inappropriate

models or drawing inappropriate conclusions (also see MacCallum, Edwards, & Cai, 2012; Steiger, 2001; Stromeyer, Miller, Sriramachandramurthy, & DeMartino, 2014).

In the pages below, we first present an overview of factor analysis and structural equation models. We also discuss necessary topics for Bayesian model estimation, including prior distribution specification and model comparison methods. The Bayesian methods are trickier than they are for other models because of the inherent parameter identification issues. We then apply the models to data from an experiment on risky choice, highlighting the unified and unique results that the models can provide. We estimate all models in JAGS and carry out related computations in R; code to replicate all results is provided at the URL referenced at the end of the paper.

Models

To describe the models, we first set up a framework that is relevant to experimental research. We assume that n individuals respond to a series of stimuli that are all intended to measure the same concept. Across the stimuli, each individual contributes p observed variables. For example, in the illustration later, the stimuli are five binary choices. For each choice stimulus, individuals rate (i) the extent to which each individual choice option is attractive and (ii) the extent to which they prefer one stimulus over the other. Here, the stimuli are all intended to measure an individual's attraction to risk, and each individual provides $p = 15$ observed variables: three ratings for each of 5 stimuli. We generally assume that y_{ij} ($i = 1, \dots, n$; $j = 1, \dots, p$) is individual i 's response to observed variable j , and we also sometimes assume that individuals are randomly assigned to a between-subjects condition. This mimics many psychology experiments, where researchers are interested in studying whether experimental conditions impacted individuals' responses across multiple trials.

In the discussion below, we assume that y_{ij} is approximately continuous (so that our models fully rely on normal distributions). However, variants of these models can also be applied to discrete data such as binary choices or ordinal ratings. These variants are highly related to the item response models that are described in Voorspoels, Rutten, Bartlema, Tuerlinckx, and Vanpaemel (in press).

Factor Model

The traditional factor model (e.g., Lawley & Maxwell, 1971; Bartholomew, Knott, & Moustaki, 2011) assumes that each individual has a stable underlying trait (e.g., attraction to risky choice) across all experimental stimuli. An individual's response to each stimulus is driven by this trait. The stimuli are unique, however, so that stimulus characteristics also impact the individual's responses. Further, like many other models, noise impacts the responses.

The above ideas can be formally represented by:

$$y_{ij} | \theta_i, \gamma_j, \lambda_j, \psi_j \sim N(\mu_{ij}, \psi_j) \quad \text{for all } i, j \quad (10.1)$$

$$\mu_{ij} = \gamma_j + \lambda_j \theta_i \quad (10.2)$$

$$\theta_i \sim N(0, \phi). \quad (10.3)$$

Within this model, θ_i is the value of individual i 's stable trait. This value arises from a normal distribution with mean 0 and variance ϕ , reflecting the distribution of the trait across the population of interest. The observed response y_{ij} is perturbed from the trait by the sources mentioned in the previous paragraph: stimulus characteristics γ_j and λ_j , and random noise with variance ψ_j .

The stimulus characteristics include the intercept γ_j , reflecting stimulus j 's mean response in the population of interest, and slope λ_j , reflecting the extent to which stimulus j "taps into" the trait of interest. If λ_j equals 0, individual i 's standing on the trait has no impact on his/her response y_{ij} . Conversely, if λ_j is large (in the positive or negative direction), individual i 's standing on the trait has a large impact on his/her response. In traditional applications, the λ_j are called *factor loadings* and the trait of interest is called the *factor*.

Multiple factors

The above model assumed that a single latent trait was associated with the observed responses y_{ij} . In some studies, multiple traits will have an association. For example, undergraduate participants' responses to a recognition memory task may be influenced both by their working memory capacity and by their attentional functioning during the experiment. If a participant has both high working memory and high attention, he/she should exhibit good test performance. Conversely, low values of either trait will result in lower performance.

Assuming that m traits impact observed responses, we can modify the previous model to arrive at an " m -factor model:"

$$y_{ij} | \boldsymbol{\theta}_i, \gamma_j, \Lambda_j, \psi_j \sim N(\mu_{ij}, \psi_j) \quad \text{for all } i, j \quad (10.4)$$

$$\mu_{ij} = \gamma_j + \sum_{k=1}^m \lambda_{jk} \theta_{ik} \quad (10.5)$$

$$\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \boldsymbol{\Phi}), \quad (10.6)$$

where $\boldsymbol{\theta}_i$ is a vector describing individual i 's standing on each of the m latent traits (with this vector now arising from a multivariate normal distribution) and Λ_j contains the m factor loadings for stimulus j . The matrix $\boldsymbol{\Phi}$ contains information about the variance of each trait along with covariances between pairs of traits. For the next section, it is helpful to combine the Λ_j into a $p \times m$ matrix Λ .

Exploratory versus confirmatory models

From inspection of the above models, it is evident that there are some unidentified parameters. For example, Equation (10.2) involves the multiplication of two unknown parameters $\lambda_j \times \theta_i$. If we multiply each λ by a constant and divide each θ by the same constant, our model predictions remain unchanged. This lack of identification is compounded by the fact that the $\boldsymbol{\Phi}$ matrix typically contains additional free parameters. To address problems such as this, we need to fix some model parameters to constants (typically to zero or one). Additionally, when we consider the m -factor model from Equations (10.4) to (10.6), it is possible to use matrix algebra to illustrate a *rotational indeterminacy* problem. This implies that

specific linear transformations (more complex than multiplication by a constant) of the λ parameters, coupled with inverse transformations of the θ parameters, lead to unchanged model predictions.

The m -factor model requires us to systematically constrain m^2 parameters in order to achieve identification (this does not count the mean of $\boldsymbol{\theta}_i$, which is fixed to $\mathbf{0}$). If we constrain only the minimum m^2 parameters, we obtain an *exploratory* factor model. This model allows for the possibility that all latent traits (contained in $\boldsymbol{\theta}$) are associated with all observed measures (the y_{ij}). We can obtain unique parameter estimates based on the specific m^2 parameter constraints employed, but there are an infinite number of other, equally-good sets of parameter values coupled with alternative parameter constraints (these arise from the rotational indeterminacy issue described in the previous paragraph). Because of this, the λ (and possibly Φ) parameters are typically transformed following model specification so that they are easily interpretable. This “interpretability” step is called *rotation*, and we provide some further information on it in the General Discussion (also see Browne, 2001). Regardless of the rotation step, however, the exploratory model can be useful because it allows us to compare models with different values of m . This can provide information about the number of latent traits associated with the observed measures, as well as about which latent traits are associated with which measures.

If we fix more than m^2 parameters, we obtain a *confirmatory* factor model. In this model, we assume that some latent traits have no association with some observed measures (consequently, some of the parameters in Λ are fixed to zero) and also assume fixed, nonzero values for some parameters. Rotational indeterminacy is no longer a problem here, though alternative parameter constraints can still impact parameter estimates. These are further described below.

Identification constraints for factor models

Regardless of whether we have an exploratory or confirmatory model, the first m^2 parameter constraints must be set in a systematic fashion (see Jöreskog, 1969, 1979; Peeters, 2012a, 2012b) to identify the likelihood. Popular strategies are described below; all strategies assume that the resulting Λ matrix is of full rank (e.g., that we cannot obtain one column of Λ through a linear combination of other columns) and that the resulting Φ matrix is positive definite.

Often, we start by setting $\Phi = \mathbf{I}$, which means that each latent trait has a variance of 1 with each pair of traits having a 0 correlation. If we employ this constraint, we then systematically fix $m(m - 1)/2$ parameters in Λ : one column can have no zeros, one column must have one zero, one column must have two zeros, . . . , and one column must have $(m - 1)$ zeros.

