Amy Willis
adw96 [at] cornell [dot] edu
1173 Comstock Hall, Cornell University
Ithaca NY 14853 USA

September 17, 2016

Professor Nigel Stallard
Editor
Journal of the Royal Statistical Society

Dear Professor,

On behalf of myself and my coauthors of JRSS-OA-SC-May-16-0104 I would like to thank the editorial team of JRSS-C for their consideration of our manuscript. We were delighted to receive such thorough reviews and that both referees were complimentary about the paper, considering the problem to be important and our solution to be an "exciting contribution." We are thrilled that JRSS-C is considering a revision.

The reviewer and editors outlined a number of specific points for improvement. Without exception, we have concurred with their recommendations and modified the manuscript to incorporate their suggestions. Detailed responses to the Associate Editor's and reviewers' comments is attached to this letter, and the revised manuscript is also enclosed.

We hope that the editorial team finds our revisions to be satisfactory and once again thank the editorial team and both reviewers for their feedback and excellent critique.

Kind regards,

Amy Willis

Referee 1

- *The paper addresses real-life problems as well as introduces novel and for this reason it is suitable for this journal. The problem is well presented and the paper introduces a great and exciting contribution to this common issue.*

Thank you very much for your thorough review of the paper! We are thrilled that you consider the work to be an exciting contribution.

- *The comparison of methods for the virtual simulations is a good approach when presenting new work though it seems the authors could have been more explicit with the real data analyses, perhaps with a few figures showing how the new approach compares to other approaches to further validate its relevance and advantages. The use of tests with both virtual and real life datasets is also a good approach.*

I'm very glad that you deemed our balance of simulation and real data to be appropriate. I absolutely agree that we should further compare our approach to the existing method of linear regression on the observed richness. With respect to the gut microbiome dataset, we have included the results of a mixed effects model in Section 5: "In contrast, a mixed effects linear regression on the observed richness suggests that only 351 species are lost due to the antibiotic, and that a difference can be concluded with far more confidence ($p = 0.0004$). This highlights that failing to account for uncertainty in species richness estimation leads to overstated confidence in tests for covariate influences..." We have also included this information in the caption for Figure 2.

Part of the motivation for this paper was that no homogeneity test for species richness existed, and thus there are no other approaches with which to compare for the soil dataset analysis. To clarify this, which was also suggested by Reviewer 2, Section 4 now contains the sentence "Note that if only observed richness is considered, the samples appear to have different richnesses (Figure 1), though no inferential method for determining this was previously available."

We were instructed by the editor not to increase the length of the manuscript, hence we incorporated your suggestion into the text rather than with additional pictures. We hope that you find this reasonable. We think that the above inclusions significantly clarify the importance of our method and are very grateful for the suggestion.

- *I find the overall structure of the paper a bit patchy. This is mostly due to the introduction of Fig. 1 at the initial parts of the paper only to be explored in more detailed later on together with the other dataset. I feel some careful thought on how to better present the paper's overall argument in a manner that flows more seamlessly is required.*

On re-evaluation, the introduction of Figure 1 early in the manuscript and without a proper discussion of the data seriously compromises the flow of the paper. We have removed the incomplete discussion in Section 2, opting to explore it more thoroughly in Section 5. We hope that you think this restructuring, along with the clarifications you requested below, improves the organisation of the paper.

- *The paper fulfils the intentions laid out in the introduction and the conclusions are well validated and conservative. The use of previous datasets is well accomplished and the paper has practical importance. There is so awkward phrasing and some structural shortcomings. The paper has sufficient text and the graphs are informative and visually appealing.*

Thank you for the very constructive feedback! We have revised the specific phrases highlighted below. Furthermore, we agree that the introduction ended abruptly, and for this reason have now included a brief overview of the remainder of the paper to better connect Section 1 to the following sections and provide the reader with a roadmap for the manuscript.

