Expert Opinions on How to Conduct and Report Bayesian Inference

B. Aczel[1], R. Hoekstra[2], A. Gelman[3], E.-J. Wagenmakers[4], I. G. Klugkist[5], J. N. Rouder[6], J. Vandekerckhove[6], M. D. Lee[6], R. D. Morey[7], W. Vanpaemel[8], Z. Dienes[9], and D. van Ravenzwaaij[2]

Eotvos Lorand University[1]
University of Groningen[2]
Columbia University[3]
University of Amsterdam[4]
Utrecht University[5]
University of California, Irvine[6]
Cardiff University[7]
University of Leuven[8]
University of Sussex[9]

Abstract

Compared to the relatively standard way of conducting null hypothesis significance testing, there seem to be fairly large differences in opinion among experts in Bayesian statistics on how best to conduct Bayesian inference. Employing Bayesian methods involves making choices about prior distributions, likelihood functions, and robustness checks, as well as on how to report, visualize, and interpret the results. This wide range of choices might make it daunting for social scientists to make the transition to conducting Bayesian inference in their own research. In this review, we conducted an expert survey in which nine of the most prominent Bayesian statisticians in the behavioural sciences shared their thinking on seven key choices that need to be made when conducting and reporting Bayesian inference. This paper highlights the areas of their agreements and the arguments behind their disagreements. The results of an iterative survey show experts agree on many more topics than they disagree on. The overall message is that instead of following rituals, researchers should understand the reasoning behind the different positions and make their choices on a case by case basis.

Introduction

In the social sciences, the dominant way of conducting statistical inference has been through the use of Null Hypothesis Significance Testing (NHST). This approach, however, has long been criticized, and discouraged as the preferred tool for statistical inference (e.g., Cohen, 1995; Cumming, 2008, 2014; Gigerenzer, 2004, 2018; Greenwald, 1975; Hoekstra, Finch, Kiers, & Johnson, 2006; Trafimow, 2003; Wagenmakers, 2007). Proposed alternatives to its current use include, but are not limited to, replacing dichotomous testing with uncertainty intervals (Cumming, 2008), and lowering the threshold for "statistical significance" (Benjamin et al., 2018; de Ruiter, 2018; but see Lakens et al., 2018). A progressively larger part of the statistical community advocates adopting *Bayesian statistics* as an alternative to NHST

(Dienes, 2008; Etz & Vandekerckhove, 2018; Kruschke, 2014; Lee & Wagenmakers, 2005; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007).

Bayesian statistics involves combining what one believes before having seen the data through a *prior* with what the data tell us we should believe through a *likelihood* to end up with a new (typically more informed) belief in the form of a *posterior*. Bayesian statistics can be used for testing (using *Bayes factors*) or for estimation (using *credible intervals*). A full exposition of Bayesian statistics is outside the scope of this article, but for a glossary of the main concepts see Box 1, and for further explanations we refer the interested reader to Kruschke (2014), Lee and Wagenmakers (2013), or Etz and Vandekerckhove (2018).

Bayes factor
The relative support provided by the data for one model over another model in the form of an odds ratio.
(Bayesian) testing
Branch of (Bayesian) statistical inference in which competing hypotheses are tested.
Credible intervals
A probabilistic interval that is believed to contain a given parameter.
Likelihood
The probability of the data given a model or (set of) parameter(s).
Posterior (distribution)
Used in Bayesian inference to quantify a researcher's updated state of belief about some hypotheses (such as parameter values) after observing data.
Prior (distribution)
Used in Bayesian inference to quantify a researcher's state of belief about some parameters *given a model* before having observed any data. Typically represented as a probability distribution over different states of belief.
Prior model probability
Used in Bayesian inference to quantify a researcher's state of belief about the plausibility of a given model before having observed any data.

Box 1. Glossary for the main statistical concepts discussed in this review.

Despite its many advocates, Bayesian inference is currently still only employed by a minority of social scientists. Often-quoted reasons for this still relatively small position in scientific literature are computational difficulties in executing these analyses (van Ravenzwaaij, Cassey, & Brown, 2018), the absence of good easy-to-read educational materials, and the absence of easy-to-use software (but see JASP, 2018).