Alternatively to above, we can fix the diagonal of Φ (reflecting latent trait variances) to equal 1 but allow other parameters in Φ to be unconstrained. This allows for the possibility that latent traits are correlated. If we opt for these constraints, we must fix more parameters in Λ to zero. In particular, each column of Λ must now contain $(m - 1)$ zeros, for a total of $m(m - 1)$ zeroes in Λ . These restrictions, proposed by Jöreskog (1979), achieve “local” identification: they identify the model up to sign changes in columns of Λ . Peeters

(2012b) describes how to make these constraints globally identified: in addition to the above constraints, one must force a nonzero parameter in each column of Λ to assume only positive or negative values.

Finally, another strategy involves placing all the restrictions in Λ . Now Φ is completely unconstrained, with each column of Λ containing $(m - 1)$ zeros. Each column of Λ must also contain a nonzero constant, typically one (with each one appearing in different rows), compensating for the unconstrained Φ . These restrictions, also proposed by Jöreskog (1979), achieve global identification.

The above constraints might make some readers uncomfortable, leading them to wonder whether estimated parameters can be interpreted given all the identification issues involved. We suggest first comparing exploratory models via Bayes factors (or other global model statistics), which should be approximately invariant to the specific identification constraints that are chosen. The Bayes factors will provide information about the best value of m or about which latent traits impact which observed measures. Interpretation of a single model may then proceed by studying the pattern of λ estimates across latent traits and observed variables; see Hoyle and Duvall (2004); Peeters, Dziura, and van Wesel (2014), and Preacher and MacCallum (2003) for further discussion of related strategies. In addition to these strategies, the models can be extended so that we obtain estimates of experiment-specific parameters. This motivates the use of structural equation models, described later.

Exploratory models and priors

In ML contexts, one often fits the model to the observed covariances between the p variables and integrates out the latent variables θ . In Bayesian contexts, one typically fits the model to the raw data, and θ is usually sampled simultaneously with other model parameters via MCMC. This sampling of the latent variables introduces some identification issues that are specific to Bayesian factor models. If these issues are not addressed, the sampled parameters will never converge to a stationary distribution.

To address the identification issues, we adopt a parameter expansion approach (Gelman, 2004, 2006; Ghosh & Dunson, 2009) to fit exploratory factor models. Parameter expansion involves sampling from a “working model” with unidentified parameters. At each iteration of the sampling, we transform the sampled parameters to correspond to the “inferential model” with uniquely identified parameters. This improves the speed at which the sampled chains converge to the posterior distribution, while also handling some of the parameter identification issues inherent to the model.

The working model is intended to help us sample the factor loadings λ and the latent trait values θ . We allow the entire covariance matrix Φ^* to be unconstrained, and we also fix $m(m - 1)$ parameters in Λ^* to zero (see p. 282). We then place the following prior distributions on the rest of the (unidentified) parameters:

$$\lambda_{jk}^* \sim N(0, 1) \text{ for } \lambda_{jk}^* \text{ free} \quad (10.7)$$

$$\theta_i^* \sim N_m(\mathbf{0}, \Phi) \quad i = 1, \dots, n \quad (10.8)$$

$$\Phi^{*-1} \sim \text{Wishart}(\mathbf{V}, m), \quad (10.9)$$

where \mathbf{V} is an $m \times m$ covariance matrix, and the asterisks on the θ s, λ s, and ϕ s imply that these parameters are unidentified. At each iteration, these unidentified working parameters are transformed to identified versions in the inferential model via

$$\lambda_{jk} = \text{sign}(\lambda_{kk}^*) \lambda_{jk}^* \phi_{kk}^{*1/2} \quad (10.10)$$

$$\theta_{ik} = \text{sign}(\lambda_{kk}^*) \phi_{kk}^{*-1/2} \theta_{ik}^* \quad (10.11)$$

$$\phi_{k\ell} = \frac{\text{sign}(\lambda_{kk}^*) \text{sign}(\lambda_{\ell\ell}^*) \phi_{k\ell}^*}{\sqrt{\phi_{kk}^* \phi_{\ell\ell}^*}}, \quad k, \ell = 1, \dots, m, \quad (10.12)$$

where $\text{sign}(\lambda_{kk}^*)$ equals either 1 or -1 , depending on whether λ_{kk}^* is positive or negative. These transformations automatically yield the remaining constraints necessary to globally identify the model parameters: in addition to the $m(m-1)$ parameters of $\boldsymbol{\Lambda}$ fixed to zero, we now have that parameters along the diagonal of $\boldsymbol{\Phi}$ equal 1. Further, we have restricted one parameter in each column of $\boldsymbol{\Lambda}$ to be positive, which is one way to fulfill the sign restriction described by Peeters (2012b). This procedure results in chains of sampled parameters that quickly converge to their posterior distributions for small values of m (say, 5 or less).

For large values of m , the specific λ parameters that are fixed to zero can impact convergence. If these constraints cause the posterior associated with one of the λ_{kk} to overlap with zero, then the chains cannot converge (because, under the parameter expansion approach, λ_{kk} can only be positive). This is only problematic if many of the observed variables are not associated with the latent traits of interest.

Structural Equation Models

The factor model can be extended in a variety of ways, many of which are relevant to experimental psychologists. For example, suppose that each participant i is assigned to one of two experimental conditions. We can allow unique parameters for each condition, which can provide information about the extent to which the parameters differ across conditions (e.g., Millsap, 2011; Verhagen & Fox, 2013). These types of models are often called *multiple-group models*.

We can also allow specific latent traits to influence other latent traits. For example, in situations where subjects choose between risky and riskless options, subjects' latent attractions to each type of option should influence their choice. Models that include directional influences of latent traits on other latent traits are no longer factor models; they are instead *structural equation models*. Structural equation models subsume factor models and can be written as

$$y_{ij} | \boldsymbol{\theta}_i, \gamma_j, \boldsymbol{\Lambda}_j, \psi_j \sim N(\mu_{ij}, \psi_j) \quad \text{for all } i, j \quad (10.13)$$

$$\mu_{ij} = \gamma_j + \sum_{k=1}^m \lambda_{jk} \theta_{ik} \quad (10.14)$$

$$\theta_{ik} = \alpha_k + \sum_{\ell=1}^m B_{k\ell} \theta_{i\ell} + \zeta_{ik} \quad \text{for all } i, k, \quad (10.15)$$

where α_k is the mean of latent trait k (often fixed to zero) and ζ_{ik} is a normally-distributed, zero-centered residual term associated with latent trait k . Equation (10.15) may look confusing to some because the θ parameters appear on both sides of the equation. This is the part of the equation that allows some latent traits to have directional influences on others. We require $B_{kk} = 0$, $k = 1, \dots, m$, which simply means that no latent trait exerts a directional influence on itself. Many of the other B parameters are also fixed to zero; the only nonzero B parameters correspond to latent variables that have a directional influence on another latent variable.

Finally, the structural equation model presented above can be further extended to allow for interactions between latent variables (e.g., Klein & Moosbrugger, 2000). To do so, we maintain most of the previous model, replacing only Equation (10.15) with

$$\theta_{ik} = \alpha_k + \sum_{\ell=1}^m B_{k\ell} \theta_{i\ell} + \sum_{q_1=1}^m \sum_{q_2=q_1}^m \xi_{kq_1 q_2} \theta_{iq_1} \theta_{iq_2} + \zeta_{ik} \quad \text{for all } i, k, \quad (10.16)$$

where ζ_{ik} is still a zero-centered residual and many of the ξ parameters are fixed to zero (as are many of the B parameters). This model is difficult to fit via ML methods and simpler to fit via Bayesian methods (Lee, Song, & Tang, 2007). Bayesian methods allow us to sample the latent variables $\boldsymbol{\theta}_i$, which in turn allow us to deal with the conditional distribution of \mathbf{y}_i given $\boldsymbol{\theta}_i$. These conditional distributions are normal, regardless of whether or not we employ latent variable interactions. In contrast, ML methods work on the marginal distribution of \mathbf{y}_i , integrating out the latent variable $\boldsymbol{\theta}_i$. When there are latent variable interactions in the model, this marginal distribution is no longer normal. Consequently, special ML estimation methods must be developed specifically for models that include latent variable interactions. For more discussion of these issues and specific ML model estimation methods, see Cudeck, Harring, and du Toit (2009), Klein and Moosbrugger (2000), Klein and Muthén (2007), and Marsh, Wen, Hau, and Nagengast (2013).