- *Line 56-61 awkward phrasing*

Agreed. The relevant lines now read "A key advantage of the method is that it permits comparison across different sample sizes. Furthermore, rigourous inference regarding the effects of covariates on biodiversity made possible, and adjustments for simultaneous inferences arise naturally. Finally, it provides the first inferential method for assessing homogeneity of samples with respect to total biodiversity."

- *Line 68-71 awkward phrasing*

Agreed; there was a missing word in the original submission. It now reads "We conclude that a highly significant decrease in species richness of the human gut occurs in response to an antibiotic (concurring with the original paper), but observe a post-treatment recovery of the richness of the ecosystem (in opposition with the original paper)."

- *Line 95 There's a typo*

Thank you! The word "attributed" is now spelt correctly.

- *Caption in Fig. 1 Check grammar. Also wondering if it would be possible to see how the same dataset is estimated with other methods, in order to make a more convincing argument that the method presented in this paper is comparable if not superior. Same applies to Fig. 2*

We were following the biological sciences protocol of a passive caption. However, in accordance with your suggestion we have changed it to first person plural: "Here we observe estimates..." rather than "Comparison of estimates..." The same alteration was made to Figure 2. We similarly concur with the suggestion regarding comparison with other methods, which was discussed on the previous page.

- *Line 297-301 Is there any results supporting this argument?*

Yes! However, due to space constraints we omitted the supporting table from the manuscript. This is a very reasonable request however, and to accommodate it and justify the claim we have created a short Supplementary Appendix to house a table of Type 1 error rates under the reversal of continuous and discrete with medium and high diversity datasets. The table (Table 5) clearly shows that the error rates are substantially lower for the high diversity (Whitman *et al.*) dataset, regardless of whether the covariate is discrete or continuous. We hope that you find the explicit inclusion of these results more satisfactory than the previous allusion to them. We have also uploaded the script used to generate this table as a Supplementary File.

Thank you again very much for your comprehensive and thoughtful review. All of your suggestions were very reasonable and we believe that they substantially improved the structure and clarity of the paper. We have added a note of our gratitude at the end of the manuscript.

Referee 2

This well-written paper deals with a substantive and interesting problem in ecology. The applied problem is well-stated, and the methodology that is proposed seems to be reasonable and statistically defensible. The authors illustrate the use of the proposed approach clearly through the use of an extensive simulation study and through two real-world examples.

Thank you! We appreciate your very constructive suggestions and are very grateful for your review.

I think, however, that the paper does still need some additional work in two main areas:

Major

1. *The key thing that I felt did not come across very clearly from the paper was exactly which aspects of the paper were new, from a statistical (methodological) perspective - I think that this needs clarification, and to be outlined explicitly within the abstract and discussion. The authors do explain in a clear and concise way that they are taking methodology from one area of application (meta-analysis, rom clinical trials) and applying it to a different area of application (ecology), but what I found less clear from the text is the extent to which the approach they are using for modelling and inference, as described in Section 2.1-2.3, constitutes established methodology (within the meta-analysis literature) and the extent to which it is a refinement or extension of that methodology.*

This is an excellent request for clarification and we are more than happy to concur. The key methodological advance of this paper was to make species richness usable and comparable, especially for microbial ecologists who are interested in diversity but unable to rigorously discuss it. Point estimates and standard errors for species richness have never been endowed with an analysis procedure, and this paper describes a first method. It is generally not understood in ecology that species richness estimates are random, because they are functions of the observed data. Thus incorporating this randomness into modelling is the methodological advance of this paper. Indeed, we did not modify the standard metaanalysis methodology, because this literature is very well developed and extending it was not necessary for our purposes.

To clarify that the methodological advance of this paper is more related to species richness than metaanalysis, we have made the following changes:

1. We have clarified in the abstract that the meta-analysis literature is used for estimation only.
2. The introduction refines our original course description of "draws together" species richness and meta-analysis to instead state that the key development here is in modelling richness estimates. We hope this removes any ambiguity about the role of metaanalysis in the novelty

of the paper.

