Bayesian Benefits for the Pragmatic Researcher

Eric-Jan Wagenmakers[1], Richard D. Morey[2], and Michael D. Lee[3]
1 University of Amsterdam
2 Cardiff University
3 University of California Irvine

Correspondence concerning this article should be addressed to:
Eric-Jan Wagenmakers
University of Amsterdam, Department of Psychology
Weesperplein 4
1018 XA Amsterdam, The Netherlands
E-mail may be sent to EJ.Wagenmakers@gmail.com.

Abstract

The practical advantages of Bayesian inference are demonstrated through two concrete examples. In the first example, we wish to learn about a criminal's IQ: a problem of parameter estimation. In the second example, we wish to quantify support in favor of a null hypothesis, and track this support as the data accumulate: a problem of hypothesis testing. The Bayesian framework unifies both problems within a coherent predictive framework, where parameters and models that predict the data successfully will receive a boost in plausibility, whereas parameters and models that predict poorly suffer a decline. Our examples demonstrate how Bayesian analyses can be more informative, more elegant, and more flexible than the orthodox methodology that remains dominant within the field of psychology.

Keywords: Parameter estimation, updating, prediction, hypothesis testing, Bayesian inference.

On a sunny morning in Florida, while the birds were singing and the crickets chirping, Bob decided to throw his wife from the bedroom balcony, killing her instantly. The case is

This work was supported by the ERC grant "Bayes or Bust!" from the European Research Council. We thank Quentin F. Gronau for his help in constructing the figures, and Helen Braithwaite for her suggestions about the literature on the IQ of criminals. Supplemental materials are available on the Open Science Framework at `https://osf.io/dpshk/`. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

clear-cut and the prosecution seeks the maximum penalty – execution by lethal injection. In a last-ditch attempt to save Bob's life, the defense argues that Bob is intellectually disabled, with an IQ lower than 70, meaning that he is not eligible to receive the penalty (Duvall & Morris, 2006). Indeed, twenty years earlier, when Bob had been incarcerated for a different crime, an entry-level group-administered IQ test had indicated he was intellectually disabled. In response, the prosecution points out that entry-level group-administered IQ tests are known to underestimate IQ (Spruill & May, 1988) and that Bob's true IQ may therefore be much higher than 70. The judge rules that more certainty about the status of Bob's IQ is required, and three additional IQ test are administered individually, yielding scores of 73, 67, and 79. Given this information, what is the probability that Bob's IQ is lower than 70? To answer this question—or, indeed, any worthwhile question about Bob's IQ at all—we cannot use standard p-values and classical confidence intervals (e.g., Pratt, Raiffa, & Schlaifer, 1995). This is a practical problem, not just for Bob, but also for clinicians and researchers who face statistically similar challenges on a regular basis. Below we will demonstrate how questions about Bob's IQ, unanswerable using classical or orthodox statistics, can be addressed effectively through what is known as inverse probability or *Bayesian* inference.

Consider another concrete problem with a little less gravitas. In South Park episode 166, one of the series' main protagonists, Eric Cartman, pretends to be robot from Japan, the "A.W.E.S.O.M.-O 4000". When kidnapped by Hollywood movie producers and put under pressure to generate profitable movie concepts, Cartman manages to generate thousands of silly ideas, 800 of which feature Adam Sandler.[1] We conjecture that the makers of South Park believe that Adam Sandler movies are profitable despite poor quality. For concreteness, we put forward the following *South Park hypothesis*: "For Adam Sandler movies, there is no correlation between box office success and movie quality (i.e., freshness ratings on Rotten Tomatoes)." Our goal is to assess the degree to which the data support the South Park hypothesis. As we will outline below, the orthodox statistics framework is unable to address the question: it does not produce a measure of evidence, and it does not apply to data that become available over time, indefinitely, inevitably, and beyond the control of any experimenter (e.g., Berger & Berry, 1988, Example 1). In contrast, the Bayesian framework coherently updates one's knowledge as new information comes in, seamlessly and in a straightforward manner, without requiring the existence of a sampling plan or a stopping rule.