Another possible barrier preventing the widespread adoption of Bayesian methods is the seemingly large differences in opinions among experts in Bayesian statistics on how best to conduct such analyses (e.g., Albers, Kiers, & van Ravenzwaaij, 2018). Employing Bayesian

methods involves making choices about prior distributions, likelihood functions, robustness checks, reporting, visualization of results, and interpretation. In the context of Bayesian testing, one should decide on testing point hypotheses or interval hypotheses and on the use of evidence thresholds. Why, a potential new user of Bayesian statistics might ask, should I change my practices if experts cannot even agree among themselves on any of the above-mentioned issues? It seems likely that researchers prefer to adhere to practices they are familiar with, until the dust on this perceived debate has settled. Researchers reading only a few sources on Bayesian statistics can easily find diverging recommendations and become discouraged from investing additional time into learning about these methods.

A possible solution to these problems would be a publicly accessible summary of expert opinions on how to conduct and report Bayesian inference. Here, we define experts loosely as those who have published one or more influential articles or books on Bayesian statistics, and who are typically regarded in the field as authorities. By having the experts answer the same questions on how to conduct Bayesian analyses, reading each other's answers, and (potentially) updating their position, it becomes possible for readers to understand the substance of the disagreements among experts. It also allows readers to identify the types of choices for which there is consensus among experts and the type of choices for which experts have different opinions.

Would such an endeavour lead to simple and easy-to-implement heuristics on how to learn and use Bayesian methods? Our answer to this is a categorical no: Users of Bayesian statistics should not hope for any "mindless ritual" any more than users of more traditional methods should (Gigerenzer, 2004, 2018). Instead, a condensed form of the experts' opinions on choices on priors, likelihoods, robustness, reporting, visualization, and interpretation should foster understanding behind the different reasons for making these choices and present them in an accessible form in a single source. This paper presents the opinions of exactly such a Bayesian expert panel. We solicited opinions from nine experts regarding nine topics on how to conduct Bayesian statistical inference by means of iterative surveys.

The remainder of this paper is organized as follows. The methods section lays out how the experts were selected, how the questions were constructed, and how answers were revised. The results section presents summaries of the answers to the nine questions across the panel. The discussion highlights similarities and differences among the experts on each of these main themes and ends with some concluding thoughts.

Method

Participants

This project employed an iterative survey method to explore the agreements and disagreements among experts on conceptual and practical questions in Bayesian analysis. The first two and last authors facilitated the project (henceforth facilitators). The facilitators approached seven Bayesian statisticians (henceforth experts). The criteria for the selection of these experts were that Bayesian inference constituted a central topic of their scientific work in recent years and that they were active in the social sciences. One expert declined, and three new experts were suggested and subsequently approached by the facilitators.

Ultimately, the following nine experts agreed to participate in this study (henceforth they will be addressed by their initials): Andrew Gelman (AG), Eric-Jan Wagenmakers (EJW), Irene Klugkist (IK), Jeffrey N. Rouder (JR), Joachim Vandekerckhove (JV), Michael D. Lee (MDL), Richard D. Morey (RM), Wolf Vanpaemel (WV), and Zoltan Dienes (ZD).

Materials

The first version of the survey consisted of eight questions regarding topics the facilitators deemed relevant. The first wave of experts was asked whether these questions were clear, and whether any important issues were missing. The experts were given the opportunity to suggest modifications in the phrasing of the questions or recommend new questions, which they did in a few instances. Eventually, the survey consisted of nine questions about the following topics: testing vs. estimation, choosing priors and robustness, point null vs. small intervals, reporting of results, visualization, interpretation, and decision making. The survey questions are presented in Appendix 1.

Procedure

Once the survey was finalized, the participating experts were asked to provide their answers to the questions. After all of the responses were collected, the facilitators summarized the answers to each question. If required, the experts were asked to clarify their positions. In the second round, a summary for each question and the detailed responses of all of the experts were shared with the panel members. Thus, experts were given the opportunity to modify their original answers. Following this, the facilitators used the experts' comments to amend and extend the summary text. The facilitators implemented modifications in the summary until all of the experts were satisfied with the text. The first round of the study took about two months and the second round took about one month. The preregistration of our procedure is available here: https://osf.io/q37as/.

