Discussion points for Bayesian inference

Authors:

Balazs Aczel[1*], Rink Hoekstra[2], Andrew Gelman[3], Eric-Jan Wagenmakers[4], Irene G. Klugkist[5], Jeffrey N. Rouder[6], Joachim Vandekerckhove[6], Michael D. Lee[6], Richard D. Morey[7], Wolf Vanpaemel[8], Zoltan Dienes[9], and Don van Ravenzwaaij[2]

Affiliations:

[1]ELTE, Eötvös Loránd University, Budapest, Hungary

[2]University of Groningen, Groningen, The Netherlands

[3]Columbia University, New York, USA

[4]University of Amsterdam, Amsterdam, The Netherlands

[5]Utrecht University, Utrecht, Utrecht, The Netherlands

[6]University of California, Irvine, USA

[7]University of Cardiff, Cardiff, UK

[8]University of Leuven, Leuven, Belgium

[9]University of Sussex, Brighton, UK

*Correspondence should be sent to aczel.balazs@ppk.elte.hu

Standfirst

Why is there no consensual way of conducting Bayesian analyses? We present a summary of agreements and disagreements of the authors on several discussion points regarding Bayesian inference. We also provide a thinking guideline to assist researchers on conducting Bayesian inference in the social and behavioural sciences.

Debates among Bayesians

Despite its many advocates, Bayesian inference is currently employed by only a minority of social and behavioural scientists. One possible barrier is a lack of consensus on how best to conduct and report such analyses. Employing Bayesian methods involves making choices about prior distributions, likelihood functions, and robustness checks, as well as on how to present, visualize, and interpret the results (for a glossary of the main Bayesian statistical concepts see Box 1). Some researchers may find this wide range of choices too daunting to use Bayesian inference in their own study. This paper highlights the areas of agreement and the arguments behind disagreements, established on the back of a self-questionnaire explained in detail in the Supplement.

The overall message is that instead of following rituals[1,2], researchers should understand the reasoning behind the different positions and make their choices on a case by case basis. To assist the reader in this task, we provide a summary of our views on seven discussion points in Bayesian inference, serving as an inspiration for a 'thinking guideline' as a guide towards conducting Bayesian inference in the social and behavioural sciences.

Our paper attempts to highlight the degree of debate that persists around the topic and explains why there are no easy-to-implement heuristics on how to use Bayesian analyses. Information about the genesis of this project can be found in the Supplementary Information and on OSF (https://osf.io/6eqx5/).

--- Insert Box 1 about here ---

Discussion Points

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?

There are (mathematical) similarities between testing and estimation, although the two approaches often have different goals in practice. Bayesian testing is generally used to test whether an effect is present; in contrast estimation is used to assess the size/strength of the effect. 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. An argument to consider estimation more informative, especially when credible intervals are calculated, is that it provides information about the uncertainty of the estimated parameter(s). Bayes factors are generally considered suitable to assess evidence for or against competing hypotheses (or models). Researchers tend to use estimation when they want to examine a single model or several models very similar to each other but testing when they examine (at least two) models that differ from each other.

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

Typically, there is a lot more emphasis on the choice of prior than on the choice of likelihood in Bayesian inference, but it is just as important to use the right model -- instantiated by the likelihood function -- for the data. Some Bayesian statisticians favour subjective priors over objective/default/uninformative ones, because uninformative priors are unrealistic, or because every scientific endeavour begins with an (informed) choice of both prior and likelihood. Uninformative priors should be chosen when assessing evidence for certain parameter values, but informative priors should be chosen when assessing evidence for one model over another. When using informative priors, uninformative priors can serve a role in fitting baseline models for comparison. A slightly less wide-spread strategy is choosing priors and likelihoods iteratively, obtaining prior predictive distributions of the model, and checking whether they lead to plausible data patterns. For example, it can be valuable to choose a sceptic's prior, a believer's prior, and a personal prior, and compare the possibly diverging results to determine how much the obtained results are influenced by prior beliefs.

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

Robustness checks are performed to verify whether the obtained results are affected by for modest variations of the prior distribution but should also be used to verify the influence of the choice of the likelihood function on the obtained results. The main argument for the importance of performing robustness checks over reasonable variations in modelling choices is to increase confidence in the obtained results: ideally results should be reasonably unaffected by a researcher's idiosyncratic choice of prior or likelihood function when reasonable alternatives exist. When performing robustness checks, it is crucial to determine first which modelling choices may impact the results and perform your checks accordingly. They are primarily important when working with non-informative, and therefore more arbitrary priors.