3. We have emphasised this again in the penultimate paragraph (Section 6.4): "Parallels with meta-analysis were used to inform the parameter estimation procedure, but the key innovation of the paper was accounting for species richness estimate randomness rather than modifications to the meta-analysis framework."

Thank you for the suggestion; we believe that it substantially clarifies the key novelty of this paper.

2. *The approach that is proposed in the paper essentially constitutes a form of two-stage modelling: estimating the species richness (and associated uncertainty) separately for each sample, and then modelling how the SR estimates vary between samples. This two-stage approach seems natural in the context of meta-analysis, where estimates and standard errors will often be the only information that are available for each study, but it is less clear to me why a two-stage approach is desirable/appropriate here, given that the raw data are available for all samples - wouldn't a more flexible, and potentially more efficient, alternative approach be to model the raw data values from within each sample directly, using a single hierarchical model that captures both the sampling mechanism (within each sample) and the variation between samples? I think this needs to be commented upon within the paper, and a defensible rationale for the use of the two-stage approach needs to be given.*

This is an excellent question. We completely agree that integrating species richness and comparison is an interesting and potentially fruitful idea. We interpret your approach to signify specifying a particular model structure common to all samples, but modelling the model parameters for each sample as a function of the covariates attributed to that sample. It seems very possible that species richness model parameter estimation is more stable than estimating species richness itself, and that this approach could be very productive. However, I suspect that any method that leverages the raw data to see how it is driven by the covariate information would be substantially less computationally efficient, because the model would need to explore the effect of each covariate on each observed OTU (or at least a reasonably large collection). This substantially increases the dimension of the parameter space in comparison to our approach, because we only consider "downstream" estimates of species richness. We are aware of some work in this direction for diversity indices, but none that consider unobserved species. Nonetheless, because this is a very interesting suggestion, we have included a discussion of it as a potential extension to the method in Section 6.3. We have been instructed not to substantially increase the length of the paper, but hope that the new paragraph adequately addresses your suggestion for discussing this approach.

Minor

1. *Section 2: I think it's worth stating explicitly at an early stage of this section that the standard way of doing these sort of analyzes in the ecology literature is using linear regression - this is mentioned later on, but I think it would make the paper easier to follow if this were stated more clearly at an early stage.*

We agree completely. The opening paragraph of Section 2 now contains the sentence "The current standard approach to comparing species richness (a linear regression on the observed richness) critically fails to account for the latter component.”

2. *Section 2: It is mentioned (lines 151-155) that linear regression is the most commonly used approach at present, but that wording implies that other approaches are also currently used in the literature - I think these alternative existing approach need to be briefly commented upon (ideally with references).*

We cut this paragraph in accordance with a suggestion from Reviewer 1. However, to incorporate the suggestion we have also mentioned the less common approach approach of a regression on the Chao1 index, along with two recent papers that employed this technique, in Section 3.1.

3. *Section 2, Lines 92-93: I wasn’t clear why it was appropriate here to assume that this is a linear function of the covariates, given that species richness must be positive. Wouldn’t it be more defensible, for example, to assume that log(species richness) is a linear function of the covariates, to ensure that the model will always predict non-negative values?*

Great question! While many estimation procedures (e.g. maximum likelihood, weighted least squares) guarantee asymptotic normality of the estimate around the truth, in general we do not have guarantees of normally distributed errors of transformed estimates. Normally distributed errors are critical for the pivot distributions of the test statistics, and cannot be dispensed with. We believe that this is more important than preventing negative species richness predictions, because species richness estimates are normally large enough that this would only occur far from the range of the observed covariates (and extrapolation should never be performed here anyway). However, if a species richness estimate could be proved to be asymptotically log-normally distributed, fitting the model on the log scale would be critical. This is an excellent request for clarification and we have elaborated on this point in a footnote on Page 4.

