How to become a Bayesian in eight easy steps: An annotated reading list

Alexander Etz, Quentin F. Gronau

University of Amsterdam

Fabian Dablander

University of Tübingen

Peter A. Edelsbrunner

ETH Zürich

Beth Baribault

University of California, Irvine

Abstract

In this guide, we present a reading list to serve as a concise introduction to Bayesian data analysis. The introduction is geared toward reviewers, editors, and interested researchers who are new to Bayesian statistics. We provide commentary for eight recommended sources, which together cover the theoretical and practical cornerstones of Bayesian statistics in psychology and related sciences. The resources are presented in an incremental order, starting with theoretical foundations and moving on to applied issues. In addition, we outline an additional 28 articles and books that can be consulted to gain background knowledge about various theoretical specifics and Bayesian approaches to frequently used models. Our goal is to offer researchers a starting point for understanding the core tenets of Bayesian analysis, while requiring a low level of time commitment. After consulting our guide and the outlined articles, the reader should understand how and why Bayesian methods work, and feel able to evaluate their use in the behavioral and social sciences.

How to become a Bayesian in eight easy steps: An annotated reading list

Introduction

In recent decades, significant advances in computational software and hardware have allowed Bayesian statistics to rise to greater prominence in psychology. 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 steep learning curve associated with using and interpreting Bayesian statistics. Many seminal texts in Bayesian statistics are dense, mathematically demanding, and assume some background in mathematical statistics (e.g., Gelman, Carlin, Stern, & Rubin, 2014). Even texts that are geared toward psychologists (e.g., Lee & Wagenmakers, 2014; Kruschke, 2015), while less mathematically difficult, still require 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 three 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. For those who would like to continue studying, the Appendix provides a number of supplemental sources that would be of interest to researchers new to Bayesian methods. 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 A1).

Overall, the guide is designed such that a researcher might be able to read all eight of the highlighted articles[1] and some supplemental readings within a few days. 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 (Kruschke, 2015) between possible states of nature. The third source demonstrates how the concepts established in the previous sources lead to many practical benefits for experimental

[1] Links to freely available versions of each article are provided in the *References* section.

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).

1. 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—colluding with Dr. Bristol's fiancé, Dr. William Roach—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 likely 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.

[2] As a historical note: Distinguishing milk-first from infusion-first tea preparation was not a particular affectation of Dr. Bristol, but a cultural debate that has persisted for over three centuries.

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 .1, 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 .03, 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 the differently made teas starts by assigning *a priori* probabilities to each value 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 Lindley's Figure 2, and note the separate amount of probability assigned to $p = .5$). The remaining probability is distributed among 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

[3]If the null hypothesis is not initially considered tenable, then we may assign no separate probability to

probabilities using *Bayes' rule*. If the observations better fit the null hypothesis (pure guessing), then the probability assigned to the null hypothesis will increase; if the data better fit 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 is the *Bayes factor* (Etz & Wagenmakers, 2015; Kass & Raftery, 1995; see also Rouder et al., 2009, and Dienes, 2011, below).

A key takeaway from this paper is that Lindley's 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.

2. 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 possibilities. Kruschke uses an example featuring

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). Alternatively, if other values of the parameter are considered impossible—such as rates that are below chance—then these may be given zero prior probability.

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 $\frac{1}{4}$ 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 $\frac{1}{3}$ (Figure 2.1, p. 17). What Holmes has done is reallocate the credibility across the different possible causes based on 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, 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 are interested in which values for the mean are 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., 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. 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 key takeaway is the idea of sequential updating from prior to posterior (Figure 2.1, p. 17) as data are collected. As Dennis Lindley famously said: "Today's posterior is tomorrow's prior" (Lindley, 1972, p. 2).

3. 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 (what Dienes calls "orthodox" and we have called "classical;" we shall use these terms interchangeably) and Bayesian paradigm which have practical implications for how experimental psychologists conduct experiments, analyze data, and interpret results. Throughout the paper, Dienes also discusses "subjective" (or "context-dependent") Bayesian methods which allow for inclusion of relevant problem-specific knowledge to enter in to the formation of one's statistical model.