Identification constraints for SEMs

Structural equation models' likelihoods are identified when both the measurement part (Equation (10.14)) and the structural part (Equation (10.15)) are identified (Lee, 2007). Identification of the measurement part is typically handled in a manner similar to the factor analysis models. There exist no general rules for identifying the structural part, however (Bollen & Davis, 2009). Researchers approach this problem by satisfying some (not unique) sufficient conditions that are relatively easy to satisfy in practice (Lee, 2007). These conditions often focus on *nonrecursive* relationships between latent variables, which include reciprocal directional effects, feedback loops, and correlated residuals. The idea is that, to achieve identification, we should minimize the number of directional influences and covariances between latent variables (i.e., we want to avoid cases where most of the B parameters are estimated and where the covariance matrix associated with ζ is nearly unrestricted). See Rigdon (1995) for further discussion of these issues.

SEMs and priors

To specify the priors for structural equation models, we adopt an approach related to that described by Merkle (2011), Curtis (2010), and others. We fix the latent variable variances ϕ to 1 and then, for each factor, restrict at least one loading λ to have a prior distribution that is truncated from below at 0. This truncation solves the “sign change” issue that was described by Peeters (2012b). We could also use a parameter expansion approach here, but it adds extra complexity to the already-complex JAGS code. Further details on this issue, along with specific prior distributions used, appears in the Application section. For the JAGS code used to fit the models, see the Computational Details section at the end of the paper.

Model Comparison: Bayes Factor Computation

In a latent variable context, Bayes factors are useful for comparing models with differing numbers of latent traits (i.e., models with different values of m) or for comparing structural equation models with different covariates. We compute log-Bayes factors to compare pairs of models in this paper, with positive values generally conveying support for more complex models (which contain extra effects or factors). Kass and Raftery (1995) provide some rules of thumb for their interpretation, suggesting that log-Bayes factors below 1 are “not worth more than a bare mention,” values from 1 to 3 are “positive,” values from 3 to 5 are “strong,” and values greater than 5 are “very strong.” Like most rules of thumb, these values are subject to debate.

To compare models with different values of m , factor analysts have traditionally used the method of path sampling for Bayes factor computation (Lee & Song, 2002; Ghosh & Dunson, 2009). This method was originally described by Gelman and Meng (1998), but recent work suggests that it can be problematic when it is applied to factor analysis (Dutta & Ghosh, 2013). Hence, we rely on the Laplace approximation to the Bayes factor (Kass & Raftery, 1995; Lewis & Raftery, 1997), and we also use the Savage-Dickey method (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) to check our results. We now briefly describe the Laplace approximation and the Savage-Dickey density ratio. Alternative methods for computing Bayes factors are described by, e.g., Kass and Raftery (1995), Lopes and West (2004), Rouder and Morey (2012), and Rouder, Morey, Speckman, and Province (2012).

Laplace Approximation

The Laplace approximation to the Bayes factor (which we use throughout the paper) is used to approximate the integrals involved in the candidate models’ marginal likelihoods (the ratio of which yields the Bayes factor). Given a model, the Laplace approximation to the marginal log-likelihood is (Lewis & Raftery, 1997)

$$\log(f(\mathbf{Y})) \approx \frac{q}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{J}^{-1*}|) + \log(f(\boldsymbol{\Omega}^*)) + \log(f(\mathbf{Y}|\boldsymbol{\Omega}^*)), \quad (10.17)$$

where Ω is the model parameter vector of length q , Ω^* is a posterior central tendency measure of Ω , J^{-1*} is the inverse of the information matrix evaluated at Ω^* , and $f()$ is the probability distribution of the terms within its parentheses. This approximation has been found to be accurate when the posterior distribution is unimodal without long tails (Lewis & Raftery, 1997; Tierney & Kadane, 1986), and its error in approximating the Bayes factor is of order $O(n^{-1})$ (Kass & Raftery, 1995). Raftery (1993) has previously considered the approximation's application to the types of models described here.

The beauty of the Laplace approximation involves the fact that all four terms on the right of (10.17) can often be obtained directly from MCMC output. The first term is trivial. The second term can be estimated via the covariance matrix of sampled parameters. The third term requires Ω^* , which is a posterior central tendency measure of each sampled parameter (and we are evaluating the prior distributions at Ω^*). Finally, the fourth term requires us to evaluate the likelihood at Ω^* .

The fourth term can be somewhat difficult to evaluate when there are random effects in the model that are not officially counted as model parameters; this includes the θ parameters in the factor models described here. As Lewis and Raftery (1997) describe, we must integrate out the random effects in order to calculate the fourth term. This is easy to do for factor models involving normal distributions because we can solve the integral analytically. The solution is:

$$\mathbf{y}_i | \gamma, \Lambda, \Phi, \Psi \sim N_p(\gamma, \Lambda \Phi \Lambda' + \Psi), \quad (10.18)$$

where \mathbf{y}_i is individual i 's data vector, γ is a vector of intercepts, Ψ is a diagonal matrix of residual variances, and Λ (the collection of factor loadings) and Φ (the latent trait covariance matrix) were defined previously. For SEMs involving normal distributions the solution is

$$\mathbf{y}_i | \gamma, \Lambda, \Phi, \Psi, \alpha, B, \Delta \sim N_p(\gamma + \Lambda \alpha, \Lambda B \Phi B' \Lambda' + \Lambda \Delta \Lambda' + \Psi), \quad (10.19)$$

where $\mathbf{y}_i, \gamma, \Lambda$ and Ψ are defined in the same way as Equation (10.18); α is a vector of latent trait means; B is a matrix containing the B parameters from Equation (10.15); and Δ is a matrix containing the variances of the residual terms ζ_{ik} from Equation (10.15). The matrix Φ is still the latent trait covariance matrix, but the vital difference between SEM and factor analysis involves the entries of Φ . In factor analysis, each entry of Φ is either a single parameter or a constraint (0 or 1). In SEMs, we need to derive many entries of Φ via conditional distributions. This is one reason why the Bayesian SEM framework can be extended more flexibly. For example, for latent variable interactions, we can separate observed variables that "receive" a latent interaction from those that do not, then rely on the conditional distribution of the former observed variables (conditioned on the other latent variables) to arrive at the full marginal distribution. If our models involve distributions other than the normal (as is the case for, say, logistic regressions and many item response models), then integration of the random effects is less straightforward. Further MCMC steps could be used here; see Lewis and Raftery (1997) for more detail.

Savage-Dickey Density Ratio

The Savage-Dickey density ratio is useful for computing Bayes factors associated with nested models; such models often mimic null hypotheses. In such situations, we often wish to test

a hypothesis that a model parameter (or some function of model parameters) equals zero. We can gain information related to this hypothesis by comparing two models: Model A, where the parameter in question is freely estimated, and Model B, where the parameter in question is constrained to zero. In computing a Bayes factor to compare these models, we gain information related to the chance that the model parameter actually equals zero.

The Savage-Dickey density ratio allows us to compute these types of Bayes factors by estimating only Model A (where the focal parameter is free). Assume that we wish to calculate the Bayes factor associated with $\omega = 0$ versus $\omega \neq 0$, where ω is a specific entry of the parameter vector Ω . Then we have that

$$BF_{BA} = \frac{p(\omega = 0 | \mathbf{Y})}{p(\omega = 0)}; \quad (10.20)$$

that is, the Bayes factor is the ratio between the posterior density evaluated at $\omega = 0$ and the prior density evaluated at $\omega = 0$. The prior density is known, and the posterior density becomes normal for large n (e.g., Gelman, Carlin, Stern, & Rubin, 2004). For smaller n , the posterior density can be approximated via splines or kernel-based methods; see Wagenmakers et al. (2010) for more detail. Regardless, we can evaluate the posterior density using the MCMC output and a minimal amount of extra computation.