In our first example, the focus is on estimation: we want to learn about an unobserved parameter, namely Bob's IQ. Questions related to estimation take the general form: "given that phenomenon X is present, what do we know about the size of its influence?". In our second example, the focus is on hypothesis testing: we want to quantify support in favor of an invariance or general law. Questions related to hypothesis testing take the general form: "what evidence do the data provide for the presence or absence of phenomenon X?". Specifically, in the South Park example the question is: "what is the evidence for the presence or absence of a correlation between box office success and quality of Adam Sandler movies?". As the examples demonstrate, the appropriateness of the question depends entirely on context; that is, on what we are willing to assume and what we wish to learn. Nevertheless,

[1] See `https://en.wikipedia.org/wiki/AWESOM-O`.

the testing question logically precedes the estimation question (Jeffreys, 1961; Simonsohn, 2015). For example, one would be ill-advised to estimate the depth with which people can look into the future before having ascertained the existence of the phenomenon in the first place. From Jeffreys's work, we may derive the maxim: "Do not try to estimate something until you have established that there is something to be estimated". However, estimation is easier to understand than testing, and therefore we discuss estimation first.

First Example: Estimating Bob's IQ

Bob's observed IQ scores are determined both by his latent intellectual ability and by the reliability of the IQ test. The literature shows that IQ tests are relatively reliable, with test standard deviations on the order of 7 IQ points. The literature also reports that inmates who were initially classified as intellectually disabled (because they scored lower than 70 on an entry-level group-administered IQ test) perform better when they are later re-assessed using an individually administered test. For the individually administered test, these inmates' IQ scores are approximately normally distributed with a mean of 75 and a standard deviation of 12 (Spruill & May, 1988). In Bayesian statistics, this knowledge can be captured by means of probability distributions. For Bob's true IQ—the key quantity of interest—we quantify our knowledge as Bob's IQ $\sim$ Normal(mean $= 75,$ variance $= 12^2$).[2] This prior distribution is indicated in Figure 1 by the dotted line. Note that this is a distribution of uncertainty, not a distribution of something that can be directly observed. The larger the variance of the prior distribution, the more uncertain we are about Bob's true IQ. For the reliability of the IQ test, we assign a uniform distribution to the test's standard deviation spanning the range of plausible values. Specifically, we use TestSD $\sim$ Uniform(lower bound $= 5,$ upper bound $= 15$), a distribution that indicates every value between 5 and 15 is equally likely a priori.

Having expressed our prior knowledge through probability distributions, we can learn from the data and update our prior distribution about Bob's true IQ. The updated distribution is known as a posterior distribution, and it is shown in Figure 1 by the solid line. The posterior distribution is a combination of our prior knowledge and the information coming from the data. From the prior and posterior distributions we can draw the following conclusions:

1. The posterior distribution is more narrow than the prior distribution, indicating that the Bob's data have reduced the uncertainty about his IQ.

2. Area A covers the prior mass smaller than 70, indicating a prior probability of about 1/3 that Bob's IQ is lower than 70. In other words, the prior odds of Bob's IQ being higher than 70 are about 2 to 1.

3. Area B covers the posterior mass smaller than 70, indicating a posterior probability of about 1/4 that Bob's IQ is lower than 70. In other words, the posterior odds of Bob's IQ being higher than 70 are about 3 to 1.

[2]The $\sim$ symbol, called a tilde, indicates "is distributed as" and indicates that uncertainty about the true value is being treated using the laws of probability.

4. The data have changed the odds that Bob's IQ is higher than 70 by a factor of about $3/2 = 1.5$.

5. Square C highlights the most likely value for Bob's IQ, which is 73.31.

6. Ratio D indicates that the value of 73.31 is 1.37 times more probable than the value of 70.

7. Interval E is a central 95% credible interval, meaning that one can be 95% confident (i.e., the posterior probability equals 95%) that Bob's true IQ falls in the interval ranging from 63.69 to 83.19.

