|  |  |  |  |
| --- | --- | --- | --- |
|  | To Do | How To Do | Done |
| As pointed out by Reviewer 2, OMM is hardly a well established statistical method. For a relatively new method, your review of this method fails to show its connection to and superiority over existing similar methods such as bootstrap and fails to convey enough details for the audience to gain working knowledge about OMM beyond a specific design. |  | Should we expand this? Don’t we already send them to other papers as a reference to clarify methods and procedures? |  |
| Both Reviewers 1 and 3 point out that in your review of NHST specific assumptions tied to a specific model or procedure are counted as negative features of the broader NHST framework. This is inadequate. Both frequentist and Bayesian frameworks may invoke certain assumptions on the distribution of the data, and both of them have nonparametric or robust procedures that do not make these assumptions. | Yeah, clarify difference between issues with NHST and assumptions of ANOVA |  |  |
| **Reviewer 1(Grice)** | (I basically agree with everything he says…) |  |  |
| Second, the *American Statistical Association* just held a conference in Bethesda, Maryland (http://ww2.amstat.org/meetings/ssi/2017/index.cfm) on significance testing in which the main theme was getting beyond NHST. | Include this tidbit in the intro and discussion |  |  |
| My primary concern is with the paper's overall organization, which fails to present the authors' arguments in the clearest manner possible and which also leads to a number of specific difficulties. As an example of a specific difficulty, on page 5 the authors write "Given the mathematical constraints associated with the current NHST procedure as delineated above, there are a number of assumptions that must be met..." The problem I'm having with this statement is that the “current NHST procedure” is not clearly identified at this point in the manuscript. The “mathematical constraints” and assumptions listed on page six are tied more directly to the statistical analysis to be conducted in the simulations (viz., a repeated measures ANOVA) than to NHST. As the authors point out, one can violate the assumptions and switch to a nonparametric test which can still be framed in NHST terms (e.g., computing a p-value for a chi-square test). Moreover, as the authors point out, different statistical tests have different assumptions, but this fact is separable from NHST and its accompanying philosophy. The problem here is that, on the one hand, NHST is a generic way of framing hypotheses. On the other hand, the hypotheses in NHST are population parameters, and the researcher uses specific statistical tests that come packaged with their own unique assumptions to choose between the hypotheses. Overall, then, I would encourage the authors to more clearly distinguish between NHST and the specific ANOVA they’ll be using to analyze the simulated data. | clarify limitations of NHST as separate from the assumptions necessary for ANOVA |  |  |
| To be more specific, I recommend a slight re-organizing of the manuscript as follows: |  |  |  |
| 1. As part of the introduction, point out that the three approaches will be compared via simulations using a repeated measures design with a Likert-type scale as the outcome variable. This will allow the authors to then present the specific hypotheses for each type of analysis and to also discuss their similarities and differences more generally. For instance, the section discussing the history of NHST would remain unchanged, but the section “Current NHST Procedure” would present the explicit null and alternative hypotheses: H0 : µ1 = µ2 = µ3 ; HA : µ1 ≠ µ2 ≠ µ3. Being concrete here, the authors can then reconsider their three bullet points on page 5 | I’m good with this. |  |  |
| In the first bullet point, for instance, given the null and alternative hypotheses here, what exactly is meant by “Ideally these should be hypotheses that allow for a meaningful conclusion to be reached regardless of whether the null hypothesis is rejected or not”? | Ugh. This was meant to be a discussion of the method generally where people can do dumb tests like not=0, or they can pit theories against eachother. I guess this is fodder for a different paper and so I can drop this and alter accordingly. |  |  |
| In the second bullet point, should “omnibus p-value” be replaced with “omnibus critical p-value”? Are the authors here thinking of corrections for Type I error inflation? Lastly, in bullet point #3, the explicit hypotheses make it clear that the conclusions reached in NHST involve population means. The word “means” should therefore be preceded by the word “population” (viz., “population means”) in the third bullet point. Two other points regarding NHST: | Agree. |  |  |