Throughout the application below, we report Bayes factors calculated via the Laplace approximation. We also used the Savage-Dickey method to ensure that both methods agreed; while the Savage-Dickey methods are not reported here, the code used to calculate all Bayes factors is included in the replication code. We now move to the application, which can be viewed as the start of a serious data analysis.

Application: Risky Choice

Peters and Levin (2008) studied framing effects of risky and riskless choice alternatives in variants of the Asian Disease Problem (Tversky & Kahneman, 1981). This problem requires subjects to choose between two treatments for a fictional disease that impacts 600 people. In a “positive frame” version of the problem, a riskless treatment saves 200 lives, while a risky treatment saves all 600 lives with probability 1/3 and no lives with probability 2/3. In a “negative frame” version of the problem, a riskless treatment causes 400 people to die, while a risky treatment causes 0 deaths with probability 1/3 and 600 deaths with probability 2/3. The question framing (number dead versus number alive) is known to have an impact on choice: “death” phrasing causes people to choose the risky option, while “alive” phrasing causes people to choose the riskless option.

Peters and Levin (2008) were interested in the extent to which each option’s attractiveness influenced choice, along with the impact of numeracy on the relationship between attractiveness and choice. They created five variants of the Asian Disease Problem involving human deaths due to disease, animal deaths due to wildfire, and crop loss due to drought. Subjects were assigned to one of two “frame” conditions, where the problems were phrased as described above. Subjects reported a preference between options on a seven-point scale (from “much prefer the riskless option” to “much prefer the risky option”), and they also

rated the attractiveness of each option. Thus, for each problem, subjects gave three ratings on seven-point scales: attraction to the risky option, attraction to the riskless option, and preference. For each subject, the data therefore include 10 attractiveness ratings (of the risky option and the riskless option for each of the five problems) and 5 preference ratings (for each of the five problems). These ratings are all treated as continuous below, though similar models could be specified if, e.g., choice were binary.

The original authors' focal hypothesis was that subjects' numeracy would moderate the relationship between attractiveness ratings and choice: as numeracy increases, attractiveness ratings should be more predictive of choice. To study this hypothesis, the authors employed separate ANOVAs on each of the five problem variants and on data that were averaged over the problems. They generally found statistically significant interaction effects in agreement with their focal hypothesis. In the analysis described here, we employ factor models and structural equation models to simultaneously study the experimental data across all five problems and to handle potential differences between problems.

Methods

We propose the use of factor analysis and SEM to study the authors' original hypotheses. This allows us to draw unified inferences about the full experimental data from a single model, removing the need to do separate analyses of each problem. While the modeling here can be viewed as the start of a full analysis, we omit some steps and issues due to space. We further describe these omissions in the Discussion section.

Our modeling proceeds in two steps. First, we fit 1- and 2-factor models (i.e., Equations (10.4) to (10.6), with $m = 1$ and $m = 2$) to the ten attractiveness ratings given by each subject (one risky rating and one riskless rating, for each of the five questions). The factors (latent variables) represent participants' attractions to risky and riskless alternatives in the context of this experimental study; they are used here to pool information across all five problems. In comparing a one-factor model to a two-factor model, we obtain information about the nature of attraction to risky and riskless alternatives: are these two extremes of a continuum, or are they unique dimensions? The former implies that a person who is attracted to risky alternatives is also repelled by riskless alternatives, while the latter implies that a person can be simultaneously attracted to (or repelled by) both risky and riskless alternatives. This comparison provides information about the best way to model the effect of attractiveness ratings on choice. To preview our results, we find that the 2-factor model is better than the 1-factor model.

Next, we build off the results of the first step to examine the effects of risky/riskless attraction on choice, as well as to draw inferences related to the experimental manipulations and to numeracy. The model is most easily conceptualized via the path diagram in Figure 10.1. This diagram illustrates a series of regression-like relationships (directed arrows) between observed variables (squares) and latent traits (circles). The observed variables AR1 to AR5 represent riskless attraction ratings for items 1–5; the observed variables AR6 to AR10 represent risky attraction ratings for items 6–10; and the observed variables C1 to C5 represent choice ratings for items 1–5. Double-headed arrows imply variance parameters.

The base portion of the model involves three main latent variables: a riskless variable

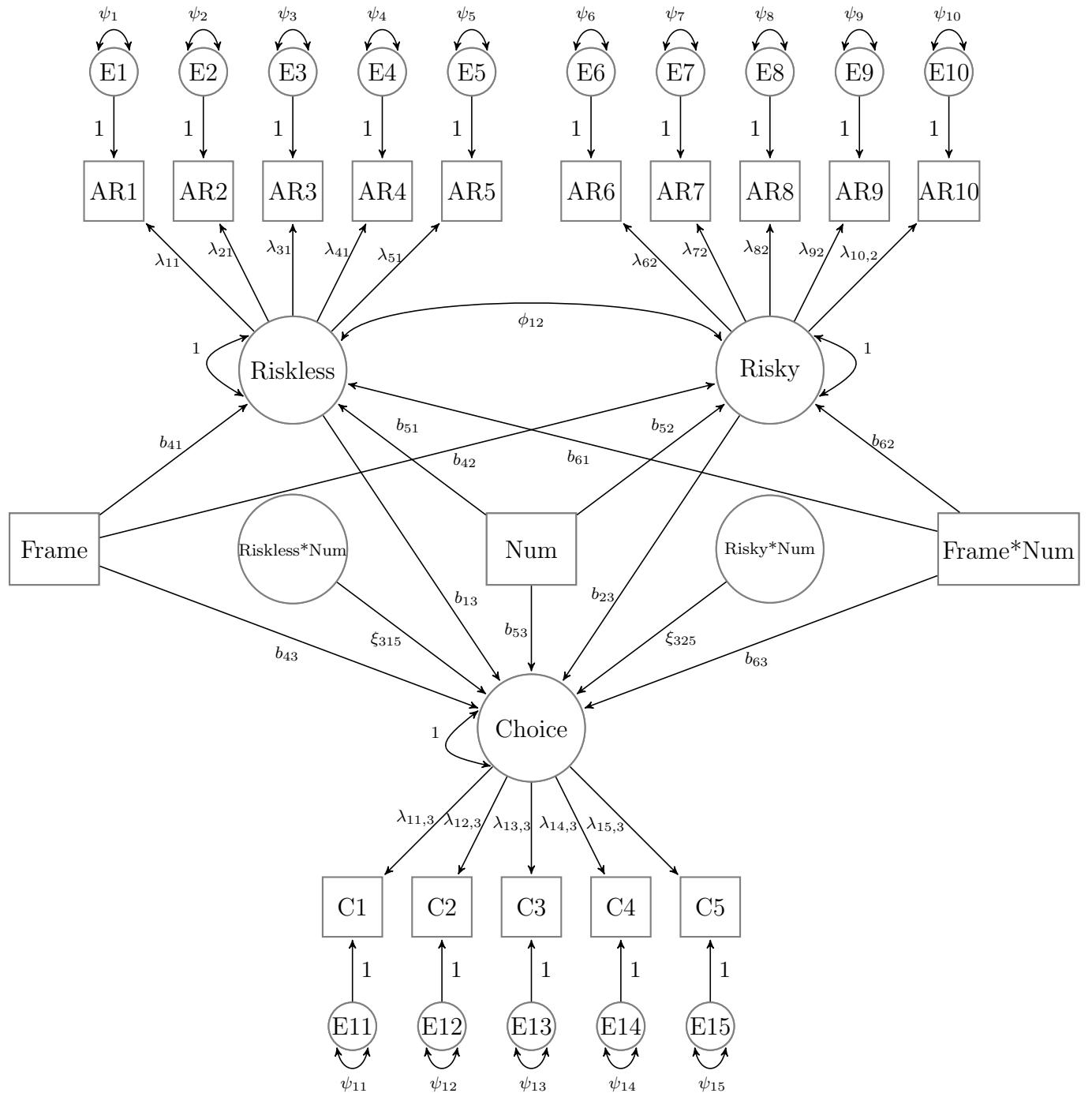


Figure 10.1: Path diagram of the structural equation model used in the application.

representing the riskless attraction ratings, a risky variable representing the risky attraction ratings, and a choice variable representing the choice ratings. The experimental manipulation (frame), numeracy, and a frame-by-numeracy interaction predict all three of the main latent variables. Finally, there appear two latent interaction variables: one between riskless attraction and numeracy, and one between risky attraction and numeracy. These are notable because (i) they help us study the original authors' focal hypotheses, and (ii) interactions involving latent variables are difficult to estimate via ML methods.