Results

The nine experts continued their participation in the study until the end of the project. All of the experts accepted the final version of the opinion summary. Their full responses to the questions are available from https://osf.io/6eqx5/.

Summary of the expert opinions per question

1. When would you recommend using Bayesian parameter estimation and when Bayesian testing (i.e., Bayes factors)? Do you think there is a fundamental difference between the two?
JR, JV, MDL, RM, WV, and ZD stress the (mathematical) similarities between testing and estimation. IK, MDL, and WV stress the different goal of the two in practice, MDL using the analogy of regression analysis and analysis of variance. IK comments that estimation is more informative, but testing may be useful if one needs to make a decision. RM and EJW point out that a big difference between the two approaches lies in the nature of the (joint) prior distribution, which tends to be discontinuous for testing, but continuous for estimation. WV,

EJW, and ZD state that testing is used for checking whether an effect is present and estimation for checking how large the effect is. JV posits we tend to use estimation when we buy into one or several similar models but testing when we buy into (at least two) different ones.

2. A. How should the prior distribution and likelihood function for Bayesian analyses be chosen?

IK points out that typically there is a lot more emphasis on the choice of prior than on the choice of likelihood in a Bayesian context, even though the latter is also very important. IK, WV and MDL favour subjective priors over objective/default/uninformative ones, IK because she believes uninformative priors are unrealistic, and MDL because every scientific endeavour begins with an (informed) choice or guess of both prior and likelihood. RM and ZD point out that uninformative priors should be chosen when assessing evidence for certain parameter values, but informed priors should be chosen when assessing evidence for one model over another. AG typically favours informative priors but thinks that uninformative priors can serve a role in fitting baseline models for comparison. WV stresses that choosing priors and likelihoods should be an iterative process, guided by obtaining prior predictive distributions of the model, and checking they lead to plausible data patterns. JR and WV state that prior and likelihood should be chosen together. JV specifically advocates choosing a sceptic's prior, a believer's prior, and a personal prior, and discuss the (possibly diverging) results. All panel members seem to agree on the following general answer: that depends!

2.B. When and how do you think robustness checks should be performed in Bayesian analyses?

IK warns that reporting robustness analyses may lend to the reported results a false sense of trustworthiness. She thinks robustness checks should be performed, but the crucial step is to determine first which modelling choices may impact the results and perform your checks accordingly. IK, JR, JV, MDL, WV and RM all stress the importance of performing robustness checks over reasonable variations in modelling choices (as opposed to a blanket policy). MDL, IK, and WV specifically warn that robustness checks are typically thought of as applying to prior sensitivity analyses but should also habitually be used to verify the influence of the choice of the likelihood function. RM warns that robustness analyses typically employed in more traditional analysis strategies should still be performed when conducting Bayesian analyses. EJW, and to a lesser extent IK, argue that robustness checks are primarily important when working with non-informative, and therefore more arbitrary, priors. AG stresses practical considerations when deciding to conduct robustness checks. ZD implies that robustness checks may be coupled to decision thresholds, so that one may conclude for what range of prior assumptions a certain decision would be taken.

3. What do you think about using point null hypotheses versus (small) interval hypotheses when testing within the Bayesian framework?

IK recommends first considering if testing is what the researcher wants at all (as opposed to just estimating), after which a researcher should consider what a practically relevant effect is before having seen the data and set up an interval test accordingly. JR, WV and MDL stress the practical usefulness of the point null as a model to reflect invariance. AG thinks point null hypotheses almost never make sense. JV and EJW argue that it would be rare for a point null

and a small interval around null to lead to practically different conclusions. MDL notes that in some cases the parsimonious point null helps flag the need for more data in case a (much) more complex model is believed to be true. RM recommends using whichever you are most interested in (or both to test robustness). ZD notes that the point null is a useful model as an approximation of a near-zero interval.

4. How would you recommend reporting Bayesian analysis results (potentially with a robustness test)? Please provide an example text.

