From: **PeerJ** peer.review@peerj.com
Subject: Decision on your PeerJ submission: "A distribution-based effect size is more reproducible than hypothesis testing when analyzing high throughput sequencing datasets" (#2021:04:59967:0:2:REVIEW)
Date: June 28, 2021 at 11:09 AM
To: Gregory Gloor ggloor@uwo.ca

PeerJ

Dear Dr. Gloor,

Thank you for your submission to PeerJ.

It is my opinion as the Academic Editor for your article - **A distribution-based effect size is more reproducible than hypothesis testing when analyzing high throughput sequencing datasets** - that it requires a number of **Major Revisions**.

My suggested changes and reviewer comments are shown below and on your article '*Overview*' screen.

Please address these changes and resubmit. Although not a hard deadline please try to submit your revision within the next 35 days.

Resubmission

1. **Use the line numbers in your review PDF** when reading the comments from your editor and reviewers, and when writing your rebuttal letter:

Download review PDF

2. Download your resubmission checklist:

Download checklist

3. Edit and resubmit when ready:

Edit and resubmit

With kind regards,

Andrew Gray
Academic Editor, PeerJ

Editor comments (Andrew Gray)

MAJOR REVISIONS

As you will see below, three reviewers have assessed your manuscript and made a collection of useful and thoughtful comments. I'd like to thank them for the obvious effort they have collectively spent on their reviews. You might be able to address all of their (and my) comments in a revised version of your manuscript, but I will note that I think that this is likely to require substantial revisions. If you feel you can address these comments, I will ask you to consider and respond to each of their comments separately, highlighting and explaining the changes you've made to your manuscript in response to each or explaining why you have not made any changes where you feel that is appropriate. I think that they have covered the important aspects already, but I will add a few points myself further below, some of which are very minor, for you to also address in the revised version of your manuscript. I look forward to seeing a new version of your work in due course.

Reviewer #1 has raised some important questions and asked for more information about your model, the data used, and the methods, along with clarifying some of your nomenclature (which they note is not always used consistently). I think that you will find all of their comments to be very useful in revising your manuscript.

Reviewer #2 is more positive, but has noted the exploratory nature of your work and asked some questions about its performance given data set characteristics and alternative approaches.

Reviewer #3 has also raised important questions that I think will require careful thought, including about the structure of your manuscript. As with Reviewer #1, they ask you to clarify your terminology and your notation. (Also, like Reviewer #1, they ask you to think about your title.) They ask some interesting and thoughtful questions about how you motivate your approach and I think your manuscript would be greatly improved if you are able to address (and so anticipate from readers) these questions, in particular the crucial one of why a researcher would switch to your new method. (You might feel that you have answered this question already, but from a reader’s perspective, I agree that there is work to be done here.)

As suggested in your introduction and its references, replicability (in terms of NHST) will always be problematic in underpowered studies (e.g., Lines 14–15 and implied by Line 24) due both to the low power to detect actual effects of an interesting size and the exaggerated effects identified when statistically significant results are observed despite small to absent true differences. Underpowered studies are prone to replication issues because of their very design, but the point that statistical power is always conditional on the size of the effect of interest is sometimes omitted in the literature (including Halsey, et al., note that Box 1 defines power without mentioning this, although this is remedied in Box 2). In short, a study cannot be under- or well- (or over-) powered in absence of identifying an effect of interest, and given the implausibility of most null identifying an effect of interest, and given the implausibility of most null hypotheses being true in the point sense, we need to carefully consider what a practically meaningful effect size would be.

I wonder if what I interpret as your intention that this method would be applied to small/pilot studies could be made explicit in the abstract and if you could elaborate on usual sample sizes around Line 24. The need for a “suitable non-parametric alternative” (Line 15) seems to need justification to me, despite Lines 44–45. What is your argument that a suitable alternative ought to be non-parametric given this situation? There are alternatives to NHST; would these be useful, in your view, to researchers in the situation described on Lines 28–30? I’m not sure that the “often” on Line 32 is justified, perhaps “sometimes”?

It seems to me that formal tests of normality (e.g. Lines 87–90) cannot be used in the way described here. In small samples, visible departures from Normality (such that they will not provide reassurance of Normality) will not necessarily produce statistically significant test results; in large samples, where the Central Limit Theorem will make the sampling distribution approximately Normal for very non-normal distributions, even minor departures of the kind that hardly threaten the interpretation of arithmetic means and SDs will produce evidence against Normality.