Crucially, none of the statements above—not a single one—can be arrived at within the framework of orthodox methods (e.g., Pratt et al., 1995), no matter how many tests Bob completes, and no matter what prior knowledge might and might not be available.[3] Yet, these statements may be vitally important for quantifying uncertainty, for predicting future events, and for making life-or-death decisions. As is apparent from the above analysis, Bob's data are anything but conclusive, and the judge may well decide that more data are needed in order to make a decision with confidence. In this case, the posterior distribution from Figure 1 will take on the role of prior for the subsequent data set. Such sequential updating will play an important role in the analysis of the South Park hypothesis, to which we turn next.

Second Example: Testing the South Park Hypothesis

The top panel of Figure 2 shows the relation between box office success (in millions of US dollars) and freshness ratings (in proportion of "fresh" judgments) for all Adam Sandler movies from 2000–2015 listed on `www.rottentomatoes.com`. A visual impression supports the South Park hypothesis. A standard Bayesian analysis proceeds as follows. The South Park hypothesis posits that there is no correlation between box office success and freshness ratings, $\mathcal{H}_0 : \rho = 0$. The alternative hypothesis $\mathcal{H}_1$ relaxes the restriction on ρ. However, to quantify evidence the alternative hypothesis $\mathcal{H}_1$ must make *predictions*, and hence our assumptions about ρ should be made precise, by means of a prior distribution. Here we adopt the default assumption that every value of ρ is equally likely a priori (Jeffreys, 1961; for alternative specifications see Wagenmakers, Verhagen, & Ly, in press).

The middle panel of Figure 2 shows the prior and posterior distribution for ρ. At $\rho = 0$, the posterior distribution is 4.429 times higher than the prior distribution, indicating that the data provide support in favor of $\mathcal{H}_0$ (e.g. Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Specifically, the observed data are 4.429 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$; that is, the data shift our prior beliefs about the relative plausibility of the competing hypotheses by a factor of 4.429. This measure of evidential support is known as the Bayes factor (Dienes, in press; Kass & Raftery, 1995;

[3] For instance, an orthodox one-sided t test does not take into account prior information and does not quantify evidence for or against $\mathcal{H}_0$. In addition, the orthodox framework delivers bounds for x% confidence intervals, but it cannot deliver confidence for a desired interval with specific bounds. For a detailed discussion of the differences between confidence and credible intervals see Morey, Hoekstra, Rouder, Lee, and Wagenmakers (in press).