The probabilities of data given theory and 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 is a result of the fact that frequentist statistics only allows 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 $P(\text{theory} \mid \text{data})$: 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 classic statistical paradigm, it is necessary to specify in advance how the data will be collected. In practice, one usually has to specify how many

[4] The conditional probability (P) of data given ($\mid$) theory.

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 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. 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 stops when the p-value is sufficiently low. In contrast, if the null hypothesis is true, then the Bayes factor, the Bayesian method of hypothesis testing, will approach infinite support in favor of the null hypothesis as more observations are collected. 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 a theory does not necessarily need to precede the data when a Bayesian approach is adopted: The evidence will be the same.

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 as 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, in press), Dienes provides an online calculator[5] 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. To properly evaluate authors' modeling decisions, reviewers and editors should see the appendix of Dienes' article for a short review of how to appropriately specify prior distributions that incorporate relevant theoretical information (see also Dienes, 2014, for more details and worked examples).

4. 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., van 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 not possible in general, since significance tests are asymmetric; they can only serve to *reject* the null hypothesis and never to affirm it. One benefit of

[5]http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm

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, in the *Further Reading* appendix). This is made possible by the use of *Bayes factors*.[6] The 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 $f(\text{data} \mid H_i)$.[7] The Bayes factor for H_0 versus H_1 is simply the ratio of $f(\text{data} \mid H_0)$ and $f(\text{data} \mid H_1)$ 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.

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 comparing a point null hypothesis to a point alternative hypothesis is too constraining, so Bayesians specify a range of plausible values that the parameter might take under the alternative hypothesis. This *prior distribution* serves as a weighting function when computing the probability of the data under the alternative hypothesis (for an intuitive illustration, see Gallistel, 2009 in the *Further Reading* appendix).

[6]Readers for whom Rouder and colleagues' (2009) treatment is too technical could focus on Dienes' mathematical ideas and motivations underlying the Bayes factor.

[7]The (average) probability (f) of the observed data given ($\mid$) 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.

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 (or "objective") Bayes (Berger, 2006) and context-dependent (or "subjective") Bayes (Goldstein et al., 2006). 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 accurately encode our prior information about the effect under study, do not necessarily have the same desirable mathematical properties, and are often represented with raw effect sizes.

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 σ. If σ is too large, the Bayes factor will always favor the null model, so such a choice would be unwise (but see also DeGroot, 1982). 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 value. 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 a narrower 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). 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 et al. 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 prior distribution to reflect more contextually relevant information while maintaining the desirable properties attached to prior distributions of this form. The latter shows why the Bayes factor will often show evidence in favor of the null hypothesis when 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 their 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 (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 issue), and the use of cognitive models for theory testing (Lee, 2008).

5. 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 and 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 the 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 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 models with distributed parameters. 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. Vandekerckhove and colleagues (2015) introduce two methods to ease the computational burden: importance sampling and the Savage-Dickey density ratio (see also Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010, in the appended *Further reading* section). They also provide code to estimate the multinomial processing 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 then the sources reviewed above, this article can greatly influence how one thinks about models and methods for selecting among them.

6. Bayesian estimation

Source: 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.

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). After explaining this alternative approach (p. 848), Bayesian estimation of focal parameters and the specification of prior distributions is discussed through 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 time-series model 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 empirical data. If the model is making appropriate predictions, the simulated data and the empirical 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 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.

7. Prior elicitation

Source: Lee and Vanpaemel (this issue) — 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 issue) 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 is an extension of general purpose statistical modeling (e.g., linear regression). Cognitive models make explicit and instantiate a theory, and thus 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).

The parameters in such a cognitive model usually have a direct psychological interpretation. To make this point clear, the authors discuss three cognitive models and show how the parameters instantiate relevant information about psychological processes.

What information can we use to specify a prior distribution? First, because there is a mapping between the parameters and psychological interpretation, theory constrains parameter values. For example, a parameter controlling attention generally cannot be smaller than zero. Lee and Vanpaemel also discuss cases in which all of the theoretical content is carried by the prior, while the model itself (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 in seconds or milliseconds). Recently, iterated learning methods, originally proposed to model the evolution of language, have been employed to unravel common knowledge that is available in groups of participants. These methods can also be used to elicit information that is subsequently formalized as a prior distribution.

