Response to Reviewers' Critiques

R Journal article 2020-141

"krippendorffsalpha: An R Package for Measuring Agreement Using Krippendorff's Alpha Coefficient"

by John Hughes

I thank the reviewers for their thoughtful and constructive comments. I have made a concerted effort to address all of their concerns.

Reviewer 1

This manuscript introduces a package to compute Krippendorff's alpha coefficient for inter-rater reliability. Inter-rater reliability has a wide variety of use cases and would likely be of interest to a broad audience. I thought the manuscript did a good job of motivating the derivation of the statistic by starting with a model that is more familiar to researchers, a one-way mixed model, with random coders and the intra-class correlation.

The package enhances prior implementations of this statistic to follow many R specific implementation practices, i.e., summary(), confint(), etc, which should help R users familiar with other R model fitting functions to use more readily. I thought the examples were clear and the primary function, krippendorffs.alpha(), was clear and follows typical R model fitting procedures. I took a peek at the package code and have a few questions/concerns.

1. First, the package does not have any unit tests, therefore the consistency on how it performs or testing for places where it should break are not being performed. I have concerns about using packages without unit tests for these reasons.

   I used the `testthat` package to create tests that are executed automatically during `R CMD check`. The test code exercises functions `krippendorffs.alpha`, `confint`, and `influence` using the nominal data that are analyzed in the article and in some of the package's examples. To test `confint` it was of course necessary to set the seed for the pseudorandom number generator so that the confidence limits would be the same for each execution (assuming the code is still correct). Thank you for suggesting this enhancement of the package.

2. Secondly, there are a few places where the '...' is used within functions, but these are not actually implemented in the functions themselves. In thinking about this, I think the most problematic is within the 'confint()' function where users may want to specify non-default arguments to the 'quantile()' function.

   Thank you for pointing this out. For a while now I have been using ... only for compatibility with the S3 framework. In light of your comments, my `confint` and `summary` functions now pass the ... arguments to the `quantile` function, and my new `plot` function passes the ... arguments to the `hist` function.

3. Third, although not a complete deal breaker to me, but it would be nice to have a presence for tracking any issues with the package rather than just an email address. An email may not be persistent, could be more easily ignored, or could prevent some from reaching out to the author.

   I changed the package's description file to include the URL of my academic website: `http://www.johnhughes.org`. My site has a page devoted to my various software packages (Perl and R): `http://www.johnhughes.org/software.html`. I have maintained this site since 2011.

   I also created a GitHub repository. The repository can be found at `https://github.com/drjphughesjr/krippendorffsalpha`.

4. Finally, although the package only has a few functions, a package vignette would also be nice to have. Perhaps this manuscript will aim to be that vignette in the future, but that is another element that I think most R users expect to see.

   I do intend to make this article the package vignette. As soon as we agree on a publishable version of the manuscript, I will add the finished article to the package. Thank you for suggesting this.

Specific Comments

There were a few other specifics that I found to be missing that I wondered about as reading through the manuscript. Some of these are highlighted in more detail below, but I provide an overview of those areas here. First, specifics of the package implementation regarding missing data or unbalanced data would be helpful for readers. Second, I was looking for a bit more context for the specific examples and why these statistics would be helpful or answer an informative question. Third, the bootstrap figures are very informative and helpful, but I wonder if they warrant a bit more interpretation in the text to guide readers on informative elements to look for.

5. Are there any notable differences in the computation or details for the unbalanced version?

   Handling the unbalanced case is messier since it requires adjustment of the weights and the introduction of indicator variables to reflect the unequal subsample sizes. If you feel that something is to be gained by presenting the equations for the unbalanced case, I will be happy to add them to the article.

6. Also, the first example had missing data, how is missing data handled and how does this impact the statistic of interest? Are cases dropped with missing data and if so, it is listwise or pairwise? Details of this would be helpful for readers. I assume this is present in the primary resource, but how missing data is handled in the package implementation would be helpful to detail here at the very least.

   The revised version of the article now explains how the package handles missingness. Please see the bottom of page 7. Thank you for pointing out my oversight.