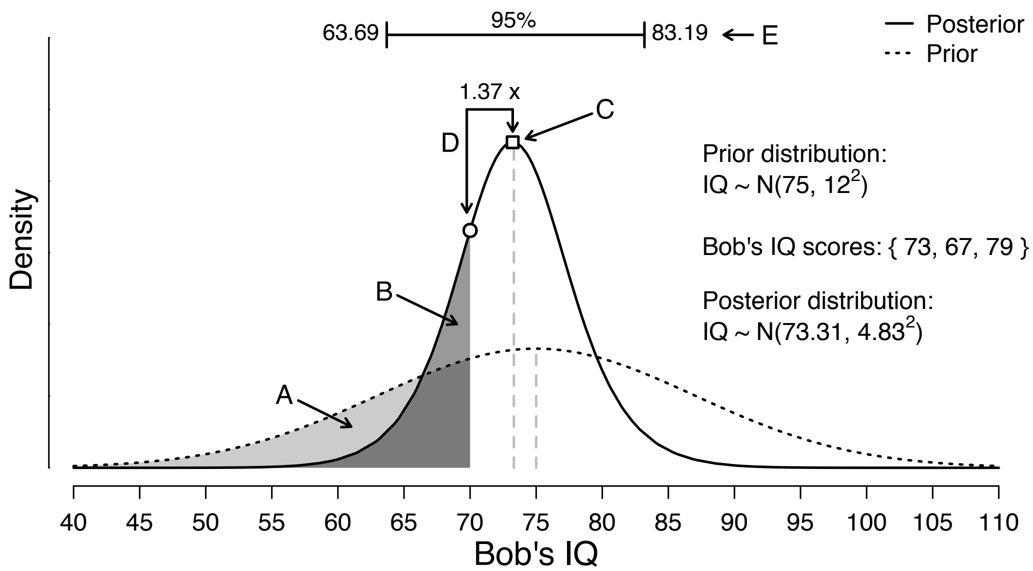
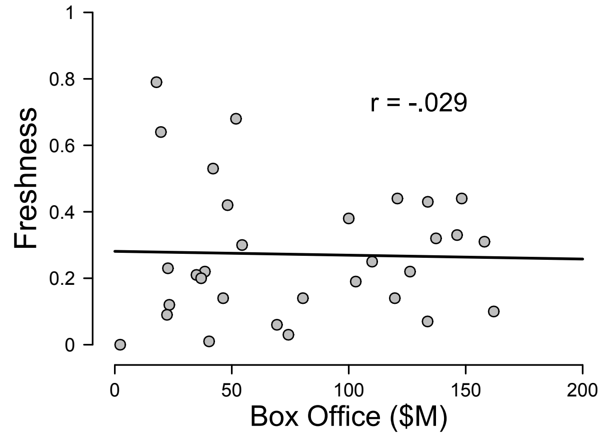
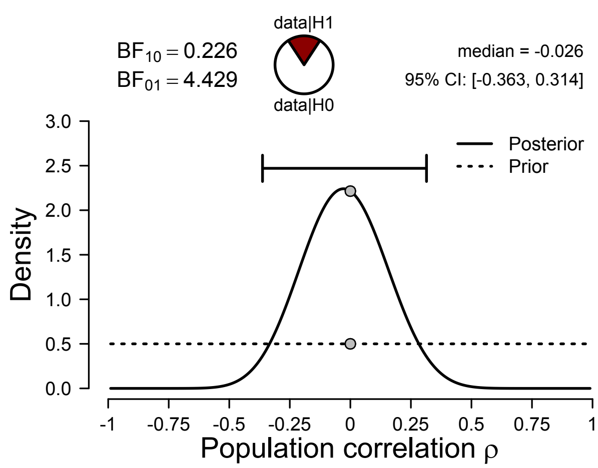


Figure 1. Prior and posterior distributions quantify uncertainty about Bob's IQ. The normal distribution is a close approximation to the posterior. R code is available at `https://osf.io/dpshk/`. Figure available at `http://tinyurl.com/jl5v7p9`, under CC license `https://creativecommons.org/licenses/by/2.0/`.

Mulder & Wagenmakers, in press; Jeffreys, 1961), and it quantifies the ability of each hypothesis to *predict* the observed data (Wagenmakers, Grünwald, & Steyvers, 2006).

The bottom panel shows how the Bayes factor develops as Adam Sandler movies accumulate. This evidential flow can be monitored indefinitely, and does not depend on the knowledge or existence of a sampling plan. An orthodox statistician, in contrast, may refuse to analyze these data at all, arguing—quite correctly—that without knowing how the data came about, the sample space is undefined and no orthodox inference is possible (Berger & Berry, 1988). This limitation is especially relevant whenever researchers study data in a non-experimental context, and it is acute for fields such as astronomy, geophysics, economics, and politics — fields where experiments are rare or impossible. However, the limitation is also relevant for fields where experiments are the norm: monitoring the evidential flow allows researchers to stop the experiment early whenever the evidence is compelling, or continue data collection whenever the evidence is weak. Such sequential designs result in experiments that are more efficient and arguably more ethical than those conducted within the dominant tradition of fixed-N designs.