Line 103: Perhaps pedantically, while the IQR can be estimated from these figures, in a sense they “show” the 25th and 75th percentiles and not the IQR (the difference between these two values).

I think the definition of “sensible estimates” (e.g. Lines 109 and 141) needs formal definition.

In the Discussion, it would be helpful to identify the limitations of the present work (not just the E statistic) and outline some future directions.

Reviewer 1 (Anonymous)

Basic reporting

Given in attached pdf

Experimental design

Given in attached pdf

Validity of the findings

Given in attached pdf

Annotated manuscript

The reviewer has also provided an annotated manuscript as part of their review:

Download Annotated Manuscript

Reviewer 2 (Anonymous)

Basic reporting

The paper describes a new statistic E that is expected to be more robust to non-normal data, particularly when the sample size is small. The notion "median of the maximum absolute deviation" is novel, so the statistic E. The paper then followed up with examples to illustrate the robustness, in extremely small sample sizes. Using the criterion E>1, the authors that the FDR can be controlled as low. The statistic is presumably useful as an exploration tool. But caution should be exercised when rigorous inference is needed and this method is not reliant on rigorous inference.

Experimental design

The simulation experiment is rigorously presented.

Validity of the findings

The finding is rigorously presented. Yet, it should be more explicitly stated that this tool is useful for exploration, not in place for rigorous inference.

Comments for the Author

The paper describes a new statistic E that is expected to be more robust to non-normal data, particularly when the sample size is small. The notion "median of the maximum absolute deviation" is novel, so the statistic E. The paper then followed up with examples to illustrate the robustness, in extremely small sample sizes. Using the criterion E>1, the authors that the FDR can be controlled as low. The statistic is presumably useful as an exploration tool. But caution should be exercised when rigorous inference is needed and this method is not reliant on rigorous inference.

I hope that the authors could explore a bit more on the properties of the E statistic when the sample size is decent say >20. How does it compare with a t-statistic when the data is/is not normal, and how does it compare with using a t-statistic but get p-values using permutations?

Reviewer 3 (Riko Kelter)

Basic reporting

The paper is written in unambiguous, professional English throughout. Literature references are sufficient, but the article requires restructuring. It requires extensive study of the supplementary material.

Experimental design

The research question is not motivated sufficiently. A key problem I encountered when reading the paper is that well-defined statistical terminology is often used inappropriately and it is difficult to understand what meaning a sentence should convey. Examples include conflations between “midpoints” and “medians” of distributions, and details are given below. The experimental design is clear although I hesitate to accept the conclusions drawn without further clarification. Details below in the general comments section.

Validity of the findings

See above, the underlying data, code and algorithms are available. Results can be replicated. However, I hesitate to accept the conclusions drawn in full generality without further explanations, see below.

Comments for the Author

I enjoyed reading the paper and found the approach interesting (although the comments below sound a little negative; however, this is primarily because of the fact that notational issues severely complicate the understanding of your approach). My major concern is that it remains unclear why someone would not consider equivalence testing approaches from the outset, and what is the justification of E, when p-values come with type I error bounds, consistency properties, et cetera. Details below.

Major points:
a) The paper argues that it introduces a novel non-parametric standardized effect size estimate for high-throughput sequencing datasets. While the introduction section was written well, the primary problem seems to be that the authors consider multimodal distributions and deem Cohen’s d inappropriate in such situations. The methods section needs to be restructured to clarify this. Also, the title is cluttered and should focus on this aspect.
b) From line 91 on, the definition of the new statistic is entirely unclear based on the provided information. Only after reading the supplement the intention became clear. In the supplement sections “Effect size” the authors without any obvious reason avoid the use of well-defined statistical terminology:
In equation (1), the denominator is termed dispersion, although Cohen’s d uses the pooled standard deviation of both groups. Also, the parameters in equation 1 are model parameters and not summary statistics as stated thereafter. Also, z scores are defined for a single distribution, where the observed values are centered and rescaled by the mean and standard deviation of this distribution. Equation (1) clearly uses two distributions, so it is unclear why any reference to z scores is made here.

In the section “Effect size vs Difference” a non-standard definition of Student’s t-statistic is given: Student’s t-distribution can be expressed as a ratio involving normal and chi-square distributions and the t-statistic for two-samples uses the pooled standard deviation, but there is no “dispersion” parameter in it. Dispersion measures include the standard deviation, variance and the variance coefficient, but the t statistic includes the pooled standard deviation of both groups in a two-sample setting.