7. I would have liked to see a bit more detail about the examples and what disagreement means. This is particularly true for the cartilage area in which I do not have experience in. For example, why is it of interest to compare inter-rater reliability for the two different measurements?

   I added a bit of information regarding the potential risks associated with the use of gadolinium-based contrast agents (GBCAs). And I provided the URL of the University of California, San Francisco's policy regarding the use of GBCAs. The UCSF webpage provides detailed information and many references. The following addition can be found at the bottom of page 9 of the revised manuscript. Thank you for pointing out this issue with the previous version of my article.

   > We see that $\hat{\alpha} = 0.84$ and $\alpha \in (0.81, 0.86)$. Thus these data suggest that raw T2* measurements agree almost perfectly with contrast-enhanced T2* measurements, perhaps rendering gadolinium-based contrast agents (GBCAs) unnecessary in T2*-based cartilage assessment. This finding could have clinical significance since the use of GBCAs is not free of risk to patients, especially pregnant women and patients with impaired kidney function. For much additional information regarding the potential risks associated with the use of GBCAs, we refer the interested reader to the University of California, San Francisco's policy on MRI with contrast: `https://radiology.ucsf.edu/patient-care/patient-safety/contrast/mri-with-contrast-gadolinium-policy`.

8. I also wonder if more interpretation of the figures showing the bootstrap results could be expanded upon to highlight informative areas.

   The revised manuscript includes additional interpretation of both bootstrap sample plots.

9. Since Krippendorff's alpha was defined from the one-way mixed-effects ANOVA model and intra-class correlation, does the interpretation of explained variance due to the random effect (in this case coders) carry over to the inter-rater reliability? If so, I think this would further add to the interpretation of the statistic for readers.

   If I understand you correctly, the answer is yes, the interpretation of explained variance due to the random effects does carry over to the inter-rater reliability measure. This is made explicit in the first equality of the second equation on page 2 of the article:

   $$\alpha = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}.$$

   And this of course carries over to the definition of the estimator $\hat{\alpha}$, which follows

   $$1 - \frac{\sigma_\varepsilon^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}$$

   instead of

   $$\frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2},$$

   that appears as equation (1) on the same page.

10. Finally, I found the code included in the manuscript to be clear, well and consistently styled (i.e., spaces, consistent syntax), but I wonder if the examples could have better more descriptive names instead of fit, fit2, fit3, etc.

    I changed all of the variable names in the paper to names that I hope are more illuminating yet not too long. I also changed the variable names in the package's examples. Thank you for suggesting this.

Reviewer 3

The author provides a very interesting and valuable overview about Krippendorff's reliability measure α: Deriving the measure from a (special case of a) one-way mixed-effects ANOVA, the author lays down how α can be generalized either nonparametrically to form a modified mul tiresponse permutation procedure (MRPP) or parametrically to arrive at a Gaussian copula-based measure of agreement. With regard to statistical inference, the author proposes a Monte Carlo based resampling procedure as a means to obtain valid confidence intervals and cautions about the risk of influential cases which may lead to erroneous conclusions. In the remainder of the paper, the author illustrates how "krippendorffsalpha" may be used to (1) compute point and uncertainty estimates of α, (2) detect influential data points, and (3) display and extract key statistics using S3-methods.

The article makes a valuable contribution to the R community by putting a key inter- and intra- coder agreement measure into a larger methodological context, discussing ways to obtain uncertainty estimates and dealing with influential cases. The approach taken to describe Krippendorff's α and to illustrate its computation with "krippendorffsalpha" is generally sensible, as well the choice of technology and methods.

I identify several points which, once addressed, will further improve the paper:

1. Despite the very interesting methodological discussion of α, it does not appear as clear and natural to the reader as it might why the discussion and derivation of α is necessary and how it is linked to the package. I hence suggest explaining more clearly what situating α among statistical procedures contributes to our understanding of Krippendorff's α and its use as an agreement measure. How do α's characteristics motivate the different functions and methods of "krippendorffsalpha"? Furthermore, what makes Sklar's ω being a parametric generalization of α (with a squared Euclidian distance) relevant for working with and conducting inference using Krippendorff's α and "krippendorffsalpha"?

    I am reminded of the following Victor Hugo quotation.