| a. P. 13, top. Specific to the null hypothesis, the authors state “The null hypothesis states that there are no significant differences between groups, and…” As shown above, the null hypothesis explicitly states the three *population* means are exactly equal. Also, “…but does not specify where using the F distribution focusing on p-values” is not clear. | Not sure what this is talking about…? |  |  |
| b. To be clear on page 13, the authors ran two pairwise comparisons but divided the critical p-value by three. If this is so, perhaps it can be stated more plainly? | I thought this was pretty clear, should we just say we only *discuss* 2 comparisons, but ran and corrected for all 3 because that’s what psych people are trained to do? |  |  |
| 2. The revised Bayesian section can again include a review of how one approaches data from this perspective, and the “Current Procedure” section can then be written to address the repeated measures data with three levels. Are the hypotheses written in the same manner as NHST? Are they different? If so, how? In NHST population parameters are considered fixed. Is this the case with Bayesian analysis? Is random sampling required in Bayesian analysis? I believe that it is required, as it is in NHST. What other assumptions are required? | Write out actual hypotheses with equations, include more of this type of information. |  |  |
| 3. The revised OOM section can include a review like those for NHST and Bayes, and the “Current Procedure” can address the same questions posed above (e.g., Does OOM rely on the estimation of population parameters and random sampling?). Again, my belief is that by setting up the particular design in the beginning (viz., repeated measures with three levels), the authors can address both specific and general differences among the three approaches. Being familiar with OOM, I also have a number of specific comments regarding these sections: | Cool. Do all the things below, since, you know, he made this thing. |  |  |
| a. Having “James W. Grice” written out is awkward, and should be replaced with simply “Grice” throughout. | ok |  |  |
| b. P. 9. What is meant by “…all theories have an underlying truth”? | This is like straight from the summarization of philosophies, but I can rejig so it makes more sense. |  |  |
| c. P. 9-10. “…in terms of forming structures and processes for phenomena.” I recommend rewriting as “…in terms of structures and processes in order to explain phenomena.” | ok |  | JS |
| d. P. 10. “…focus on population parameters and their various underlying assumptions.” Consistent with my points above, perhaps rewrite as “…focus on population parameters and the various assumptions underlying statistical tests (e.g., random sampling, normality, homogeneity of population treatment differences, etc.).” | ok |  | JS |
| e. P. 10. “…to take a step back and to focus on observations at the level of the individual.” I’m not sure how this is a “step back” but it is clear that OOM is person-centered. | clarify |  | JS |
| f. P. 10. PCC stands for Percent Correct Classification rather than “Percent Correct Matches.” | ok |  | JS |
| g. P. 10. The statement “Complete matches are the proportion of observations that match the researcher-designated patter on all dimensions” is not clear without an example. For the 3-level OPA in this paper, a person would be tallied as a “complete match” if the ordinal pattern of his/her data matched the expected ordinal pattern across all three levels. | ok |  | JS |
| h. P. 10. “…by randomizing observations a researcher set number of times…” is awkwardly stated. In simplest terms, the c-value is a type of randomization test, and the researcher can determine the number of randomized trials for the test (e.g., 1000 or 5000 randomized versions of the actual observations). | ok |  | JS |
| i. P. 11. “…are indicative of distinct observations that are not likely due to chance.”  Perhaps rewrite as “…indicate a pattern of observations that is improbable (i.e., unlikely produced by chance) when compared to randomized versions of the same data.” | ok |  | JS |
| j. Page 11, under Current Procedure, the authors should discuss more details of the OPA, particularly the idea of complete matches and exactly what a PCC value would indicate from the analysis. In order to highlight the person-centered and strictly ordinal nature of the OPA, example responses from several cases could be provided and shown how they match or fail to match the hypothesized ordinal pattern. | ok |  |  |