4. *Section 2, paragraph beginning line 109: some additional overview at this stage as to how species richness is commonly estimated for each sample would be very useful - the text mentions the method of inference, but not the form of the model for estimating SR from each sample.*

This is an excellent suggestion. I have included a statement about three very broad classes of models that are commonly used for estimating richness in microbial studies: homogeneity, mixed Poisson, and non-mixed Poisson. While many different implementations exist for these model classes, I have also referenced their most commonly used sources in microbial studies; namely the Chao1 estimator, CatchAll, and breakaway, respectively.

5. *Section 2, paragraph beginning line 109: Does the proposed approach rely on species richness being estimated in the same way for all samples within an analysis? Are there any potential problems if different approaches to estimation are used for different samples?*

I have spent a lot of time thinking about this question since we started working on this paper. My current opinion is that in theory, it is not necessary that the same model for richness be used for all samples. All models estimate the same target, and (in theory) give reasonable estimates of their uncertainty. However, in practice this can induce substantial problems. Certain models consistently produce lower estimates than others, and lower standard errors. In some cases this is well understood (e.g. the popular Chao1 estimate is known to be severely downwardly biased unless every taxa has the same relative abundance), but in general can only be explained by model flexibility, with more flexible models giving larger estimates, but also larger standard errors due to more parameters. As such, my advice to `betta`'s users is to only run the procedure on richness estimates obtained from the same method. This is because I think that introducing bias via different estimation procedures is worse than "letting the estimates compete on the same field," even if it entails model misspecification.

However, in the interests of full disclosure, I have been considering investigating the efficacy of model-based metaanalysis for species richness in an attempt to soothe my worries about model misspecification in the species problem. It's very possible that this approach will raise more problems than it solves, but I do not want you to think that I am inconsistent.

Thank you for raising this very excellent point. To ensure that the reader of the manuscript has adequate information, we have included in Section 6.1 (*Model selection and diagnostics*) the suggestion "Furthermore, since richness estimates are often highly sensitive to the chosen model, we encourage `betta`'s users to only compare estimates obtained from the same estimation procedure." Given that you raised this question in Section 2, we have noted in this section that a discussion can be found in Section 6.1.

6. *Section 2, paragraph beginning line 156: I think it is worth noting somewhere in this paragraph that the homogeneous version of the model is simply a weighted linear regression model.*

That's an excellent suggestion. The last sentence of the paragraph now reads "Note that in the absence of covariates, the model simplifies to a weighted linear regression with the estimates weighted by the inverse of their variance estimates."

7. *Section 3, lines 299-301: "are due to the differing data structures (high versus medium latent diversity)": is it due to the differing data structures themselves, or to the different models that are used to estimate species richness in these two cases? Or some combination of the two?*

Great question! It is due to the differing data structures themselves. The models for species richness were merely chosen to reflect the data structures. While it is possible to also reverse the richness estimates on the datasets, this would in fact constitute model misspecification: CatchAll is far too inflexible to capture high diversity datasets (documented in Willis & Bunge, 2015), but breakaway is overparameterised in the medium diversity case. To clarify, we have changed the sentence to read "due to the differing data structures (that is, high versus medium rare diversity observed in the samples)..." The question of species richness being estimated in the same way was addressed previously; we hope you find these two changes to satisfactorily address your questions about data structures and models.

8. *Section 3: Table 3 caption: "The homogeneity test captures the effect of absent predictors,*

> *thus no covariates were employed": I didn't understand this sentence, I'm afraid: what does it mean to say that the test captures the effect of absent predictors? Is the 'homogeneity test' that is mentioned here the test that is described on lines 156-170? If so, that surely is conditional upon the set of predictors that are included within the model, so I couldn't see in what sense it could be said to capture the effect of absent predictors?*

I absolutely concur that the statement needs clarification, and am grateful for the request. The purpose of the sentence was to capture that the microscale heterogeneity of microbiomes mean that there are unobservable covariates associated with any sample, including replicates, but that the variability arising from this source (non-modeled covariates) is captured by the variance estimate σ_u^2, and is thus tested for by the heterogeneity test. However, the sentence is completely misplaced in this caption, and I have changed it to read "No covariate information was modeled in this simulation."