    > Phenomena intersect; to see but one is to see nothing.

    In any case, the beginning of Section 2 of the manuscript now appears as follows.

    > Since Krippendorff's α is defined in terms of discrepancies (Krippendorff, 2013), at first glance one might conclude, erroneously, that α is a measure of *dis*-agreement and so answers the wrong question. In Sections 2.2.1–2.2.3 we will show, by examining Krippendorff's α's place among statistical procedures, that α is, in fact, a sensible measure of agreement. Also, establishing a context for α may help practitioners make educated decisions regarding α's use.

> The UML class diagram (Fowler et al., 2004) shown below in Figure 1 provides a conceptual roadmap for our development. Briefly, a special case of α (which we denote as Alpha(SED) or α_{SED}) arises naturally in the context of the one-way mixed-effects ANOVA model. Alpha(SED) can then be generalized in a nonparametric fashion to arrive at Krippendorff's α as it has been presented by Hayes and Krippendorff (see Gwet (2015) for a development of nonparametric α in terms of agreement rather than discrepancies); this nonparametric form of α is a (slightly modified) multiresponse permutation procedure. Alternatively, α_{SED} can be generalized in a parametric fashion to arrive at Sklar's ω, a Gaussian copula-based methodology for measuring agreement.

I also added the following sentence to the section on α as an MRPP.

> Note, however, that although α can be viewed an an MRPP (as we are about to show), α has been modified for the purpose of measuring agreement rather than discerning differences.

I hope these additions clarify sufficiently my reasons for relating α to other statistical procedures, especially the two parametric procedures, both of which model agreement explicitly as positive correlation. Thank you for pointing this out as a potential cause for confusion.

2. How do α's characteristics motivate the different functions and methods of "krippendorff-salpha"?

   Alpha's relationships to other statistical procedures do not motivate the various functions offered in package `krippendorffsalpha`. I chose to provide the S3 methods `confint`, `influence`, `plot`, and `summary` so that package `krippendorffsalpha` would conform to the usual way of doing things in R. The only aspect of `krippendorffsalpha` that was motivated by α itself is the package's support for user-defined distance functions. This feature was motivated by the most general form of α, in which the distance function $d^2(\cdot, \cdot)$ can take any of a great many forms.

3. Gwet (2014) provides a methodological discussion of Krippendoff's α as well, how does it relate to the author's discussion?

   Since Gwet (2015) develops nonparametric α in terms of agreement as

   $$\alpha = \frac{p_a - p_e}{1 - p_e},$$

   my article now mentions this fact and cites Gwet (2015) as well as Gwet's book. Thank you for pointing out my oversight.

4. The author lists two extant R packages implementing Krippendorff's α, however, at least two more packages exist (icr, irrCAC). I therefore suggest checking whether there are more packages implementing α and to add them to the discussion. Ideally, provide a table comparing those packages along several functional characteristics.

Thank you for pointing this out. Instead of providing a table, I elected to discuss the other packages at various places throughout the paper. For example, the revised paper discusses interval estimation at the top of page 5. The added paragraph appears below.

> We carried out a number of realistic simulation experiments and found that this approach to interval estimation performs well in a wide variety of circumstances. When the true distribution of $\hat{\alpha}$ is (at least approximately) symmetric, Gwet's closed-form expression for $\hat{\mathbb{V}}(\hat{\alpha})$, which is implemented (for categorical data only) in package `irrCAC`, also performs well. By contrast, we found that the bootstrapping procedure recommended by Krippendorff (2016), which is implemented in packages `kripp.boot` and `icr`, generally performs rather poorly, producing intervals that are far too narrow (e.g., 95% intervals achieve 74% coverage).

If you feel strongly that a table is necessary, I will be happy to add one.

5. Note that as of writing this review, Proutskova and Gruszczynski's "kripp.boot" has been removed from CRAN and is only available from GitHub (https://github.com/MikeGruz/kripp.boot).

   Thank you for pointing this out. The article now reflects the change in location of package `kripp.boot`.

6. "krippendorffsalpha" is (to my knowledge) unique in that it allows user-defined distance functions and provides standard S3-methods for confidence intervals, and to obtain influential values. I believe these unique aspects should be emphasized more strongly in the paper, as these may be the reason a user might want to read the paper and use the package in the first place.