| k. As alluded to above, like NHST and Bayesian approaches, the authors can also discuss what the hypotheses are for OOM. This might be tricky, but the essential point is that population parameters are not necessarily inferred from the analysis. Instead, through OOM the researcher is normally seeking an inference to causal structures (as mentioned by the authors). In this case, the expectation is that the causal mechanism is operating at the individual level rather than the aggregate level. The main point of the analysis, then, is to see how many people fit the expected pattern which is based on a causal theory. If all causes are accounted for in the study and observations have been made with sufficient precision and accuracy, then 100% of the persons should fit the expected pattern; otherwise, a lower PCC value will be expected and it is up to the researcher to determine how high a PCC must be (e.g., is 75% high enough?) to support an inference to the causal mechanism.  The philosophical and technical differences between modern Bayesian statistics and OOM are vast given their complexity, but I am hopeful that by placing focus on the repeated measures design with three levels up front, the authors can demonstrate and discuss some of the essential differences more clearly. | ok |  |  |
| Simulations and Discussion   * 1. 1. I found the presentation of the simulations to be clear and their design to be sensible. I appreciate the authors’ using whole numbers ranging from 1 to 7 from a Likert-type scale as such data are most common in psychology. Rarely do psychologists work with truly continuous quantities. | cool |  |  |
| * 1. 2. The graphs are too crowded both vertically and horizontally. Horizontally, in particular, it is difficult to see the differences between sample sizes of 10, 30, and 100. I recommend reformatting the graphs to show the differences more clearly. | Fix it |  |  |
| * 1. 3. The label for Figure 2 needs to be clear. The word “significant” means different things for the three analyses. The current label implies that Bayes and OOM analyses used p < .05 as well. | Fix it |  |  |
| * 1. 4. It would be interesting to see graphs of sample mean effect sizes (d), Bayes factors, and PCC indices. I am particularly curious to examine how the BFs and PCCs compare across the different effect sizes. The cut-point for “significant” PCCs is low (> .50), and the c-value from a randomization test will be small for large sample sizes, like a traditional p-value. If the PCCs for the large effects are moderately large (e.g., 75%), that would put them more in line with the Bayesian results. This is important for OOM because PCCs are the focus of the results, not the c-value nor an arbitrary cut-point (as the authors note previously in the manuscript). In the larger context of OOM, the PCCs are only meaningful in the context of an observed or predicted pattern, the latter of which would ideally be understood through a causal (integrated) model. | Use some of this language if necessary, and make some cute graphs. Of like, PCC/BF scatter, c/p scatter (ick, do we want to even put that in people’s heads? Or maybe we do because that would be akin to saying unless we have a specific d value we won’t care about p………)  I think this would be worth looking at regardless of whether it’s necessary in the paper or not…..but now I’m thinking this idea of look at effect size first and significance second is important……. |  |  |
| * 1. 5. I would like to see a discussion of effect size included in the Discussion section. I understand this will be tricky given the generic 1 – 7 rating scale, but the difference between the first and third population means is only 1 point (2.5 vs. 3.5). Would a “typical” researcher find such a difference psychologically meaningful? How do the PCC indices, as effect sizes, compare? As the authors note, the OPA ignores differences in magnitude, looking only at ordinal differences. What does this mean in terms of measurement and how we think about a 1 – 7 scale? | Would they find it meaningful—yes, definitely  How to pcc and d react?—that’s what graphs above will tell us | We should think on this las point. I’m unsure (but also currently tired) |  |
| * 1. 6. Lastly, perhaps a more appropriate title would be: Beyond p-values: Utilizing Multiple Methods to Evaluate Evidence. | I’m ok with this. |  | JS |
| **Reviewer 2: I will be you $20 this is the NHST person** |  |  |  |
