Thank you for your thorough revisions. As you can see, many issues have been resolved and several of the reviewers are satisfied.
Nevertheless, there remain issues related to the analysis that require further consideration. That is, the setting of the prior distribution and the inferences that can be drawn from it. Please read the reviewer's comment carefully for more details. I agree that this issue should be resolved before granting IPA. I look forward to seeing the revised manuscript again.

Yuki Yamada

Thanks again to all reviewers and the editor. We've learned a lot in the journey. Now we addressed Dr. Dienes' comments by providing simulations and revised the protocol as well.

Reviews

Reviewed by Zoltan Dienes, 20 Jan 2023 22:49

The planned analyses are now more clearly presented. But a main issue I raised has not been dealt with. The authors are using uniform priors. Such default priors do not typically reflect what a plausible theory would predict; which means no plausible theory is typically being tested. The way to answer the concern is to show that in this situation the default uniform is reasonable. How can one show that? One way is to indicate how the BF performs on imaginary data showing plausible ways H1 may be true. For example, give a sex ratio that deviates from the H0 by a relatively small amount - what is the smallest amount for which the BF still gives good evidence for H1? Age is more complex as it is multi-df; but one may proceed as the authors have by assuming age is normally distributed and find the smallest difference in means for which one just gets evidence for H1. (If one assumes normality one could argue one should use Bayesian t-tests for age. But one could also treat normality as just one possibility for checking how the test behaves.) This approach would be the simplest. More thorough would be showing what population effect would lead to say a 80% chance of being detected. Conversely one should show that if there is no effect, there is sufficient N to obtain evidence for H0. This is of course unlikely to be a problem with the planned sample size, but one should always show this in a registered report (the logic of planning for a severe test is given here: https://psyarxiv.com/yc7s5/).

Response 1: Thanks very much for this brilliant suggestion. We followed the suggested approach and conducted simulations to address the concern.

For the Bayes factor analysis of sex ratio, we conducted the following simulation (Which is also described between line 170 and line 380 in R Notebook at: https://gitee.com/hcp4715/chin-subj/blob/master/Notebook_Data_Analysis_CHN_Sample_Stage1_RR.rmd):

Step 1: Generate 1000 values using `rbinom(1000, size = 100, prob)`. `prob` here is the parameter for the binomial distribution (see below). The generated data represent male's counts (n_{male}). The female's count is then calculated as $n_{female} = 100 - n_{male}$.

Step 2: Calculate the Bayes factor using Bayesian multinomial test, with non-informative prior and each pair of the generated data as observed and 50/50 as the expected. This step results in 1000 BF values.

Step 3: Calculate the proportion of BF values that are greater than 3.

We iterated the above three steps for binomial parameter values ranging from 0.5 to 0.7, with step size of 0.01. The results revealed that, with BF >= 3 as the criterion, the non-informative prior can detect a deviation of 0.17 from 0.5 at 87% of the cases. The greater the deviation from 0.5, the greater proportion of BF values is greater than 3.

When the parameter of binomial is 0.5, i.e., the null effect, we obtained evidence for null effect (BF_{01} >= 3) for 85.9% of the cases. This suggests that the current setting provides both evidence for null effect and for a medium deviation from the null value.

As for the age distribution, we simulated the case of five age bins. Because the probability distribution is multinomial, quantifying the effect size is difficult. We simplified problem by generating multinomial parameter values from Dirichlet distribution. More specifically, we used the following steps:

Step 1: Generate 10, 000 probability vectors from a Dirichlet distribution with the alpha parameter as (1, 1, 1, 1, 1). This alpha value is chosen because it generates a uniformly distributed probability vectors. Code: `gtools::rdirichlet(n=10000, alpha = c(1, 1, 1, 1, 1))`

Step 2: Generate 5000 multinomial data for each probability vector from Step 1. The generated data will be the observed data. Code: `rmultinom(5000, size=100, prob)`

Step 3: Calculate the Bayes factor by comparing each observed (from Step 2) to the expected (we used equal proportion [20, 20, 20, 20, 20] for simplicity), resulting 1000 BF_{10} values for each probability vector.

Step 4: Calculate the proportions of BF_{10} values that are greater than 3 for each probability vector. This proportion is similar to the "statistical power" in Frequentists' statistics. Iterating through all probability vectors from step 1, we get 10, 000 proportions.

Step 5: Calculate the percentage of the proportions in Step 4 that are greater than 0.8.

The above simulation revealed that 93.8% of the probability vectors were generated from Dirichlet distribution. These results suggested that our Bayesian multinomial test can provide evidence that the probability vectors of interest are different from the null for 80% of the cases.

We also calculated how the current Bayes factor analysis can provide evidence for null effect, in this case, multinomial parameters equal to [0.2, 0.2, 0.2, 0.2, 0.2]. We found that in 1000 simulations, we get evidence for null effect for 96.8 of the cases when using BF_{10} >= 3 as the criterion. This suggests BF is sensitive to support the null.

To conclude, our simulation can be summarized in the following table, which suggests that non-informative prior can be used for the current purpose. We also appreciate any input from the reviewer regarding our simulation.

Situation	Evidence for null (BF_{01} >= 3)	Evidence for alternativeBF_{10} >= 3
Sex ratio	85.9%	>= 87% if deviation >= 0.17
Age bins	96.8%	93.8%

The authors responded to my point by saying they will check robustness with a different prior. This suggestion in itself leaves open inferential flexibility: What conclusions will be reached if the different priors lead to different conclusions? So the authors need to be clear on what basis they will draw particular conclusions. Simplest would be to justify one prior as most suitable (e.g. because of the way it behaves as described above), indicate all conclusions will be wrt this prior, and the other is for background information only.

Response 2: We now only used the non-informative prior and added justified for it.

Reviewed by Kai Hiraishi, 13 Jan 2023 02:47

Thank you for revising the manuscript. I am very pleased with the author's response and the revisions. I also believe that the communication between the author and the other two reviewers has greatly improved the manuscript (and greatly helped me understand the planned analysis and interpretation of the results). I would like to recommend this protocol to receive the IPA. I am looking forward to seeing the 2nd stage manuscript.

Sincerely,

Kai Hiraish

Response: Many thanks for your comments, they helped us a lot.

Reviewed by Patrick Forscher, 04 Jan 2023 13:47

I have read the revised manuscript and the authors' response to reviewers. I was largely satisfied with the previous draft of the manuscript and I'm also satisfied with this one. I'd like to see this manuscript accepted so that I can read the authors' results. :)

This is a great project -- one I'll certainly be keeping an eye on!

Patrick S. Forscher

Associate Director

Busara Center for Behavioral Economics

Response: Many thanks for your comments, we also looking forward to seeing the results.