All observed variables are assumed to be conditionally normal, so that the path diagram can also be written as a combination of Equations (10.13), (10.14), and (10.16). The parameter subscripts in Figure 10.1 match the subscripts that would be used in the equation version of the model (not shown). Additionally, to obtain the interactions from Equation (10.16), we implicitly create a latent numeracy variable that is perfectly related to the observed numeracy variable.

Prior distributions on model parameters are generally taken to be non-informative. For the factor analysis models, many of the prior distributions are listed in Equations (10.7) to (10.9). The remaining parameters include the intercepts associated with observed variables (γ_j) and residual variances (ψ_j); these classes of parameters are assigned priors of

$$\gamma_j \sim N(0, 10^3) \quad (10.21)$$

$$\psi_j \sim \text{Inv-Gamma}(0.1, 0.1), \quad (10.22)$$

where $j = 1, 2, \dots, 15$. The γ_j parameters are given exceptionally-large variances because they are essentially nuisance parameters here. They represent the mean of each observed variable and play little role in the parameters of interest, which include the λ_{jk} , ϕ_{12} , and ψ_j parameters, where $j = 1, 2, \dots, 15$; $k = 1, \dots, 6$. For completeness, note that the variances of the three observed covariates (labeled “Frame,” “Num,” and “Frame by Num”) are fixed to their sample estimates.

For the structural equation model in Figure 10.1, the $b_{k\ell}$ and $\xi_{kq_1q_2}$ parameters (paths between latent variables) can be somewhat difficult to sample. This is because, for these parameters, the chains sampled via JAGS (and related software) exhibit high autocorrelation. When the prior variances are very large, the MCMC algorithm can accept a “bad” posterior value, and the high autocorrelation implies that it will take a long time for the chain to recover from that bad value. As a result, we use “weakly informative” prior distributions on the parameters (Gelman, 2006): prior distributions with variances constrained to reflect plausible parameter values. For example, because the latent variables in Figure 10.1 are constrained to have means of 0 and variances of 1 (and because we know that, for the data considered here, no latent variable will be perfectly predictive of another), we can be nearly certain that the λ_{jk} , $b_{k\ell}$, and $\xi_{kq_1q_2}$ parameters are between -10 and 10 . Because 99% of the normal distribution lies within 3 standard deviations of the mean (and because we expect the parameters to be closer to zero than to -10 or 10), a normal distribution with mean 0 and variance 10 is mildly informative for these parameters. Note that we are supplying no information about the parameters’ signs; the prior distributions are always symmetric around zero. The only mild information we provide is related to the expected variability of the parameters.

In Figure 10.1, there are five main types of parameters: λ_{jk} , $b_{k\ell}$, $\xi_{kq_1q_2}$, ψ_j , and ϕ_{12} . There also exist observed-variable intercepts γ_j that are not displayed. The prior distributions for these classes of parameters are

$$\gamma_j \sim N(0, 10^3) \quad (10.23)$$

$$\lambda_{jk} \sim TN_{\mathbb{R}^+}(0, 10) \quad (10.24)$$

$$b_{k\ell} \sim N(0, 10) \quad (10.25)$$

$$\xi_{kq_1q_2} \sim N(0, 10) \quad (10.26)$$

$$\psi_j \sim \text{Inv-Gamma}(0.1, 0.1) \quad (10.27)$$

$$\phi_{12} \sim \text{Unif}(-1, 1), \quad (10.28)$$

where $TN_{\mathbb{R}^+}$ is a normal distribution truncated from below at 0. The truncated normals are used here as a shortcut to ensure that the sign of the loadings λ_{jk} and the signs of the latent variables θ_i do not switch on one another. The shortcut can be safely used here because we know the loadings all have the same sign and are far from zero. If this were not the case, we could instead estimate each latent variable's variance, fixing a single λ parameter to 1 for each latent variable. We could also adopt a parameter expansion approach, similar to that described for the factor analysis model.

All models are sampled for 5,000 iterations following an adaptation/burn-in of 6,000 iterations. Model convergence is assessed via time series plots and the Gelman-Rubin statistic (Gelman & Rubin, 1992); for the prior distributions outlined above, all parameters achieve values of the Gelman-Rubin statistic less than 1.1.

Results

Correlations between the 15 observed ratings (riskless attractiveness, risky attractiveness, choice) are presented in Figure 10.2 (other descriptive statistics can be found in the original authors' paper). Correlations for the negative frame condition are displayed above the main diagonal, and correlations for the positive frame condition are displayed below the main diagonal. Blue shadings imply negative correlations, and red shadings imply positive correlations. It is seen that the riskless ratings (AR1–AR5) generally display large correlations with one another, as do the risky ratings (AR6–AR10). Correlations between the riskless ratings and risky ratings are also generally positive: for a given scenario, subjects generally found the risky and riskless options to be more or less appealing. This may be because the options had the same expected value, and some subjects found the expected value to be more appealing than others. Correlations between the choice ratings (C1–C5) and the attractiveness ratings are generally smaller, and they generally differ in sign by type of option: the riskless ratings are generally negatively related to choice ratings, while the risky ratings are generally positively related. The original authors' focal hypotheses involve moderation of these latter correlations by numeracy.

Model estimation was on the order of minutes (on our computers, two minutes or less), as opposed to seconds or hours. We build up to the focal hypotheses by first comparing two exploratory factor models: a one-factor model and a two-factor model (i.e., a model with $m =$