IK and WV recommend reporting the model and its assumptions, prior distributions, potential hypotheses to be evaluated, and details about MCMC sampling when applicable. IK, JV, WV, and MDL do not think a standardized format should exist for reporting Bayesian results, but IK stresses that guidance on the interpretation of results is always useful. JV advocates reporting the prior, choice of likelihood, posterior, and robustness tests (if conducted) at the very least. JR recommends reporting results in terms of competing and completely specified models. RM states the importance of making available raw data and analysis code, as well as providing figures that show estimates with uncertainty. EJW thinks showing plots of the prior and posterior, accompanied by Bayes factors and credible intervals when applicable, are most important when reporting. ZD advocates specifying your hypotheses in terms of distributions with associated robustness regions when reporting results as a means to convey for what range of modelling choices conclusions qualitatively holds.

5. How would you recommend visualizing the results of a Bayesian analysis on diagrams?

IK and JV advocate plotting posteriors of parameters accompanied by measures of uncertainty in case of estimation but believe in case of testing reporting the results in the text will suffice (IK: through a Bayes factor). WV shares the previous sentiment but thinks even in case of estimation visuals are not always useful, leading potentially to information overload. AG, JR and WV prefer plotting marginal predictions of a model and observed data together, if possible. MDL stresses the importance of including uncertainty when visualizing results, while RM stresses the importance of summarizing information in the data in the visuals so that readers can see how authors came to their conclusions. EJW thinks showing plots of the prior and posterior, accompanied by Bayes factors and credible intervals when applicable, are most important when reporting. ZD adds it may be useful in case of testing to incorporate in plots if your Bayes factor reaches some meaningful threshold for drawing conclusions. Most of the experts stress that there should not be a standardized way to visualize data, but instead should be guided by the research topic, the intended audience, and the type of analysis.

6. How would you recommend interpreting Bayesian analysis results (with a robustness test)? Please provide an example text.

IK recommends focusing on what the results mean and what the uncertainties of the presented conclusions are and states that she is against decision thresholds for Bayes factors. JR, RM, and WV argue for focusing on the scientific interpretation over the statistical interpretation. JV prefers an interpretation chain that goes from (modelling) assumptions to observed data to conclusions drawn, possibly with a similar chain for an alternative (but plausible) set of assumptions. MDL think Bayes factors specifically are best presented through the lens of

betting, especially when accompanied by real-world examples of odds that should foster an intuition of what such Bayes factors mean (see MDL's response for some concrete examples). EJW and ZD refer to their answers on #4 where they stress illustrative visualizations and providing ranges for your qualitative conclusions respectively when interpreting results.

7. A. Should we use Bayesian analysis for making decisions about the evidence?
IK, MDL, WV, and ZD answer in the affirmative. JR and RM state that you can if you wish, but do not have to. AG, JR, JV, MDL, WV, and EJW stress the need for utilities in addition to Bayesian statistics in order to be able to do so (with MDL providing a concrete example).

7. B. Would you recommend a decision threshold, an a priori sample size, or anything else?
IK argues against standard decision thresholds, stating that the behaviour of Bayes factors for different kinds of hypotheses is insufficiently understood and that it may lead to arbitrary decision making (both about the fate of the manuscript that reports them and about the true state of the world). MDL is similarly against standardized decision thresholds, warning that the strength of evidence (and the number of data points) need to be understood within the research context. EJW thinks standard decision thresholds are a useful heuristic for evaluating the statistical evidence but does not think it should be used to base decisions about publishing papers on. AG and JV are emphatically against standard decision thresholds and JV similarly sees no statistical reasons for a-priori sample sizes. RM is also against standard decision thresholds but thinks it is useful to conduct an a-priori design analysis and define a sampling plan. WV is against standard decision thresholds, because even the smallest study can contribute useful information. ZD has the unique opinion that standard decision thresholds are useful as a convention and has been active in having journals implement them.

Discussion

Different sources on Bayesian inference sometimes advocate different practices, leaving the user potentially confused about the best way to conduct Bayesian inference. This paper seeks to remedy this situation by presenting the opinions of nine experts on Bayesian statistics on choices related to the formal modelling and reporting. Importantly, experts had access to each other's' statements, the presented summaries of all experts were collected in an iterative process. The results show experts agree on many more topics than they disagree on.

