Author’s response to reviews

Title: Bayes factors for superiority, non-inferiority, and equivalence designs

Authors:
Don van Ravenzwaaij (d.van.ravenzwaaij@rug.nl)
Rei Monden (rei.monden@gmail.com)
Jorge Tendeiro (j.n.tendeiro@rug.nl)
John Ioannidis (jioannid@stanford.edu)

Version: 1 Date: 28 Nov 2018

Author’s response to reviews:

November 22, 2018

Dear Dr. Tian,

My co-authors and I would like to submit the revised manuscript “Bayes factors for superiority, non-inferiority, and equivalence designs” (BMRM-D-18-00323), for publication in BMC Medical Research Methodology. Below is a detailed description of the ways in which we have incorporated the suggestions of the reviewers. We reproduced the relevant sections of the reviews in standard font and added our response in bold. We hope that this revision is satisfactory. We look forward to your comments.

Kind regards,

Don van Ravenzwaaij
Jorge Tendeiro
Rei Monden
John Ioannidis
Reviewer 1 (Daniel F. Heitjan)

1. 5:12 — Yes, the size of the band around is often "arbitrary", but it is not intended to be so. Rather, it should reflect what somebody versed in the science considers a reasonable range of equivalence (or noninferiority, in that context). In a sense, it is your approach that is arbitrary, because it does not provide for some wiggle room around zero. Please discuss.

R1-1: Thank you for pointing this out, as it alerted us to the fact that we needed to be clearer about the fact that the equivalence Bayes factor can quantify evidence in favor of a point null hypothesis as well as evidence in favor of an interval null hypothesis, whereas the frequentist equivalence test can only do the latter. As such, you have to pick an interval, even if clear theoretical grounds for the exact bounds are lacking. We now discuss this issue more explicitly.

2. 5:16 — The quote here appears to give the analyst license to test until he finds the result he seeks. In other contexts, this would be called "p-hacking". I think the underlying issue is that you can say these things if you are willing to abandon the principle that the testing procedure, however construed, should limit the frequentist error rate under the assumption that the null is true. Just adopting Bayesian methods of analysis does not make operating characteristics go away.

R1-2: We agree that we should have elaborated on this passage. The specific issue we wished to highlight here is that NHST has essentially one decision criterion (i.e., $p<\alpha$). As such, if one employs sequential testing, every additional test increases the chance that this criterion is reached, even if the null hypothesis is true. Bayesian testing does not require a fixed $n$ in the sampling plan because the decision criterion is symmetrical. If one were to decide, for instance, to test until the relative evidence for one hypothesis over the other is at least ten, one would stop when the $BF=10$ or the $BF=0.1$. A recent publication by Jeff Rouder (2014) nicely illustrates this property using simulation studies. We now make this point in more detail.

3. 7:36 — "is itself following" should be "follows" or "itself follows".

R1-3: Agreed, now fixed.
4. 8:19 — In the first paragraph of the previous page, you mention three different pairs of null and alternative hypotheses, but you never tell us to which of those this equation refers. It is confusing.

R1-4: Agreed, now fixed.

5. 8:35 — "$\mu_\delta$ and $g$ are the mean and [SD] of the effect size prior distribution". But I thought the prior was Cauchy, which has no mean and SD.

R1-5: Conceptually, effect size $\delta$ follows a Normal distribution with a mean $\mu_\delta$ of 0 and a variance $g$, which itself follows an inverse chi square distribution with df=1. Mathematically, the combination of these two priors is equivalent to placing a Cauchy prior (with scale parameter 1) on the effect size distribution. The equation allows modification of these original mean and variance parameters for direct calculation of the Bayes factor for additional flexibility in the choice of prior. For instance, choosing a different mean allows the calculation of non-inferiority Bayes factors further down. We modified the manuscript to clarify this.

6. 8:39 — "$r$ denotes the scale parameter of the Cauchy distribution". How is this different from the prior that you just mentioned? Please explain more clearly what this equation represents and what all its components mean.

R1-6: We have clarified this. Jeffrey’s original prior was a Cauchy with location parameter 0 and scale parameter 1, the more recent work allows shifting the location (as discussed above) and scale (for fields where large effect sizes are more or less common). We have added clarification after presentation of the equation.

7. 14:34 — "principle", not "principal"

R1-7: Agreed, now fixed.
8. 21:52 — You should indicate that you are citing the second edition of the Senn book.

R1-8: Agreed, now fixed.