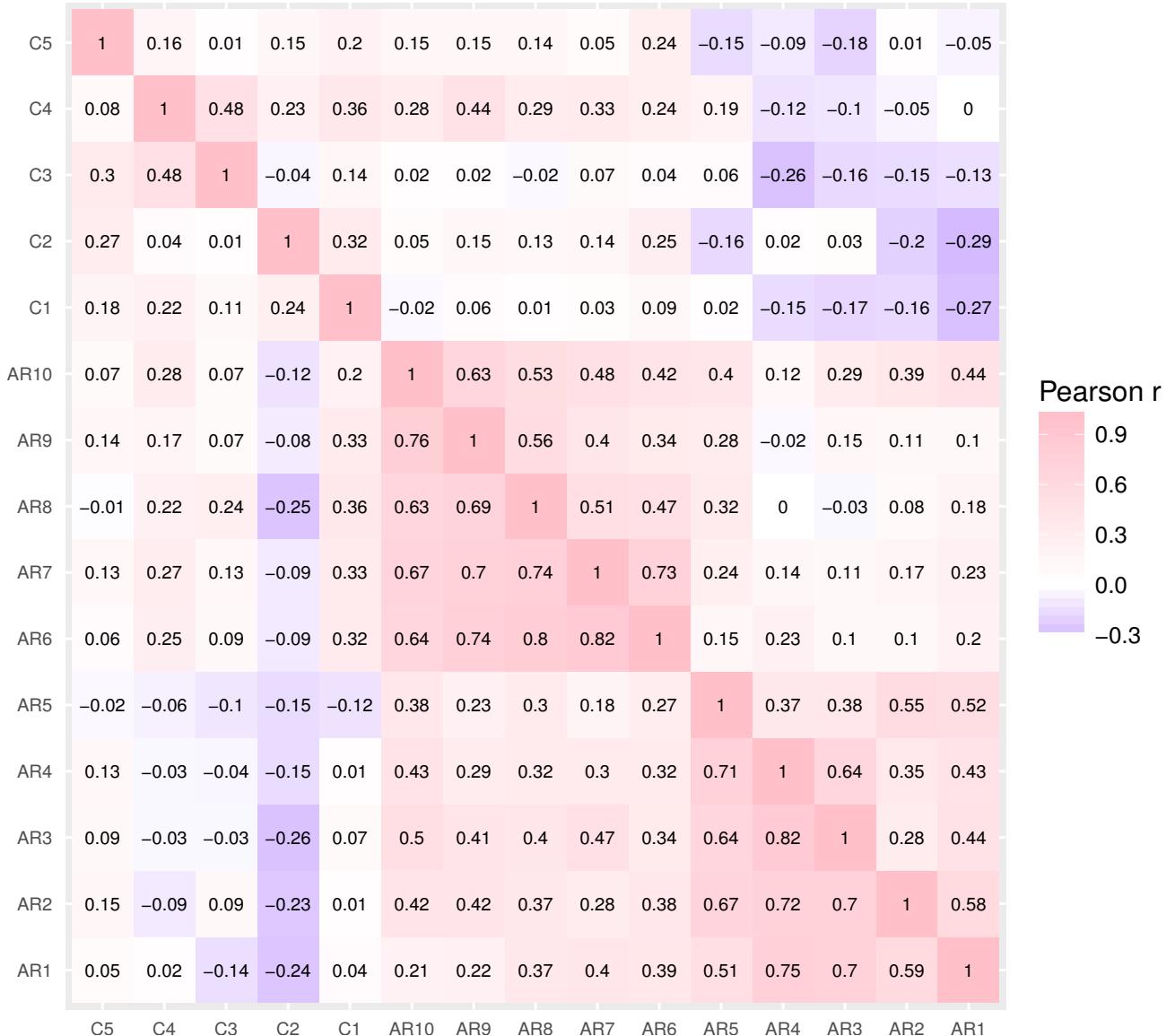


Figure 10.2: Correlation matrix of data from Peters & Levin (2008). Correlations for the positive frame condition are below the main diagonal; correlations for the negative frame condition are above the main diagonal. Variables include riskless attraction rating on items 1–5 (AR1–AR5); risky attraction rating on items 1–5 (AR6–AR10); and choice rating on items 1–5 (C1–C5). Red values signify correlations closer to 1, and blue values signify negative correlations.

Table 10.1: Estimated factor loadings from the exploratory model with $m = 2$. AR1 to AR5 represent attraction to the riskless options for items 1 to 5, and AR6 to AR10 represent attraction to the risky options. Rows correspond to the two factors, labeled “Riskless” and “Risky.”

	AR1	AR2	AR3	AR4	AR5	AR6	AR7	AR8	AR9	AR10
Riskless	1.19	1.15	1.19	1.28	1.13	0.00	0.05	-0.03	-0.02	0.30
Risky	0.00	-0.05	0.00	-0.17	-0.06	1.20	1.14	1.12	1.11	0.98

1 versus a model with $m = 2$). As stated previously, this comparison provides information about the nature of attraction to risky and riskless alternatives: are these attractions two extremes of a single dimension, or are they unique dimensions? In employing the Laplace approximation to compare models, we obtain an estimated log-Bayes factor of 29.02 in favor of the two-factor model. This implies that attraction to risky options and attraction to riskless options should be treated as separate dimensions. Table 10.1 displays the posterior mean factor loadings (i.e., the estimated λ s) and reinforces this result. The table shows that the riskless alternatives (labeled AR1 to AR5) all have large, positive loadings on the first factor and near-zero loadings on the second factor. The risky alternatives (labeled AR6 to AR10) exhibit the converse result, with the estimated correlation between factors being 0.42. The loadings that equal exactly zero (i.e., that equal 0 instead of 0.00) reflect identification constraints for the exploratory model; these constraints led to factor loadings that were easy to interpret, without the need for further rotation. Based on these results, we maintain a 2-factor confirmatory model for the remainder of the analyses. This model fixes to zero the loadings that are already near zero in Table 10.1: the risky alternatives on the “Riskless” factor, and the riskless alternatives on the “Risky” factor.

We are now prepared to additionally model the impact of numeracy on attraction, along with the impacts of frame, numeracy, and attraction on choice via the model displayed in Figure 10.1. The results are displayed in the rightmost columns of Table 10.2, including 95% posterior intervals and log-Bayes factors associated with different parameters.

There is a small effect of frame on riskless attraction, whereby people were more attracted to the riskless option in the positive frame condition (the condition where options were phrased in terms of “number alive”). The posterior interval associated with this effect hovered around zero, and the log-Bayes factor favored a model where this effect equaled zero. We also modeled effects of numeracy and numeracy-by-frame interactions on the attraction latent variables, but these effects were unsupported by both the posterior intervals and Bayes factors.

We now consider effects of frame, numeracy, and the latent attraction variables on choice. We found the standard framing effect on choice, whereby the positive frame condition led people to prefer the riskless option. The log-Bayes factor was around 5, with the posterior interval excluding zero. There was also a small numeracy \times frame interaction: numerate subjects were less impacted by question framing as compared to less-numerate subjects. This interaction had a posterior interval that hovered around zero, with the log-Bayes factor mildly preferring a model without the interaction.

Finally, to address the focal hypothesis, attraction ratings were positively related to choice

Table 10.2: Comparison of original (general linear model) results, 95% posterior intervals from the factor model, and log(Bayes factors) from the structural equation model. Negative log(Bayes factor) implies that the “no-effect” model is preferred. Some of the original results have been converted from t statistics to F statistics for uniformity.

Effect	DV	Original	Posterior interval	Log(BF)
Frame	AT (riskless)	$F(1, 106) = 5.2, p < .05$	(-0.02, 0.81)	-0.87
Frame	AT (risky)	Unreported (not significant)	(-0.58, 0.19)	-2.07
NU	AT (riskless)	$F(1, 104) = 10.8, p < .01$	(-0.07, 0.32)	-3.41
NU	AT (risky)	Unreported (not significant)	(-0.16, 0.23)	-4.31
NU × Frame	AT (riskless)	$F(1, 104) = 3.9, p = .05$	(-0.5 , 0.07)	-2.45
NU × Frame	AT (risky)	Unreported (not significant)	(-0.33, 0.22)	-4.29
Frame	Choice	$F(1, 104) = 18.9, p < .001$	(-3.04, -0.68)	5.08
NU	Choice	$F(1, 104) = 0.04, p = .85$	(-0.53, 0.22)	-2.69
NU × Frame	Choice	$F(1, 104) = 1.8, p = .28$	(0 , 1.23)	-0.8
AT (riskless)	Choice	$F(1, 100) = 6.76, p < 0.01$	(-1.46, -0.2)	1.97
AT (risky)	Choice	$F(1, 100) = 7.29, p < 0.01$	(0.36, 1.85)	5.15
NU × AT (riskless)	Choice	$F(1, 100) = 9.6, p = .002$	(-0.92, -0.12)	1.51
NU × AT (risky)	Choice	$F(1, 100) = 4.4, p = .04$	(-0.09, 0.59)	-2.79

Note: DV = Dependent variable. NU = Numeracy. AT = Attraction.

ratings, more so for numerate subjects. First, increases in riskless attraction were associated with preferences for the riskless option, and increases in risky attraction were associated with preferences for the risky option. The riskless attraction latent variable had a weaker association with choice than did risky attraction, with log-Bayes factors of 1.97 and 5.15, respectively. As Figure 10.1 shows, we also included numeracy \times riskless attraction and numeracy \times risky attraction interactions. The interaction term involving riskless attraction had a posterior interval that did not overlap with zero, while the interaction term involving risky attraction had a posterior interval that did overlap with zero. The log-Bayes factor favors the model that includes the riskless interaction (1.51) but not the risky interaction (-2.79). This suggests a small numeracy \times riskless attraction association with choice, similar to the original authors’ hypotheses.

Discussion

We now discuss some general issues associated with the results in Table 10.2, along with some limitations of the current analyses.