| The historical overview of NHST and Bayesian methods are rather long, however. In my view, OOM does not represent a new testing methodology but instead represents a rather awkward attempt to combine ideas from philosophy and the widely used statistical technique called "the bootstrap" (the proponents of OOM are apparently unaware of the vast literature on the bootstrap or its obvious connection to OOM). I do not believe OOM has been accepted or even adequately peer-reviewed in the psychology literature (it has appeared primarily in monographs by its authors); thus, the rather cursory statistical analyses of its properties and comparison to standard testing procedures weakens the current manuscript. The review of these methodologies comprise 11 out of the 17 pages of the manuscript and does not contribute new results. | Based on Grice’s comments and my own thoughts, I think the best course of action is to include a few sentences in our description of OOM that removes any wording we already have regarding bootstrapping (I think I wrote the c test was akin to boostrapping somewhere), and give a more thorough explanation of why it is not bootstrapping and why this difference is important and valid. |  |  |
| Setting aside the unnecessary coverage of OOM, I found the simulation studies and their description to be largely uninformative |  | How can we address this? Do we need more studies? Should we include the null hyp study we did? |  |
| The major problems associated with NHST focus on the connection, or lack thereof, between the probability that a null hypothesis is true and evidence provided by p-values. The binning procedures used by the authors almost completely obscure this relationship by combining all experiments with p-values less than 0.05 or 0.005, making it nearly impossible to disentangle the evidence provided by specific p-values within each range. | Talk about the purpose of binning is to simulate the decision criteria that one might use given current trends. Include in intro to begin with. Other comparisons are not the purpose of the paper. This is more about the actual decisions that a researcher would make and how often those would agree.  Pg. 11-sim, given these things, the goals are x y z blah about real researchers making decisions based on current criteria | Not sure how to deal with this one either. I mean we could combine Grice’s discussion of d above with our p values/bf/whatever and graph these things continuously, but these things are done ALL THE TIME, and these relationships are well known. I’m not sure if we just need to make our hypotheses more clear, and that we are comparing malt pole methodology is and other evidence. Or if we need to completely discuss east P values as continuous. If we did so, what other continuous variables can map them on to that would be worthwhile in novel |  |
| There was no meaningful discussion of false positive proportions, false discovery rates, false negative proportions, etc., or the impact that the prior probabilities of hypotheses have on these rates. | This makes me think we should be including the null simulation in here, then we could code and discuss real false positives, (in the case of the null or negligible effect), as well as real false negatives (in the sense of the small-large effects)  Put as a limitation of the study—d not really 0, but blah. | … “or the impact that the prior probabilities of hypotheses have on these rates” I’m not 100% clear on this, but if they are referring to the Bayes priors there have already been studies on this that (I’m pretty sure) we reference in the paper. |  |
| The presentation of figures was poor (details in the figures are almost not discernible). | Fix this—mentioned throughout |  |  |
| Discussion of the simulation results and recommendations were unfocused and conclusions drawn were not adequately supported by theory or simulation. | I think if we can clarify the hypotheses as Grice suggests and include the null setting, we can get around this one decently well. |  |  |
| I was unable to identify either any new insights regarding the shortfalls of NHST or Bayesian hypothesis testing procedures, or any novel solutions to known problems. | [insert eye-roll emoji here] |  |  |
| The assertion that changing the p-value threshold from 0.05 to 0.005 (whether one agrees with this proposal or not) is prima facie not correct. By definition such an effect has a ten-fold impact on type 1 error; the ramification of its effects on the overall efficiency of the science is a more complicated problem that depends on a variety of factors, but the choice of evidence thresholds certainly plays a central role. |  | We say this…is there a way to make that clearer? |  |
| **Reviewer 3: I will also bet you this is the Bayes person** |  |  |  |