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. (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 general statistical models rather than cognitive models see Morey (this issue).

8. 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), 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. For example, in his discussion of signal detection theory, Lee highlights how Bayesian methods are able to easily account for individual differences (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 an 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 issue; van de Schoot et al., 2014), and with the interpretation of a variety of analyses, including Bayesian analogs of classical statistical tests (e.g., *t*-tests; Rouder et al., 2009), Bayes factors for hypothesis testing (Dienes, 2011; Vandekerckhove et al., 2015), Bayesian cognitive models (Lee, 2008), and Bayesian model comparison (Vandekerckhove et al., 2015).

Reviewers and editors unfamiliar with Bayesian methods may initially feel apprehensive about evaluating empirical articles in which such methods are applied (Love et al., this issue). Ideally, the present article should help ameliorate this hesitance 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 et al., (2015) in the *Further reading* appendix), and they allow inferences that many researchers in psychology are interested in but are not able to answer with classical statistics (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, in press). In this article, we have provided an overview of sources that should allow a novice to understand how Bayesian statistics allows for these benefits, even without prior knowledge of Bayesian methods.

Acknowledgments

The authors would like to thank E.-J. Wagenmakers and Joachim Vandekerckhove for their guidance in writing this paper. (Continue to list our financial supporters, grants etc.?)

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology*, *95*(3), 631–636. Retrieved from `http://tinyurl.com/aho2014` doi: dx.doi.org/10.1890/13-1452.1

Berger, J. O. (2006). The case for objective bayesian analysis. *Bayesian analysis*, *1*(3), 385–402. Retrieved from `http://projecteuclid.org/euclid.ba/1340371035` doi: 10.1214/06-BA115

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 317–335. Retrieved from `https://projecteuclid.org/euclid.ss/1177013238`

Cornfield, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, *20*, 18–23. Retrieved from `http://www.jstor.org/stable/2682711`

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. Retrieved from `http://tinyurl.com/dienes2011`

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. Retrieved from `http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00781/full`

Dienes, Z., & Overgaard, M. (2015). How bayesian statistics are needed to determine whether mental states are unconscious. *Behavioural methods in consciousness research*, 199–220. Retrieved from `http://tinyurl.com/dienes2015`

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychology research. *Psychological Review*, *70*(3), 193–242. Retrieved from http://tinyurl.com/edwards1963

Etz, A., & Vandekerckhove, J. (in press). A bayesian perspective on the reproducibility project. *PLOS ONE*.

Etz, A., & Wagenmakers, E.-J. (2015). Origin of the bayes factor. *Preprint. Available: arXiv:1511.08180 [stat.OT]*. Retrieved from http://arxiv.org/abs/1511.08180

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. Retrieved from http://tinyurl.com/gallistel

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Taylor & Francis.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. Retrieved from http://tinyurl.com/gelman2014

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. Retrieved from http://tinyurl.com/gelman2013 doi: 10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. Retrieved from http://tinyurl.com/gigerenzer2004 doi: 10.1016/j.socec.2004.09.033

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (Vol. 20). CRC press.

Goldstein, M., et al. (2006). Subjective bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. Retrieved from http://projecteuclid.org/euclid.ba/1340371036 doi: 10.1214/06-BA116

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

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. Retrieved from http://tinyurl.com/jaynes1986

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. doi: 10.1080/14786443608561691

Jeffreys, H. (1961). *Theory of probability* (3rd 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. Retrieved from http://tinyurl.com/kaplan2012

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. Retrieved from http://tinyurl.com/kruschke2015

Lee, M. D. (2008). Three case studies in the bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, *15*(1), 1–15. Retrieved from http://tinyurl.com/lee2008cognitive

Lee, M. D., & Vanpaemel, W. (this issue). Determining priors for cognitive models. *Psychonomic Bulletin & Review.* Retrieved from https://webfiles.uci.edu/mdlee/LeeVanpaemel2015.pdf

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

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. Retrieved from http://tinyurl.com/lindley1993 doi: 10.1111/j.1467-9639.1993.tb00252.x

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*(3), 293–337.

Retrieved from http://tinyurl.com/lindley2000

Lindley, D. V. (2006). *Understanding uncertainty.* John Wiley & Sons.

Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., ... Wagenmakers, E.-J. (this issue). Bayesian benefits with jasp: A fresh way to do statistics. *Psychonomic Bulletin & Review*.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., ... Wagenmakers, E.-J. (2015). JASP (version 0.7.1.12). *Computer Software*.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (in press). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*. Retrieved from http://tinyurl.com/zyvgp9y

Mayer, J., Khairy, K., & Howard, J. (2010). Drawing an elephant with four complex parameters. *American Journal of Physics*, *78*(6), 648–649. Retrieved from http://tinyurl.com/gtz9w3q

Myung, I. J., & Pitt, M. A. (1997). Applying occam's razor in modeling cognition: A bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. Retrieved from http://tinyurl.com/myung1997 doi: 10.3758/BF03210778

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241. Retrieved from http://tinyurl.com/nickerson2000 doi: 10.1037//1082-989X.S.2.241

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. Retrieved from http://tinyurl.com/rouder2014 doi: 10.3758/s13423-014-0595-4

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. Retrieved from `http://tinyurl.com/rouder2005`

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903. Retrieved from `http://tinyurl.com/rouder2012regression` doi: 10.1080/00273171.2012.734737

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., 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. Retrieved from `http://tinyurl.com/rouder2009` doi: 10.3758/PBR.16.2.225

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. Retrieved from `http://tinyurl.com/royall2004`

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*. Retrieved from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2604513` doi: 10.1037/met0000061

Senn, S. (2013). Invalid inversion. *Significance*, *10*(2), 40–42. Retrieved from `http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00652.x/full`

Sorensen, T., Hohenstein, S., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint arXiv:1506.06201*. Retrieved from `http://tinyurl.com/zoxbsqq`

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. Retrieved from http://dx.doi.org/10.1080/01973533.2015.1012991 doi: 10.1080/01973533.2015.1012991

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. 300–317). Oxford University Press. Retrieved from http://tinyurl.com/vandekerckhove2015

van de Schoot, R., 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. Retrieved from http://tinyurl.com/vandeschoot

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498. Retrieved from http://tinyurl.com/vanpaemel2010 doi: doi:10.1016/j.jmp.2010.07.003

van Ravenzwaaij, D., 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. Retrieved from http://tinyurl.com/vanRavenzwaaij doi: 10.1037/a0036923

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. Retrieved from http://tinyurl.com/verhagen2014 doi: 10.1037/a0036731

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*(5), 779–804. Retrieved from http://tinyurl.com/wagenmakers2007

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. Retrieved from http://tinyurl.com/wagenmakers2010 doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Morey, R. D., & Lee, M. (in press). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*.

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. Retrieved from http://tinyurl.com/wetzels2011 doi: 10.1177/1745691611406923

Appendix

Further reading

In this Appendix, we provide a concise overview of 28 additional articles and books that provide further discussion of various theoretical and applied topics in Bayesian inference. For example, the list includes practical sources 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, 2015), and design (i.e., power) analysis (Schönbrodt et al., 2015). To aid in readers' selection of sources, we have summarized the associated focus and difficulty ratings for each source in Figure A1.

Recommended articles

9. **Cornfield (1966)** — Sequential Trials, Sequential Analysis, and the Likelihood Principle. *Theoretical focus (3), low difficulty (2).*

   A short exposition of the difference between Bayesian and classical inference in sequential sampling problems. Highly recommended and easy to follow.

10. **Lindley (2000)** — The Philosophy of Statistics. *Theoretical focus (1), moderate difficulty (6).*

    Dennis Lindley, a famous 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 rewarding paper. Much of the more technical mathematical exposition can be skipped with negligible loss of understanding.

13. **Myung and Pitt (1997)** — Applying Occam's Razor in Cognitive Modeling: A Bayesian Approach. *Balanced focus (6), high difficulty (9).*

    This paper largely reintroduced Bayesian methods to modern psychologists, discussing the allure of Bayesian model comparison for non-nested models and providing worked examples. The authors provide great discussion on the principle of parsimony, thus this paper serves as a good followup to the discussion in Section 5 above.