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

First of all, it is important to consider if the research question is best served by testing rather than estimating. A researcher should consider what a practically relevant effect is before having seen the data and set up an interval test accordingly. There is some agreement regarding the practical usefulness of the point null as a model to reflect invariance, but the viewpoint is open to critique: In the end, it may not matter that much, it would be rare for a point null and a small interval around null to lead to practically different conclusions, since the point null is a useful model as an approximation of a near-zero interval. 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. Ultimately, researchers should use whichever they are most interested in (or both, to test robustness).

4. How would you recommend reporting Bayesian analysis results?

Although there is no agreement on a necessary reporting format, there are some important markers that are considered helpful in assessing the evidence. These include the model and its assumptions, prior distributions, choice of likelihood and posterior, potential hypotheses to be evaluated, details about samples from the posterior[3] when applicable, and robustness tests. It is helpful to report results in terms of competing and completely specified models. Providing figures that show estimates with uncertainty, accompanied by Bayes factors when applicable is important.

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

For Bayesian estimation, it is good practice to plot posteriors of parameters as a measure of uncertainty in case of estimation. Unless it creates an information overload, marginal predictions of a model and observed data should be plotted together, so that readers can see how authors came to their conclusions.

For Bayesian testing, plots can include information on whether the Bayes factor reaches a meaningful threshold to facilitate the reader in drawing conclusions. It may be unwise to standardize data visualization as no solution fits all purposes.

6. How would you recommend interpreting Bayesian analysis results (with a robustness test)?

There are good arguments why it may be better to focus on the scientific rather than on the statistical interpretation because it helps the reader understand what the results mean and what the uncertainties of the presented conclusions are. One helpful chain of interpretation would go from (modelling) assumptions to observed data to conclusions, possibly with a similar chain for an alternative (but plausible) set of assumptions. When interpreting Bayes factors, presenting them through the lens of betting, especially when accompanied by real-world examples of odds (i.e., Team A is deemed three times more likely to win than Team B) may be a helpful way of providing an intuition of the meaning of a Bayes factor. The same holds for providing illustrative visualizations and ranges for your qualitative conclusions when interpreting results.

7. A. Should we use Bayesian analysis for making decisions about the evidence?

One option for making decisions involves using Bayes factors. As an example, consider a researcher who obtains a Bayes factor of 10 for the hypothesis that a new medicine against migraine reduces symptoms over the hypothesis that the new medicine does not reduce symptoms. Should this Bayes factor be used to make a decision (i.e., endorse the new medication, so that it can be sold by pharmacies)?

Some Bayesian statisticians think we should, offering that Bayes factors are suitable to do so. This, however, requires reliance on related utilities as well as probabilities (see supplementary materials for a concrete example). A second option involves doing Bayesian utility analysis based on the posterior from a single fitted model. Other Bayesian statisticians state that making decisions about the evidence is optional and perhaps better left to policy makers rather than researchers. This echoes similar debates among frequentists[4].

7. B. Would you recommend a decision threshold, an a priori sample size, or anything else?

There are arguments speaking against decision thresholds, e.g., (1) the behaviour of Bayes factors for different kinds of hypotheses is insufficiently understood such 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; (2) the strength of evidence (and the number of data points) needs to be understood within the research context; (3) even the smallest study can contribute useful information; (4) basing a decision on decision thresholds alone does not incorporate utilities. One of us believes that standard decision thresholds are useful as a convention because it facilitates making a decision about the evidence (see previous question) and has been active in having journals implement them. Perhaps a compromise is to consider standard decision thresholds a useful heuristic for evaluating the statistical evidence, without using them as a basis for publishing papers.

Questions to consider

This list of discussion points shows some of the disagreement that exists on major discussion points, but also that differing opinions are supported by arguments. The bottom line, endorsed by all authors, is: Use common sense. To assist the reader in this task, we compiled a ‘thinking guideline’ (Box 2) which aims to orient the attention to the questions that should be considered when conducting Bayesian statistics.

--- Insert Box 2 about here ---

To conduct statistical inference is to make choices, for Bayesian inference, this dilemma remains. We hope that the thinking guideline that we present here is able to guide some of the choices necessary for analysing work in the behavioural and social sciences and informs researchers of some of the opinions of those in the field.

Author Contributions

B.A., R.H., and D.v.R. conceptualized the project, conducted the study survey and wrote the manuscript. A.G., E-J.W., I.G.K., J. N.R., J.V., M.D.L., R.D.M., W.V., and Z.D. contributed to the summary of this review and added suggestions to the manuscript. The authorship order follows the alphabetical order of their first names. All authors reviewed and approved the final version of the manuscript.