Prior Sensitivity and Comparison Focusing on results in Table 10.2, the magnitude of the Bayes factors is impacted by the prior distributions used (for the model here, prior distributions with smaller variances generally produce larger Bayes factors). This property of Bayes factors is known, and it has been the subject of both praise and criticism (Liu & Aitkin, 2008; Vanpaemel, 2010). The fact that Bayes factors are sensitive to choice of prior distribution is helpful in that they summarize the data while also accounting for our previous

knowledge of the phenomenon under study. Others point out that, if one needs to surpass a Bayes factor threshold for a journal publication (similar to $p < .05$), then it is disconcerting that the threshold can be met by manipulating the prior distribution. To address this issue, one can fit the model under multiple prior distributions to examine the results' sensitivities.

We can further use Table 10.2 to compare results arising from the Bayesian models to those of the original authors. The third column of the table displays the original authors' ANOVA-based results. Some of these results are not based on exactly the same data as our results, because the original authors sometimes focused on three of the five scenarios and sometimes aggregated data (whereas we consistently include all five scenarios). Nevertheless, we see that the posterior intervals and Bayes factors often agree with the ANOVA results, in that intervals fail to overlap with zero and log-Bayes factors are positive when p -values are small. When there are disagreements, the Bayes factors and posterior intervals appear to be more conservative than the other measures.

There are also a few situations where the posterior intervals fail to overlap with zero, yet the Bayes factors favor a model that excludes the effect. These different conclusions stemming from posterior intervals versus Bayes factors arise from their different intents: posterior intervals summarize the value of a single model parameter, while Bayes factors signify whether or not the associated parameter(s) generally improve our model. While the former are readily interpreted and may be more familiar to researchers (due to the similarity between posterior intervals and confidence intervals), the latter seems more useful for summarizing the extent to which a parameter is generally important to one's theory (somewhat similarly to a frequentist likelihood ratio test). Bayes factors additionally account for the complexity afforded to us by extra model parameters, which is relevant here because some types of SEM parameters may afford us more complexity than other types. For further discussion of Bayes factors and of model complexity, see Jefferys and Berger (1992) and Spiegelhalter and Smith (1982).

Limitations If the focus of this paper were risky choice (as opposed to Bayesian latent variable models), further model extension and study would be warranted. The model used here posits three separate latent variables for risky attraction, riskless attraction, and choice. Conditional on these latent variables, the observed variables are posited to be independent. This assumption is likely to be incorrect for observed variables stemming from the same problem. That is, even after conditioning on the latent variables, there may remain a correlation between the three observed variables associated with any one problem (e.g., the variables labeled AR1, AR6, and C1 in Figure 10.1). To account for these issues, we can allow the residual terms associated with these three variables (E1, E6, and E11) to be correlated. The introduction of residual correlations poses a difficult problem for Bayesian SEM estimation (Barnard, McCulloch, & Meng, 2000; Chib & Greenberg, 1998; Palomo, Dunson, & Bollen, 2007), though model estimation is possible (e.g., B. Muthén & Asparouhov, 2012) and will become easier in the future (e.g., Merkle & Rosseel, 2016).

Aside from residual correlations, several issues deserve more attention than they received here. First, it would be useful to include more descriptive statistics that supplement the model. Peters and Levin (2008) provided many tables that are useful along these lines.

Second, it would be useful to study the absolute fits of the models in addition to the relative fits. This may involve posterior predictive checks (e.g., Gelman et al., 2004) or calculation of a Bayes factor comparing the proposed models to saturated models (i.e., multivariate normal models with a free parameter for each mean and covariance). Finally, the statistical issue of suppression may be considered in more detail. The SEM from Figure 10.1 involves two positively-related latent variables predicting a third latent variable with opposite signs. This can lead to instability in the associated regression weights and standard errors. While the signs of the regression weights involving latent variables are the same as the majority of relevant, observed-variable correlations in Figure 10.2, some instability may still exist.

In the General Discussion below, we provide detail about further model extensions, uses, and needs.

General Discussion

In this paper, we have demonstrated the utility of applying Bayesian latent variable models to multivariate experimental psychology data. We first provided background on factor analysis models, structural equation models, and issues specifically related to Bayesian estimation of these models. We then considered some model extensions and applied them to data from a decision making experiment on risky choice. The models allowed us to draw conclusions across multiple stimuli in a unified manner, removing the need for data aggregation and for separate, stimulus-specific analyses. Bayesian methods further allowed us to easily estimate models with latent variable interactions, which are difficult to estimate via ML methods. In the paragraphs below, we address additional Bayesian model extensions and needs. These include general comparisons to ML models, considerations of exploratory factor analysis, and cognitive psychometric models.

Bayesian versus ML Advantages of the Bayesian approach to structural equation modeling include easy extension to complex situations, along with non-asymptotic estimates of the variability in parameter estimates. Expanding on the former point, the methods described here were easily extended to handle latent variable interactions, and they are also easily extended to handle models with both continuous and ordinal observed variables. In particular, Song and Lee (2012b) provide a tutorial on Bayesian estimation of these types of models, with further details in their book (Song & Lee, 2012a). These methods are very powerful because they provide a single framework for estimating many structural equation models that a researcher may conceptualize. This framework is not necessarily the best for building an intuition of structural equation models, however, because it is so general. Researchers' desired models can often be written in a manner that is more concise than the general framework.

Disadvantages of Bayesian structural equation models, as compared to ML models, lie in model specification and estimation. Model specification in, e.g., JAGS is very different from model specification in traditional ML software like Mplus or LISREL, so that users coming from these traditions will find that their prior software experience is not terribly helpful. Relatedly, Bayesian methodology has more “moving parts,” including prior distri-

bution specification and assessments of model convergence. These represent new topics for users of ML software, but we are optimistic that users can master the new topics.

Bayesian exploratory factor analysis Bayesian methods offer novel ways to estimate exploratory factor models. In this paper, we always fixed specific parameters to zero in order to identify model parameters. If the factor model is exploratory, we could then rotate model parameters' (the λ s) posterior means to have a more-interpretable structure. This is a two-step process, just like the ML case: we first estimate the model, and we then rotate parameters. Under the Bayesian approach, we could simultaneously rotate and sample the factor loading matrices at each iteration of the MCMC algorithm, which may lead to more-interpretable solutions than the ML algorithm (which can only use the single set of factor loadings arising from the ML estimates).

Related methods have been the focus of recent work. In particular, Peeters (2012a) describes a method by which factor models with multiple values of m are simultaneously estimated, rotated, and compared to one another via marginal likelihoods (the building blocks of Bayes factors). Conti, Frühwirth-Schnatter, Heckman, and Piatek (2014) propose a method whereby the value of m is allowed to change from iteration to iteration, with the possibility that some latent traits are associated with no observed variables. Importantly, the latter authors assume that every observed variable has a nonzero loading associated with only one latent trait; this assumption is restrictive and results in a model that, using the definitions in this paper, would be called “confirmatory” instead of “exploratory.” Finally, Ročková and George (2014) propose an alternative method whereby m is set to a large value, and “spike-and-slab” priors are used to enforce simplicity of the Λ matrix. These prior distributions fix weak loadings to zero, so that smaller values of m are obtained by fixing to zero all loadings associated with a latent trait. This method is computationally difficult but appears promising for future application.

Cognitive psychometric models The methods described here are highly related to the development of *cognitive* latent variable models (e.g., Nunez, Srinivasan, & Vandekerckhove, 2015; Turner et al., 2013; Turner, van Maanen, & Forstmann, 2015; Vandekerckhove, 2014), where the latent variables are used to tie multiple types of data (response time, BOLD response, survey data, etc) together in a single model. For example, in Vandekerckhove’s application, a diffusion model was used to describe response times from an executive functioning task. Participants also completed two depression scales, with the latent variables in the model predicting both the depression scales and diffusion model parameters. This allowed for novel information about specific aspects of performance on the response time task that are related to the depression scales. Cognitive psychometric models generally offer new avenues for combining psychometric knowledge with cognitive modeling (also see, e.g., Tuerlinckx & De Boeck, 2005; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011).