   Thank you for pointing this out. The revised manuscript mentions these advantages in numerous locations.

7. Krippendorff defined a bootstrapping procedure (see https://www.asc.upenn.edu/sites/default/files/documents/boot.c-Alpha.pdf, implementations in SPSS and SAS can be found here https://www.afhayes.com), which is also implemented in "kripp.boot". Gwet (2014) proposes a closed-form formula to compute the variance of α. Given the extant literature and packages regarding confidence interval estimation, I suggest the author elaborate on the differences between the Monte Carlo resampling procedure described in the paper and the other approaches (it seems the author's approach is somewhat similar – yet not identical – to Krippendorff's algorithm).

   Please see my response to your fourth comment.

   I am not sure why Krippendorff's bootstrapping algorithm yields a very poor estimate of the true distribution of $\hat{\alpha}$, but I suspect that, by resampling pairs of scores instead of rows of scores, his procedure fails to account for the variability due to the random effects (viewing $\hat{\alpha}$ in the context of the one-way mixed-effects ANOVA model or, more generally, Sklar's ω).

Specific Comments

The code within "krippendorffsalpha" is well written and organized. I believe the package may however be even further improved by:

8. Using braces consequently, even in the case of single-line if statements and after for-loops (the R "styler"-package provides functions to reformat a package's source code at once).

   I am sympathetic to your concern but respectfully disagree that wrapping single-line if statements and single-line loop bodies improves readability or makes code easier to maintain. Consider the following function, for example.

```
krippendorff.total = function(data, dist)
{
    m = rowSums(! is.na(data))
    D.e = 0
    n.u = nrow(data)
    n.c = ncol(data)
    for (i in 1:n.u)
        for (j in 1:n.c)
            for (k in 1:n.u)
                for (l in 1:n.c)
                    D.e = D.e + dist(data[i, j], data[k, l])
    n = sum(m)
    D.e = D.e / (2 * n * (n - 1))
}
```

   In my opinion, adding braces for each of those loops would merely clutter the code a bit. The current version, by contrast, allows the reader to ascertain at a glance that a quadruply-nested loop is the heart of this function.

   In any case, Google's (well-respected) style guide for R permits the omission of curly braces for single-line if statements and single-line loop bodies. And the same is true for other programming languages, e.g., C++, Python.

9. Using underscores instead of points in variable names.

   I am sympathetic to your concern regarding naming, but I prefer to use dots instead of underscores for both variable names and function names. This is because a great many of the most commonly used R functions, arguments, and variables use dots, e.g., `is.matrix`, `na.rm`, `as.data.frame`, `confint.glm`, `conf.level`, `R.version`, and many more. Also, R's S3 framework uses dots.

   Interestingly, there does not appear to be a consensus regarding naming in R. For example, the tidyverse style guide recommends that variable and function names use only lowercase letters and separate words with an underscore while Google's style guide recommends `BigCamelCase` (for function names, at least).

10. Using message()/warning() instead of cat().

    I changed all instances of `cat` to `message`, except in function `summary.krippendorffsalpha`. Thank you for suggesting this.

Furthermore, the author may consider

11. Providing automated tests to ensure results remain consistent after changes to the code.

    I used the `testthat` package to create tests that are executed automatically during R CMD check. The test code exercises functions `krippendorffs.alpha`, `confint`, and `influence` using the nominal data that are analyzed in the article and in some of the package's examples. To test `confint` it was of course necessary to set the seed for the pseudorandom number generator so that the confidence limits would be the same for each execution (assuming the code is still correct). Thank you for suggesting this enhancement of the package.

12. Providing a software development page (for instance on r-forge, GitHub, GitLab) to increase the package's visibility and allow users to more easily interact with the author.

    I changed the package's description file to include the URL of my academic website: `http://www.johnhughes.org`. My site has a page devoted to my various software packages (Perl and R): `http://www.johnhughes.org/software.html`. I have maintained this site since 2011.

    I also created a GitHub repository. The repository can be found at `https://github.com/drjphughesjr/krippendorffsalpha`.