References

1. Gigerenzer, G. *J. Socio-Econ.* **33**, 587–606 (2004).
2. Gigerenzer, G. *Adv. Methods Pract. Psychol. Sci.* **1**, 198–218 (2018).
3. van Ravenzwaaij, D., Cassey, P. & Brown, S. D. *Psychon. Bull. Rev.* **25**, 143–154 (2018).
4. Fisher, R. *Journal of the Royal Statistical Society: Series B (Methodological)* **17**, 69–78 (1955).
5. Dienes, Z. *Understanding psychology as a science: An introduction to scientific and statistical inference*. (Palgrave Macmillan, 2008).
6. Etz, A. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 5–34 (2018).
7. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. (Academic Press, 2014).
8. Wagenmakers, E.-J. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
9. Haaf, J. M., Ly, A. & Wagenmakers, E.-J. *Nature* **567**, 461 (2019).
10. Fisher, R.A. Statistical Methods for Research Workers, 2nd Edit. *Oliver Boyd Edinb.* (1928).
11. Jeffreys, H. *Theory of Probability, section 3.23*. (Oxford: Clarendon Press, 1948).
12. Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. *The Am. Stat.* 1–14 (2019).
13. Wagenmakers, E.-J. *et al. Psychon. Bull. Rev.* **25**, 58–76 (2018).
14. Matzke, D., Boehm, U. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 77–101 (2018).
15. van Doorn, J. *et al.* The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. (2019).

Figure captions

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

Box 2. Thinking Guideline for Bayesian Inference, Questions to consider when conducting Bayesian statistics.

Competing interests

The authors declare no competing interests.

Bayes factor
The relative support provided by the data for one model over another model in the form of an odds ratio.
Bayesian estimation
Branch of Bayesian statistical inference in which (an) unknown population parameter(s) is/are estimated.
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 (density) of the data given a model for a particular (set of) parameter(s).
Likelihood function
A function of the parameters of a statistical model, given specific observed data. Consider, for instance, a coin with an unknown rate probability r of coming up heads on a single flip. For the specific data of two flips, each coming up heads {H, H}, the likelihood function of r is $L(r|H,H) = Pr(\{(H,H)\}|r) = r^2$. For instance, given these observed data, the likelihood of the specific value $r = 0.6$ is $0.6^2 = 0.36$.
Posterior (distribution)
Used in Bayesian inference to quantify an updated state of belief about some hypotheses (such as parameter values) after observing data.
Prior (distribution)
Used in Bayesian inference to quantify a state of belief about some parameter values *given a model* before having observed any data. Typically represented as a probability distribution over different states of belief.
Posterior model probability
Used in Bayesian inference to quantify an updated state of belief about the plausibility of a given model after observing data. The ratio of prior model probabilities times the Bayes factor for these same models gives the ratio of posterior model probabilities.
Prior model probability
Used in Bayesian inference to quantify a state of belief about the plausibility of a given model without taking observed data into account.
Robustness check
Used in Bayesian inference to verify the extent to which the obtained results are affected by (typically modest) variations of prior distribution and/or likelihood function.

Thinking Guideline for Bayesian Inference

Questions to consider when conducting Bayesian statistics

1. Why use Bayesian statistics?

Possible reasons include: (1) given a model, the strength of evidence only depends on data that were actually observed; (2) the results do not depend on the intention of the researcher; (3) the evidence is quantified as relative for one model or hypothesis over another model or hypothesis; and (4) the possibility to include prior information or beliefs.

For general introductions to Bayesian inference, see ref [5-8].

2. Are you interested in estimation or testing?

Conduct a test when a binary question of some kind needs to be answered (e.g., “Can people see into the future?”). In such cases, a particular parameter value, such as zero, often has a special status when testing. Estimate parameters, possibly after having conducted a test, when your main interest is about the extent of the effect (e.g., “Assuming that they can, what is their predictive accuracy?”)[9,10 p 274,11 p 385].

3. How will you choose the prior distribution and likelihood function for Bayesian analyses?

If you have relevant prior information available, for example based on prior study results, incorporate this in your prior distribution[12-15]. If not, consider using a ‘default’ (testing), or uninformative (estimation) prior. When you have several plausible candidates for your likelihood function, perform model comparisons.

4. How do you plan to demonstrate the robustness of your analysis?

Examine whether similar results would be obtained for different, but plausible, choices for the prior distribution. Perform model comparison when one has different, but plausible, choices for the likelihood function. One can couple robustness checks to decision thresholds, to verify for what range of prior assumptions a certain decision would be taken.

5. How do you plan to communicate your results?

Think about whether your results are best communicated through descriptive (summary) statistics (when the results are easily presented in the main text), graphics (when a visualisation conveys the information better), or tables (when there is too much information to present in a figure)[14]. The choice should also be guided by the research topic, the intended audience, and the type of analysis.

6. Whatever you do, at each choice and decision in your analysis, be prepared to answer the ‘why’ question!

Statistical analyses are sequences of choices. Understanding the implications of these choices and carefully thinking about them on a case by case basis are the responsibility of the author. Step-by-step guidelines and rituals can never substitute statistical thinking.