| The authors present a simulation study that aims to compare three methodologies (frequentist inference NHST, Bayes Factor, and Observation Oriented Modeling), whether they differ (or not) in the strength of their conclusion to support the null hypothesis. The conducted simulations are based on a repeated measure ANOVA with three levels. **The manuscript deals with an actual and important topic in empirical research, especially since Benjamin and colleagues (2017) proposed to change the default p-value threshold for statistical significance from 0.05 to 0.005.** The results of the simulations show that the two significance levels p = .05 or p = .005 lead to very similar conclusions, even for small effect sizes. However, this manuscript has some weaknesses and I think there is considerable potential for improvement. To this end, below I list some questions and recommendations for some changes in no particular order. | Just needed some positive feedback after reading the last reviewer. You may proceed. |  |  |
| 1.) Overall, the manuscript has great need to improve accurate/precise scientific writing. For example, what is an "omnibus p-value"(line 104); or line 94: what exactly is meant by "current" in "Current NHST Procedure"? Did it change, or will it change in the next few years?; line 228: PCC = Percent Correct Classification not "percent complete match". | Easy fixes, but should re-read entire document for these types of shortfalls. |  |  |
| 2.) I miss the (section) "Aims of the Study" or "Research Questions". Please, make your research questions explicit. Sum up the previous section, state the known problems, and derive your questions logically from the theoretical background. Please, prioritize your questions that are central to your purpose, and others that are secondary. | Makes sense. |  |  |
| 3.) NHST "…assumptions that must be met…" (line 109ff): In my opinion, some statements are not correct. a) "normal sampling distribution" is not an assumption for NHST, e.g. the difference of two means is t-distributed, F-tests, or chi-square tests etc. What is necessary is to know the distribution of the test statistics under the null hypothesis. Then you can calculate the probability (p-value) for this test statistic ("Compute your given analysis" in line 105). Tabachnick and Fidell (2012) describe multivariate analyses in which multivariate normality is typically assumed. Linearity is an assumption for linear models, but is not related to NHST. NHST is used for General Linear Models, as well as for Generalized Linear Models. It seems that assumptions are mentioned at this point that are relevant only for the repeated measure ANOVA of the simulation study. | Yeah, as mentioned above just need to clarify the issues with NHST vs. assumptions with ANOVA thing. |  |  |
| 4.) Why (only) a repeated measure ANOVA (with three levels)? What is the reason for this design? | I’m pretty sure we already say why—because it’s so widely used in pysch—but should beef up this sentence/rationale |  | JS |
| 5.) Line 181-182: Why is missing data a problem for Bayes statistics? Missing data analyses are mostly based on Bayes statistics because it can handle missing data. | I can clarify here, it’s an issue with the function we used, not an issue with the statistic. |  |  |
| 6.) In Figures 2-5 "Significant" on the vertical axis: What exactly is meant by "Significant"? I'm not sure if this term is precise enough because it is not the same for the three methods. | Re-name axis |  |  |
| 7.) If I understand correctly, p-values, Bayesian factors, and PCC and c-values have been divided into three classes in order to rank the strength of the evidence for the null hypothesis? This should be more clearly illustrated in Figures 2-5. When you created three classes, why did you show only two of them? Maybe you can explain why these results are not worth being presented? | It might be easier to show all 3, just so this doesn’t keep coming up, as I know we said in the paper that they sum to 100%, but apparently this isn’t getting through. |  |  |
| 8.) Line 369ff: "Percent Agreement" "A goal of this project was to …" In the result section, I read for the first time that this was a goal of the study. This goal should already be included in the questions and its meaning should be explained in a comprehensible way. | Yeah. Put together an aims section. |  |  |
| 9.) The Figures 2-5 are currently too small and too crowded. The vertical axis is a bit too short. On the horizontal axis, the data points for N = 10 and N = 30 are indistinguishable, and difficult to distinguish for N = 30 and N = 100. A non-linear scaling of the horizontal axis would be a way to make the Figures clearer. | Fix this. |  |  |
| 10.) Please, acknowledge the limitations (and the strengths!) of your research. This is a very important part. For example a) only repeated measure ANOVA with 3 levels was investigated, b) are the classifications of p-values, Bayesian factors, and PCC appropriate? c) Are the results generalizable or only partially generalizable? How do violations of statistical assumptions affect the generalizability of the results? | Do this. |  |  |