9. *Section 3: Line 391-392: "Where replicates are available": does the word 'replicate' here effectively refer (in the context of real-world data) to technical replicates from the same biological sample?*

Not necessarily. Any manner of replication could be used to estimate standard deviations in richness estimates. However, a replication procedure that reflects true experimental variability could give the most realistic picture of uncertainty in the estimate. For example, true biological replicates would be favoured over sequencing/extraction ("technical") replicates because they incorporate both variation within the environment and within the sequencing protocol. This is an excellent request for clarification, and the relevant paragraph now begins "Biological replicates are ideal for estimating the standard deviations of richness estimates because they incorporate variability due to both environmental and sequencing sources. Technical (sequencing) replicates, while only dealing with the latter source of variability, can also provide some information on estimate variability. We encourage the use of true, biological replication to empirically confirm standard errors..."

10. *Section 4, lines 408-410: how do the samples (biological replicates) differ from each other? Are the samples collected at exactly the same time and location as each other?*

Yes, the samples were collected from the same field site, on the same day (within the same hour), and with the same fertilizer characteristics (no soil amendments). However, to induce true replication, they were sampled from different plots (i.e. they were not technical replicates from sequencing) that were laid out using a randomised complete block design. To maintain the flow while addressing your concerns, we have removed the term biological replicates from the opening sentence in this section, and provide the details of sampling in the following paragraph. The description of the study now reads "We analyse field replicates within different plots of the same field (under a randomized complete block design) that were sampled within the hour and without amendments to the soil."

11. *Section 4: Fig 1: I think it's worth stating in the caption that the samples have been ordered by the value of observed richness (assuming this is indeed the case).*

Agreed; and this indeed is the case. The relevant sentence of the caption now reads “This ordering of the replicates (increasing with the value of observed richness) suggests inconsistent levels of biodiversity...”

12. *Section 4: Lines 431-432: ”Note that if only observed richness is considered, the samples appear to have different richnesses”: It’s worth noting here, I think, that no formal hypothesis test of this is possible using existing methods.*

Agreed; the sentence now reads “Note that if only observed richness is considered, the samples appear to have different richnesses (Figure 1), though no inferential method for determining this was previously available.”

13. *Section 5: Lines 545-546: ”and bootstrap errors understate true sampling variability”: It would be useful to include a reference for this.*

Agreed; we have included a reference to the very accessible overview by Kulsea *et al.* (2015).

We are very grateful for your helpful and considerate suggestions. You raised many important questions and the resulting changes considerably clarified some key points. Thank you very much for your time! We have added our thanks to the manuscript in an Acknowledgements section.

Associate Editor

This paper uses methods from the meta-analysis of clinical trials literature to develop a way of rigorously detecting changes in species richness. The naive approach of using observed numbers of species to assess changes has obvious problems, with these values actually providing under estimates of true numbers due to imperfect detection.

The application of the paper motivates the work well, and the paper is clearly and concisely written, whilst retaining sufficient detail for non-experts in the field to follow the approach taken. The approach is supported by nicely presented simulations.

Papers for JRSS-C should have an element of statistical novelty, and I suspect there are adaptations of the methods for implementation within this paper which need to be highlighted to make it suitable for publication within this journal. Two referees have reviewed the paper and have provided useful and comprehensive reports to guide the revision of the manuscript.

Thank you very much for your review of our paper we are very happy that you find the paper well motivated and well written. We have addressed the request of yourself and Reviewer 2 to highlight the statistical novelty of the paper, which was in progressing the usual approaches to estimating species richness into a usable framework for comparison and inference. We have furthermore addressed every concern of the reviewers, and believe that their comprehensive critiques have substantially improved the manuscript. The manuscript is a single page longer than the original submission, in accordance with the Editor’s request not to substantially increase its length. We hope that our revisions are acceptable to the editorial body and the reviewers, and are very grateful for such thorough suggestions.