9. When I read the abstract, I got the idea that you would put a normal prior on the $\delta$ parameter, which — for fixed $c$ — would make the calculation of posterior probabilities (or BF) sensible for noninferiority and equivalence hypotheses as typically construed. I do not understand how you can get positive posterior probability for an alternative hypothesis of $\delta=0$, a set of measure zero. Is this a feature of the Cauchy prior? If so, it seems pathological, and not the sort of foundation on which one would wish to build a method. Please explain how your approach would compare with other approaches like the one I just outlined.

R1-9: We believe your question is about how one can calculate a Bayes factor for one hypothesis over the other when one of the two is a point hypothesis. In case of competing hypotheses a Bayes factor is defined as

$$(\frac{p(H1|data)}{p(H1)}) / (\frac{p(H0|data)}{p(H0)})$$

Now let $H0$ be that the population effect size falls in an interval around zero (say $-c<\delta<c$). The smaller one chooses $c$ (and therefore the interval around zero), the more $p(H0|data)/p(H0)$ will dominate in the calculation of the Bayes factor, as $p(H1|data)/p(H1)$ will tend to 1. In the limit of a point null hypothesis, one can get the Bayes factor by $p(H0)/p(H0|data)$, or by evaluating the ratio of the density of the prior and the posterior, evaluated at $\delta=0$. This way of calculating the Bayes factor for point null hypotheses is known as the Savage Dickey ratio, mathematical proof is presented in O’Hagan & Forster (full reference in the manuscript). We have added a paragraph to the revised manuscript to clarify this issue.

Whether or not a point null hypothesis makes sense from a philosophical perspective is another matter. In this manuscript, we provide the tools for calculating Bayes factors for both point (null) and interval (null) hypotheses, so that the reader can decide for themselves.
Reviewer 2

1.) Page 4 of the manuscript states that the p-value is underpowered. This is potentially misleading. If a test is underpowered, we cannot infer that the test results would be significant if we collected more data. Whether the test will be significant, or not, depends on what the data say. And what the data say is the empirical matter we have yet to find out.

See references: Krefeld, Schwalb, Witte, and Zenker (2017). Hypothesis testing needs trustworthy data. Frontiers in Psychology (open access)

and


R2-1: We agree that we could have reworded this more precisely, we meant to indicate that non-significant results can follow from the null hypothesis being true or from the null hypothesis being false, but the test being underpowered. We have clarified the segment.

2.) Page 5. "entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" - This is a strong statement, and particularly the second disjunction --the "run out of resources"-part--is in need of justification. (Notice that nobody would disagree with the first disjunct, to begin with!). I suggest to add a justification why this is appropriate, and here state especially what is meant by 'appropriate'.

R2-2: Agreed, we have elaborated on this passage (see also R1-2).

3.) Page 5 'For a given statistical model, say M, the prior distribution or prior p(θ|M) for a parameter θ is updated after encountering data y to yield a posterior distribution or posterior p(θ|y,M) - in the following, since you always conditionalize on M, you might state M once, then
mention that you so conditionalize, and subsequently therefore leave M implicit. It would increase the readability of your equations.

R2-3: We prefer to leave Equation 1 in with the quantities conditioned on M to be maximally precise. Equations 2 and 3 just contain the hypotheses and data.

4.) Bayes rule and the Bayesian theorem (which you state on page 5) are not exactly the same. The rule states that nothing but the prior and the likelihood determine the posterior; the theorem make this calculable. The distinction, admittedly, is subtle, but is is there!

R2-4: Agreed, now fixed.

5.) Page 5 the authors noted a comparison - suggest re-wording. between why ""compromise between""? how about ""mathematical product of""?

R2-5: Agreed, now fixed.

6.) figure 3 - If you wrote L here (for likelihood) it would be clearer that L is distinct from a probability, for a likelihood is a probability that is multiplied by a positive constant (i.e., by the prior).

again, see Witte & Zenker (2017).

R2-6: We were unsure whether the reviewer refers to the absent axis labels or to “BF” presented in text. We have added x-axis (Parameter) and y-axis (Density) labels, the “BF” text labels seem appropriate and consistent with the surrounding text, hopefully this addresses the reviewer’s suggestion to improve clarity.

7.) page 6 BF10 or BF01 (with 10, or 01, as subscripts) is perhaps the more common notation.
R2-7: Changed as suggested.

minor proof-reading work needed

R2-8: We have carefully proof-read the paper one more time and corrected several typos.