In the section “The distribution effect size E” it is stated: “the standardized difference between the means (or midpoints) of the distributions”. Means are not midpoints. Midpoints are not even a well-defined statistic of a probability distribution. Also “two distributions represented as a and b” probably refers to assuming two random variables X and Y where a=(x_1,…,x_n) and b=(y_1,…,y_m) should represent two samples of sizes n and m from the distributions of X and Y.

Also, it is stated that “the purpose of E is to determine the standardized difference between two distributions”. However, there are well-established concepts to measure differences or distances between probability distributions, e.g. the Kullback-Leibler divergence, Shannon-Entropy or Hellinger’s distance. No discussion is provided why existing approaches to measure distances between probability distributions are ignored. Also, it is not motivated why the dispersion is calculated by simply permuting the vector and building the difference. What is the justification / motivation? Why not simply taking the maximum difference between two elements of a? Or the median of differences between elements in a?

A few lines after that it is stated that “both the difference and dispersion estimates are vectors and not point estimates.” Again: A vector can be a point estimate. In fact, the dimension of a point estimate is entirely unrelated to the fact whether a function is a (point) estimator or not. Please check standard textbooks (Casella & Berger, 2002) for the proper vocabulary, otherwise the sections are nearly impossible to understand.

In Equation (4) it is unclear over which set the maximum is taken: Over all permutations of the vectors a and b? Or is it the element-wise maximum for a fixed permutation? In the latter case, disp is dependent on the permutation which is used, in the former it is not.

It also is said that “The vectors recycle if one vector is different in length than the other.” I have no clue what recycle means. Are elements repeated from element one until the length of the other vector is reached, or randomly sampled and appended? Again, it is more guessing than following precise mathematical definitions here.

following precise mathematical definitions here.

The same holds for example for Fig 2: it is not clear what the labels refer to, or what the breakdown point should measure.
c) Line 134-136: Efficiency is well defined for an estimator being unbiased and attaining the Cramér-Rao lower bound for its variance. You argue that your statistic E is as efficient as the estimator (which one?) for the difference in means of both distributions but nothing like that is shown in the paper. Line 136-138: Same problem here. You argue that the efficiency is 52% of the standard deviation in a Normal distribution. A parameter has no efficiency, an estimator has.
d) Fig 4: Your core argument regarding the reproducibility seems to be that using E together with “a fold change and an overlap metric” produces false-discovery rates similar to using the q-score and difference combination. Still, looking at Fig. 4 it seems as if the FP is equal (filled circles), but the TP (empty circles) of q-score and difference combination is higher than for the E statistic with “a fold change and an overlap metric”. I wonder why I should shift to the E statistic then.
e) Based on the simulations, your E seems to be as good as p-values under Benjamini-Hochberg corrections under additional criteria. It is not entirely clear what the fold-change and overlap metric, denoted as triple in Fig. 4A, refers to. Please clarify this. Also, p-values have an elaborate mathematical theory which provides type I error bounds, consistency under the alternative hypothesis, et cetera. As no theoretical results are provided for E regarding these properties, I wonder why someone would shift towards E? I see the appeal of quantifying substantial differences instead of focusing on significance, but there are frameworks in NHST which focus on substantial differences, which are called equivalence tests. There are entire monographies (Wellek, 2010) on the topic, and it should be clarified why no equivalence tests are considered from the outset.

Minor points:
a) Line 166-168: A distribution has two tails. Which one is meant here when defining the overlap?

when defining the overlap?

b) Line 129-130: This statement is false. There are NHST frameworks which are exactly designed to prevent such things like two-one sided tests (TOST) of Lakens et al. and others. These are standard frameworks which are widely used in statistical equivalence testing and clinical trials.

c) Line 230: TPR was TP before. Please use one abbreviation only.

d) Line 252: “consistently” is a well-defined statistical property of an estimator. You have not shown that E is a consistent estimator of the number of true features, so please rephrase or remove.

e) Line 252-256: I see the appeal for very small sample sizes, but who draws conclusions out of a study which involves only two participants?

f) Line 273: “the E metric is the mid point of the effect size distribution”. No, it is not. It is the median, compare equation (6). Even when reading (6) as the median for a sample instead of a probability distribution, the median is not the "midpoint" whenever the sample length is an even number.

g) Line 277-278: “no false positives are identified when no difference is