14. **Wagenmakers, Morey, and Lee (in press)** — Bayesian Benefits for the Pragmatic Researcher. *Applied focus (9), low difficulty (1).*

    Provides pragmatic arguments for the use of Bayesian inference on two examples featuring Eric Cartman and Adam Sandler. This paper is clear, witty, and persuasive.

15. **Rouder (2014)** — Optional Stopping: No Problem for Bayesians. *Balanced focus (5), moderate difficulty (5).*

    Provides a simple illustration of why Bayesian inferences are always valid in the case of optional stopping. A natural followup to Dienes (2011), which is covered in Section 3 above.

16. **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.

17. **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.

18. **Ly et al. (in press)** — 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 (although see Etz & Wagenmakers, 2015). The second part of the paper explains the computations in detail for t-tests and correlations. Especially the first part is essential in grasping the motivation behind the Bayes factor.

19. **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 old and makes for rather slow reading, but Jeffreys's writing is insightful nonetheless.

20. **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.

21. **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 by p values, effect sizes, and Bayes factors differ. An illuminating reference to understand the practical differences between various ways of inference.

22. **Vanpaemel (2010)** — Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor. *Theoretical focus (3), high difficulty (7).*

    Defends Bayes factors against the common criticism that the inference is sensitive to specification of the prior. Argues that this sensitivity is valuable and desirable.

23. **Royall (2004)** — The Likelihood Paradigm for Statistical Inference. *Theoretical focus (2), moderate difficulty (4).*

    An accessible introduction to the Likelihood principle, and its relevance to inference. Contrasts different accounts of statistical evidence. A more complete account is given in Royall (1997).

24. **Gelman and Shalizi (2013)** — Philosophy and the Practice of Bayesian Statistics. *Theoretical focus (2), high difficulty (7).*

    This is the target article of an excellent special issue on the philosophy of Bayesian inference. Best read in discussion groups, it promises intriguing and fundamental discussions about the nature of inference.

25. **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).

26. **Gallistel (2009)** — The Importance of Proving the Null. *Applied focus (7), low difficulty (1).*

    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.

27. **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.

28. **Sorensen et al. (2015)** — 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. The paper and code are hosted on github, which serves as a good example of reproducible research.

29. **Schönbrodt et al. (2015)** — Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Applied focus (7), 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.

30. **Kaplan and Depaoli (2012)** — Bayesian Structural Equation Modeling. *Balanced focus (6), moderate difficulty (6).*

    One of few available practical sources on Bayesian structural equation modeling. The article focuses on the Mplus software but represents a generic source beyond.

Recommended books

31. **Lee and Wagenmakers (2014)** — Bayesian Cognitive Modeling: A Practical Course. *Applied focus (7), low difficulty (3).*

    A further reading beyond Lee (2008) about Bayesian cognitive modeling methods. Friendly introductions to core principles of implementation, including many case examples.

32. **Lindley (2006)** — Understanding Uncertainty. *Theoretical focus (2), moderate difficulty (4).*

    An introduction to uncertainty and how it influences everyday life and science. A largely non-technical text, but a clear and concise introduction to the general Bayesian perspective on decision making under uncertainty.

33. **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 between Bayesian and classical inference.

34. **Stone (2013)** — Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. *Balanced focus (4), moderate difficulty (6).*

    In this clear introductory text, Stone explains Bayesian inference using accessible examples and writes for readers with little mathematical background. Provides Python and MATLAB code on the author's website.

35. **Gill (2014)** — Bayesian Methods: A Social and Behavioral Sciences Approach. *Balanced focus (5), high difficulty (9).*

    This new edition of a classic text in political science gives a mathematically detailed and extensive treatment of Bayesian statistics. Having its own package for the *R* software, and the solutions to exercises online, it is both useful for self-study and as a reference.

36. **Jackman (2009)** — Bayesian Analysis for the Social Sciences. *Applied focus (7), high difficulty (7).*

    Similar to Gill (2014), but less extensive. Also comes with its own package for the *R* software, intriguing real world data examples, and solutions to exercises.

Figure A1. *An overview of focus and difficulty ratings for all sources included in the present paper.* Sources discussed at length in the *Theoretical sources* and *Practical sources* sections are presented in bold text. Sources listed in the appended *Further reading* section are presented in light text. Source numbers representing books are italicized.