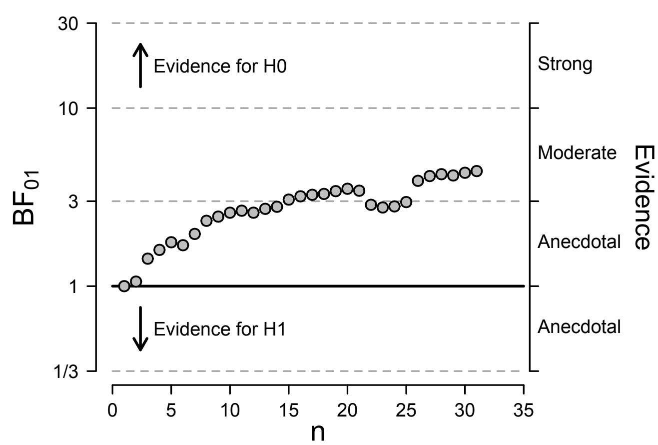


Figure 2. Movies with Adam Sandler are profitable regardless of their quality. Top panel: box office success and freshness ratings for 31 Adam Sandler movies from 2000-2015; middle panel: prior and posterior distribution for the Pearson correlation coefficient, and the evidential support for $\mathcal{H}_0 : \rho = 0$; bottom panel: development of evidential flow as Adam Sandler movies accumulate over time. The figure was created in JASP (`jasp-stats.org`) and is available at `http://tinyurl.com/pfexqhg` under CC license `https://creativecommons.org/licenses/by/2.0/`. An annotated JASP file is available at `https://osf.io/dpshk/`.

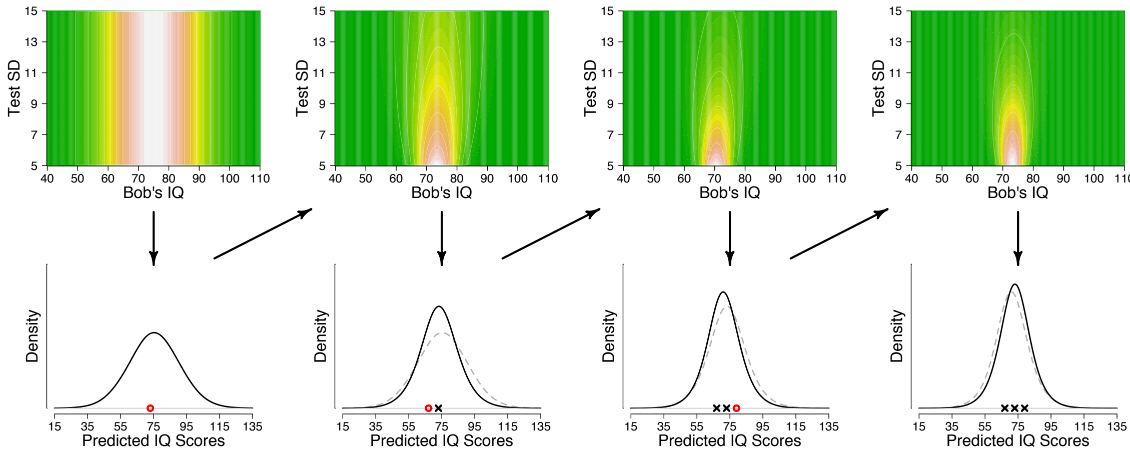


Figure 3. Bayesian inference as the inversion of a generative model applied to the example of estimating Bob's IQ. The generative model (top row panels) makes predictions about data (bottom row panels), and the resulting relative prediction error drives an optimal knowledge updating process. This predictive updating cycle continues indefinitely. Each of the red circles indicates Bob's latest IQ score; the black crosses indicate his old scores. As the data accumulate, the posterior distribution becomes more concentrated and the associated predictions are more precise. Figure available at `http://tinyurl.com/zk5mrn2` under CC license `https://creativecommons.org/licenses/by/2.0/`.