All of the panelists share one sentiment across the board: Use common sense. While some advocate default approaches in certain situations, all point out that blanket policies without thinking do more harm than good. Such advice is easy to give, but hard to follow: Following simple guidelines is simpler than carefully thinking about statistical choices on a case by case basis.

One notable result is that most of the experts argue against decision thresholds for Bayesian testing. This appears to be in opposition to a recent cry for a uniform (though stricter) level of statistical significance (Benjamin et al., 2018), which was based on an equivalence argument with upper bounds of Bayes factors (but see Lakens et al., 2018). One expert argues

in favour of fixed decision thresholds as a useful convention though one that could be over-ridden.

Another topic the panel was not completely unanimous on concerned the value of point-null hypotheses. Many of the experts seem to be of the opinion that it has some practical value as a model, though it should not necessarily be taken literally. Two experts question the value of the point-null, stating that rejecting the null hypothesis is not very interesting in itself.

All of the experts have slightly different preferences on reporting. These seem to be mostly cosmetic differences: Thorough reporting can be done in many ways, and in this sense Bayesian statistics is no different from more traditional forms of inference.

One limitation of our study is that there is no "objective" method for the selection of experts. The label "expert" is somewhat subjective, which means that there are many other potential panels that could have been constructed. Nevertheless, the facilitators believe this expert panel is diverse enough that a representative range of opinions on how to conduct Bayesian inference was reflected.

Another choice we made was to limit ourselves to Bayesian statisticians who are predominantly active in the social sciences. We purposely left out some figureheads from different scientific areas, as our aim was to present an expert opinion paper applicable to Bayesian inference in this field of research.

Throughout the project, we found many benefits of surveying expert opinions as a way to provide the readers with a review of arguments behind diverging views. The summary of opinions of experts with diverse background has probably better guarantees in creating a comprehensive and balanced picture than any review that a single author could write. We recommend the use of this approach in the future, especially for reviews on contested topics.

To conclude our review, to conduct statistical inference is to make choices. This survey demonstrates that for Bayesian inference, this dilemma remains. We hope that the presentation of opinions of leading Bayesian statisticians in the field can guide some of the choices the typical user has to make on this front.

Open Practices Disclosure

The preregistration of our procedure is available here: https://osf.io/q37as/. Although the preregistration protocol stated we would include 7-8 experts, we ended up with 9. All materials of this study are available on the OSF at https://osf.io/6eqx5/ [these will be uploaded before submission]

Author Contributions

BA, DvR and RH conceptualized the project, conducted the study survey and wrote the manuscript. AG, EJW, IK, JR, JV, MDL, RM, WV, and ZD contributed to the summary of this review. All authors reviewed and approved the final version of the manuscript. [the last sentence will be revised before submission]

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Cohen, J. (1995). The earth is round (p < .05): Rejoinder. *American Psychologist*, *50*(12), 1103–1103. https://doi.org/10.1037/0003-066X.50.12.1103

Cumming, G. (2008). Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

de Ruiter, J. (2018). Redefine or justify? Comments on the alpha debate. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-018-1523-9

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse ofp values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. https://doi.org/10.3758/BF03213921

JASP, J. T. (2018). *JASP (Version 0.9. 0.1)[Computer software]*. Amsterdam.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003)., *112*(3), 662–668.

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110*(3), 526–535. https://doi.org/10.1037/0033-295X.110.3.526

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Appendix 1

The Survey Questions Used for Collecting Expert Opinions

1. When would you recommend using Bayesian parameter estimation and when Bayesian testing (i.e., Bayes factors)? Do you think there is a fundamental difference between the two?
2. A. How should the prior distribution and likelihood function for Bayesian analyses be chosen?
2.B. When and how do you think robustness checks should be performed in Bayesian analyses?
3. What do you think about using point null hypotheses versus (small) interval hypotheses when testing within the Bayesian framework?
4. How would you recommend reporting Bayesian analysis results (potentially with a robustness test)? Please provide an example text.
5. How would you recommend visualizing the results of a Bayesian analysis on diagrams?
6. How would you recommend interpreting Bayesian analysis results (with a robustness test)? Please provide an example text.
7. A. Should we use Bayesian analysis for making decisions about the evidence?
7. B. Would you recommend a decision threshold, an a priori sample size, or anything else?