

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/126822/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Aczel, Balazs, Hoekstra, Rink, Gelman, Andrew, Wagenmakers, Eric-Jan, Klugkist, Irene G., Rouder, Jeffrey N., Vandekerckhove, Joachim, Lee, Michael D., Morey, Richard D. ORCID: <https://orcid.org/0000-0001-9220-3179>, Vanpaemel, Wolf, Dienes, Zoltan and Ravenzwaaij, Don van 2020. Discussion points for Bayesian inference. Nature Human Behaviour 4 , pp. 561-563. 10.1038/s41562-019-0807-z file

Publishers page: <http://dx.doi.org/10.1038/s41562-019-0807-z>
<<http://dx.doi.org/10.1038/s41562-019-0807-z>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies.

See

<http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Discussion points for Bayesian inference

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Authors:

Balazs Aczel^{1*}, Rink Hoekstra², Andrew Gelman³, Eric-Jan Wagenmakers⁴, Irene G. Klugkist⁵, Jeffrey N. Rouder⁶, Joachim Vandekerckhove⁶, Michael D. Lee⁶, Richard D. Morey⁷, Wolf Vanpaemel⁸, Zoltan Dienes⁹, and Don van Ravenzwaaij²

Affiliations:

¹ELTE, Eötvös Loránd University, Budapest, Hungary

²University of Groningen, Groningen, The Netherlands

³Columbia University, New York, USA

⁴University of Amsterdam, Amsterdam, The Netherlands

⁵Utrecht University, Utrecht, Utrecht, The Netherlands

⁶University of California, Irvine, USA

⁷University of Cardiff, Cardiff, UK

⁸University of Leuven, Leuven, Belgium

⁹University of Sussex, Brighton, UK

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

22 **Standfirst**

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

27

28 **Debates among Bayesians**

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

38

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

44

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

49

50 --- Insert Box 1 about here ---

51

52

53 **Discussion Points**

54

55 *1. When would you recommend using Bayesian parameter estimation and when Bayesian*
 56 *testing (i.e., Bayes factors)? Do you think there is a fundamental difference between the two?*

57 There are (mathematical) similarities between testing and estimation, although the two
 58 approaches often have different goals in practice. Bayesian testing is generally used to test
 59 whether an effect is present; in contrast estimation is used to assess the size/strength of the
 60 effect. A big difference between the two approaches lies in the nature of the (joint) prior
 61 distribution, which tends to be discontinuous for testing, but continuous for estimation. An
 62 argument to consider estimation more informative, especially when credible intervals are
 63 calculated, is that it provides information about the uncertainty of the estimated parameter(s).
 64 Bayes factors are generally considered suitable to assess evidence for or against competing

65 hypotheses (or models). Researchers tend to use estimation when they want to examine a
66 single model or several models very similar to each other but testing when they examine (at
67 least two) models that differ from each other.

68

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

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

85

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

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

98

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

101 First of all, it is important to consider if the research question is best served by testing
102 rather than estimating. A researcher should consider what a practically relevant effect is
103 before having seen the data and set up an interval test accordingly. There is some agreement
104 regarding the practical usefulness of the point null as a model to reflect invariance, but the
105 viewpoint is open to critique: In the end, it may not matter that much, it would be rare for a
106 point null and a small interval around null to lead to practically different conclusions, since
107 the point null is a useful model as an approximation of a near-zero interval. In some cases,

108 the parsimonious point null helps flag the need for more data in case a (much) more complex
109 model is believed to be true. Ultimately, researchers should use whichever they are most
110 interested in (or both, to test robustness).

111

112 *4. How would you recommend reporting Bayesian analysis results?*

113 Although there is no agreement on a necessary reporting format, there are some important
114 markers that are considered helpful in assessing the evidence. These include the model and its
115 assumptions, prior distributions, choice of likelihood and posterior, potential hypotheses to be
116 evaluated, details about samples from the posterior³ when applicable, and robustness tests. It
117 is helpful to report results in terms of competing and completely specified models. Providing
118 figures that show estimates with uncertainty, accompanied by Bayes factors when applicable
119 is important.

120

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

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

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

129

130 *6. How would you recommend interpreting Bayesian analysis results (with a robustness 131 test)?*

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

142

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

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

149 Some Bayesian statisticians think we should, offering that Bayes factors are suitable to do
150 so. This, however, requires reliance on related utilities as well as probabilities (see

151 supplementary materials for a concrete example). A second option involves doing Bayesian
152 utility analysis based on the posterior from a single fitted model. Other Bayesian statisticians
153 state that making decisions about the evidence is optional and perhaps better left to policy
154 makers rather than researchers. This echoes similar debates among frequentists⁴.

155

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

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

168

169 **Questions to consider**

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

175

176

--- Insert Box 2 about here ---

177

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

182 **Author Contributions**

183 B.A., R.H., and D.v.R. conceptualized the project, conducted the study survey and
184 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.
185 contributed to the summary of this review and added suggestions to the manuscript. The
186 authorship order follows the alphabetical order of their first names. All authors reviewed and
187 approved the final version of the manuscript.

188 **References**

- 189 1. Gigerenzer, G. *J. Socio-Econ.* **33**, 587–606 (2004).
- 190 2. Gigerenzer, G. *Adv. Methods Pract. Psychol. Sci.* **1**, 198–218 (2018).
- 191 3. van Ravenzwaaij, D., Cassey, P. & Brown, S. D. *Psychon. Bull. Rev.* **25**, 143–154 (2018).
- 192 4. Fisher, R. *Journal of the Royal Statistical Society: Series B (Methodological)* **17**, 69–78 (1955).
- 193 5. Dienes, Z. *Understanding psychology as a science: An introduction to scientific and statistical*
- 194 *inference*. (Palgrave Macmillan, 2008).
- 195 6. Etz, A. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 5–34 (2018).
- 196 7. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. (Academic
- 197 Press, 2014).
- 198 8. Wagenmakers, E.-J. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
- 199 9. Haaf, J. M., Ly, A. & Wagenmakers, E.-J. *Nature* **567**, 461 (2019).
- 200 10. Fisher, R.A. *Statistical Methods for Research Workers*, 2nd Edit. *Oliver Boyd Edinb.* (1928).
- 201 11. Jeffreys, H. *Theory of Probability, section 3.23*. (Oxford: Clarendon Press, 1948).
- 202 12. Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. *The Am. Stat.* 1–14 (2019).
- 203 13. Wagenmakers, E.-J. *et al. Psychon. Bull. Rev.* **25**, 58–76 (2018).
- 204 14. Matzke, D., Boehm, U. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 77–101 (2018).
- 205 15. van Doorn, J. *et al.* The JASP Guidelines for Conducting and Reporting a Bayesian Analysis.
- 206 (2019).

207

208 **Figure captions**

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

210

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

212 Bayesian statistics.

213

214 **Competing interests**

215 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 ref⁵⁻⁸.

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 distribution¹²⁻¹⁵. 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.