Explanation: Bayesian Inference as Learning From Predictions

There are multiple perspectives on, and interpretations of, Bayesian inference. A cognitive psychologist might consider it a theory of optimal learning from experience, a philosopher might consider it a logic of partial beliefs, and an economist might consider it a normative account of decision making. All of these interpretations are valuable. Here we focus on an interpretation, popular in machine learning, that gave the methodology its original name: inverse probability.

Consider a statistical model for a set of observed data. For a Bayesian, the crucial task is to specify this model generatively, before it has made contact with the observed data. In other words, the model needs to be specified in such a way that it generates data and thereby makes *predictions*. Without making predictions, a model cannot be tested in a meaningful way. When the generative model is then confronted with observed data, the prediction errors drive an optimal inference and updating process that reduces the uncertainty about the components of the generative model. This process is called "inverting a generative model" and it is illustrated in Figure 3. The process of inversion is automatic and described by Bayes' rule. Thus, the central aspect of Bayesian inference is learning from prediction errors by inverting a generative model, such that, upon observing particular consequences, we may learn about their latent causes.

In order to make predictions we need to specify what parameter values are plausible (i.e., the prior distribution), and how a specific set of parameters generates an observed outcome (i.e., the likelihood). Based on these predictions, incoming data can update our knowledge, both about parameters and about models.

A Predictive Perspective on Estimation

Bayes' rule determines how prior distributions are updated by means of the data to produce posterior distributions. This updating process may be given a predictive interpretation, such that parameter values that predict the data well receive a boost in plausibility, and parameter values that predict the data poorly suffer a decline (Morey, Romeijn, & Rouder, in press). The predictive interpretation is clear from rewriting Bayes' rule as follows:

$$\underbrace{p(\theta \mid \text{data})}_{\substack{\text{Posterior beliefs}\\\text{about parameters}}} = \underbrace{p(\theta)}_{\substack{\text{Prior beliefs}\\\text{about parameters}}} \times \underbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}_{\substack{\text{Predictive}\\\text{updating factor}}} \tag{1}$$

This equation shows that the change from the prior to the posterior distribution is brought about by a predictive updating factor. This factor considers, for every parameter value θ, its success in probabilistically predicting the observed data – that is, $p(\text{data} \mid \theta)$ – as compared to the average probabilistic predictive success across all values of θ – that is, $p(\text{data})$.[4]

A Predictive Perspective on Testing

Bayes' rule also determines how data update the relative plausibility of competing models. As with estimation, this updating process may be given a predictive interpretation, as follows:

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs}\\\text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs}\\\text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive}\\\text{updating factor}}} \tag{2}$$

This equation shows that the change from prior to posterior odds is brought about by a predictive updating factor that is commonly known as the Bayes factor. The Bayes factor considers the average predictive adequacy of $\mathcal{H}_1$ and compares it against that of $\mathcal{H}_0$. It should be stressed that these are true predictions, in an out-of-sample sense, since are made without advance knowledge of the data. Predictions can be made sequentially, as the data accumulate one datum at a time. Thus, two models make predictions about the first observation, then receive that datum, update their parameters, make predictions about the second observation, receive that datum, update their parameters, make predictions about the third observation, and so on. The Bayes factor equals the relative cumulative total of the resulting predictive errors. Importantly, this predictive interpretation of the Bayes factor shows that its interpretation does not depend on whether either of the models is true in some absolute sense (see also Feldman, 2015).