The model extensions mentioned here, coupled with computational advances in model estimation and increased communication between experimental psychologists and psychometrists, yield fruitful prospects for the use of Bayesian latent variable models in cognitive science. We encourage readers to explore these prospective models and the novel inferences

that they can provide.

Computational Details

All results were obtained using the R system for statistical computing (R Development Core Team, 2014) version 3.2.3 and JAGS software for Bayesian computation version 3.4.0, employing the helper package *rjags* 4-5 (Plummer, 2014). R and the package *rjags* are freely available under the General Public License 2 from the Comprehensive R Archive Network at <http://CRAN.R-project.org/>. JAGS is freely available under the General Public License 2 from Sourceforge at <http://mcmc-jags.sourceforge.net/>. R and JAGS code for replication of our results is available at <http://semtools.R-Forge.R-project.org/>.

References

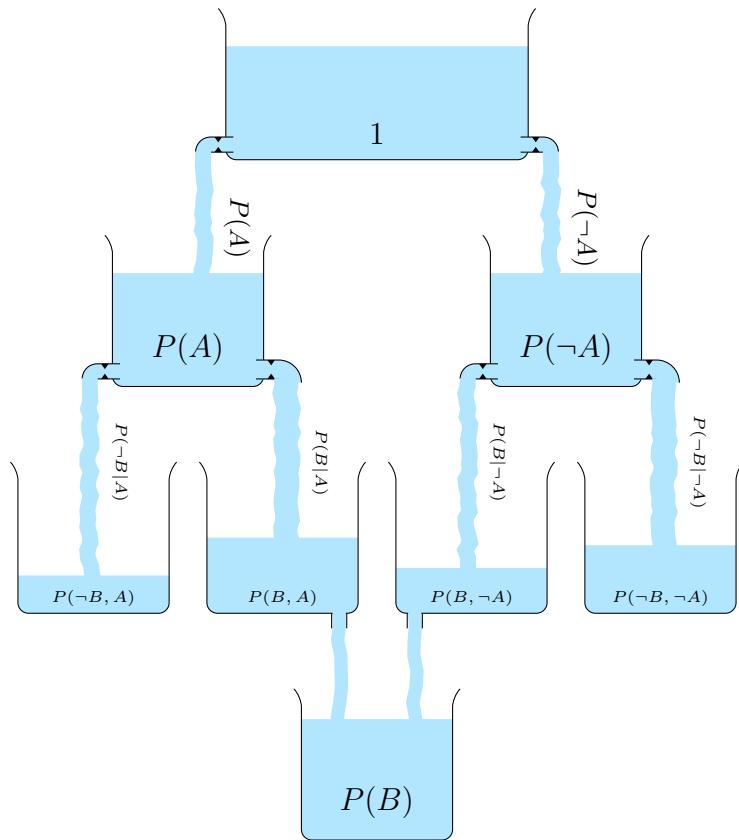
- Bagozzi, R. P., & Yi, Y. (1989). On the use of structural equation models in experimental designs. *Journal of Marketing Research*, 26, 271–284.
- Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10, 1281–1311.
- Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology*, 34, 93–99.
- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Chichester, England: Wiley.
- Bollen, K. A., & Davis, W. R. (2009). Two rules of identification for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 523–536.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150.
- Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85, 347–361.
- Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, 183, 31–57.
- Cudeck, R., Harring, J. R., & du Toit, S. H. C. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, 34, 131–144.
- Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36, 1–34.
- 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.
- Donaldson, G. W. (2003). General linear contrasts on latent variable means: Structural equation hypothesis tests for multivariate clinical trials. *Statistics in Medicine*, 22, 2893–2917.

- Dutta, R., & Ghosh, J. K. (2013). Bayes model selection with path sampling: Factor models and other examples. *Statistical Science*, 28, 95–115.
- Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99, 537–545.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.
- Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13, 163–185.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–511.
- Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18, 306–320.
- Hoyle, R. H., & Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 301–316). Thousand Oaks, CA: Sage.
- Jefferys, W., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80, 64–72.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Jöreskog, K. G. (1979). Author's addendum. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models* (pp. 40–43). Cambridge: Abt Books.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65, 457–474.
- Klein, A., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, 42, 647–673.
- Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). New York: American Elsevier Publishing Co.
- Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, 46, 153–160.
- Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester: Wiley.
- Lee, S.-Y., & Song, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, 29, 23–39.
- Lee, S.-Y., Song, X.-Y., & Tang, N.-S. (2007). Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling*, 14, 404–434.
- Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, 92, 648–655.

- Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52, 362–375.
- Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14, 41–67.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman & Hall/CRC.
- MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, 17, 340–345.
- Marsh, H. W., Wen, Z., Hau, K.-T., & Nagengast, B. (2013). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 267–308). Information Age Publishing.
- Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, 40, 505–517.
- Merkle, E. C. (2011). A comparison of imputation methods for Bayesian factor analysis models. *Journal of Educational and Behavioral Statistics*, 36, 257–276.
- Merkle, E. C., & Rosseel, Y. (2016). blavaan: Bayesian structural equation models via parameter expansion. *Manuscript under review*.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335.
- Muthén, L. K., & Muthén, B. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nunez, M. D., Srinivasan, R., & Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in Psychology*, 8(18), 1–13.
- Palomo, J., Dunson, D. B., & Bollen, K. (2007). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 163–188). Elsevier.
- Peeters, C. F. W. (2012a). *Bayesian exploratory and confirmatory factor analysis: Perspectives on constrained-model selection*. Unpublished doctoral dissertation, Utrecht University.
- Peeters, C. F. W. (2012b). Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, 77, 288–292.
- Peeters, C. F. W., Dziura, J., & van Wesel, F. (2014). Pathophysiological domains underlying the metabolic syndrome: An alternative factor analytic strategy. *Annals of Epidemiology*, 24, 762–770.
- Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, 3, 435–448.
- 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*.
- Plummer, M. (2014). *rjags: Bayesian graphical models using MCMC*. (R package version 3-13)
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2, 13–43.
- Press, S. J. (1972). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. New York: Holt, Rinehart, and Winston.
- R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-07-0)
- Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Beverly Hills, CA: Sage.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30(3), 359–383.
- Ročková, V., & George, E. I. (2014). *Fast Bayesian factor analysis via automatic rotations to sparsity*. (Obtained Jan 8, 2015 from <http://www.ssc.upenn.edu/~fdiebold/papers/misc/RockovaAndGeorge2014FactorAnalysis.pdf>)
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877–903.
- 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.
- Russell, D. W., Kahn, J. H., Spoth, R., & Altmaier, E. M. (1998). Analyzing data from experimental studies: A latent variable structural equation modeling approach. *Journal of Counseling Psychology*, 45, 18–29.
- Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37–52.
- Song, X.-Y., & Lee, S.-Y. (2012a). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. Chichester, England: Wiley.
- Song, X.-Y., & Lee, S.-Y. (2012b). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, 56, 135–148.
- Spiegelhalter, D. J., & Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B*, 44, 377–387.
- Stan Development Team. (2014). *Stan modeling language users guide and reference manual, version 2.5.0*.
- Steiger, J. H. (2001). Driving fast in reverse. *Journal of the American Statistical Association*, 96, 331–338.
- Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2014). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management*.
- Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

- Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, 70, 629–650.
- Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72, 193–206.
- Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, 122, 312–336.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.
- van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118, 339–356.
- Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, 60, 58–71.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.
- Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, 66, 383–401.
- Voorspoels, W., Rutten, I., Bartlema, A., Tuerlinckx, F., & Vanpaemel, W. (in press). Sensitivity to the prototype in children with high-functioning autism spectrum disorder: An example of bayesian cognitive psychometrics. *Psychonomic Bulletin & Review*.
- 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.
- Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, 89(5), 696–716.



An illustration of the sum rule of probability. Probabilities are often conveniently thought of as volumes. If the volume in the top reservoir is a standard unit and the rate of flow through the valves is proportional to the unconditional probabilities $P(A)$ and $P(\neg A)$ and the indicated conditional probabilities, then the volume in the bottom reservoir is equal to the unconditional probability $P(B)$.