In sum, Bayesian parameter estimation and hypothesis testing are based on the same principle of predictive updating. Indeed, there exist statistical scenarios in which parameter

[4]The fact that $p(\text{data})$ is the average predictive success can be appreciated by rewriting it as $\int p(\text{data} \mid \theta)p(\theta)\,\mathrm{d}\theta$.

estimation and hypothesis testing seem to coalesce. For instance, in the case of Bob's IQ one could reformulate the estimation question ("what do we know about Bob's IQ?") in terms of a directional hypothesis test which contrasts $\mathcal{H}-$: Bob's IQ < 70 with $\mathcal{H}+$: Bob's IQ > 70. A strict separation can be achieved when one reserves the term "hypothesis test" for point hypotheses only (Jeffreys, 1961, p. 387).

Concluding Comments

The Bayesian statistical framework offers substantial practical advantages. A Bayesian researcher is able to enrich statistical models with prior knowledge, and this allows the models to make meaningful predictions about data (Myung & Pitt, 1997). The quality of these predictions then drives an optimal process of knowledge updating: parameters and models that predict the data well receive a boost in plausibility, whereas parameters and models that predict poorly suffer a decline. The Bayesian researcher updates the plausibility of parameters and models in a single coherent framework, motivated by relative predictive success. This theoretical foundation allows a clear answer to important practical questions. What is the probability that a parameter is less than some value of interest? What is the relative support for one hypothesis over another? How does this support change as data accumulate over time? These important questions fall outside the purview of the orthodox framework.

For a long time, Bayesian analyses did not find widespread practical application as only a subset of specific models allowed Bayesian results to be obtained in analytic form. However, the development of Markov chain Monte Carlo (MCMC: Gilks, Richardson, & Spiegelhalter, 1996; Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012) has revolutionized the field. Instead of having to derive the posterior distribution mathematically, the MCMC routines can obtain samples from it, and the resulting histogram approximates the posterior distribution to arbitrary precision. Because of MCMC, Bayesian models are now said to be "limited only by the user's imagination".

Psychologists who wish to apply Bayesian analyses to their own data have access to several books and software packages. For books, we recommend Dienes (2008), Lee and Wagenmakers (2013), McElreath (2016), Lindley (2006) and the references therein, and we refer the reader to Etz, Gronau, Dablander, Edelsbrunner, and Baribault (submitted) for more elaborate advice. For software packages, we recommend JASP (`jasp-stats.org`), the BayesFactor package in R (Morey & Rouder, 2015), and the popular programs BUGS, JAGS, and Stan (e.g., Lunn et al., 2012). As more Bayesian course books and user-friendly software packages become available, we expect researchers will increasingly take advantage of the additional possibilities that Bayesian modeling has to offer.

References

Berger, J. O., & Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics: Vol. 4* (pp. 29–72). New York: Springer Verlag.

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. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* New York: Palgrave MacMillan.

Dienes, Z. (in press). How Bayes factors change scientific practice. *Journal of Mathematical Psycholology.*

Duvall, J. C., & Morris, R. J. (2006). Assessing mental retardation in death penalty cases: Critical issues for psychology and psychological practice. *Professional Psychology: Research and Practice*, *37*, 658–665.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (submitted). How to become a Bayesian in eight easy steps: An annotated reading list.

Feldman, J. (2015). Bayesian inference and "truth": A comment on Hoffman, Singh, and Prakash. *Psychonomic Bulletin & Review*, *22*, 1523–1525.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Boca Raton (FL): Chapman & Hall/CRC.

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.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lindley, D. V. (2006). *Understanding uncertainty.* Hoboken: Wiley.

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.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan.* Boca Raton (FL): Chapman & Hall/CRC Press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (in press). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review.*

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (in press). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology.*

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11-1.* Comprehensive R Archive Network.

Mulder, J., & Wagenmakers, E.-J. (in press). Editor's introduction to the special issue on "bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology.*

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory.* Cambridge, MA: MIT Press.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Spruill, J., & May, J. (1988). The mentally retarded offender: Prevalence rates based on individual versus group intelligence tests. *Criminal Justice and Behavior*, *15*, 484–491.

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., Verhagen, A. J., & Ly, A. (in press). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*.