Review of: "Comments on “The roles, challenges, and merits of the p value” by Chén et al. (Patterns, 2023, 4(12), 100878)"

Cristiano Ialongo[1]

1 Sapienza University of Rome, Italy

Potential competing interests: No potential competing interests to declare.

The debate over the “scientific” utility of the p-value, and the ease with which it can be misinterpreted, is well established within the scientific community.

Reading this work, I personally find it unclear whether the issue lies in the interpretation of the p-value or in its usefulness. Let's start with the example of the z-test. The author lucidly acknowledges the connection between the z-test and Cohen’s d, which is the effect size measure of a systematic difference. However, while the z-test (or z-score) represents the difference after accounting for sampling bias (as the divisor in the formula is the standard error of the estimate), Cohen’s d represents the mean difference after accounting for observed variability (hence the divisor is the standard deviation). Therefore, there is an algebraic factor of 1/sqrt(N) for converting between the two, allowing for the translation of d into z, and vice versa. That said, it is clear that the use of the standard error aims to give an inferential aspect to the use of this statistic, which does not happen with Cohen’s d, which remains limited to the sample level. Thus, the problem of p-value interpretation here seems to stem more from the inferential nature of the z statistic, which is absent in the effect size measure. On this point, the author seems to get confused or fails to explain it clearly.

Additionally, I am unable to grasp the subtle difference between the interpretation of a one-tailed or two-tailed p-value. The directionality of the test simply verifies the relative position of the two observed means. What seems to emerge from the discussion of the two examples is more the paradigm with which the hypothesis is formulated and the intention behind verifying it: the "probability of compatibility," as the author calls it, which characterizes the two-tailed test, strongly resembles the equivalence of means (Anderson and Hauck). However, the equivalence test of means tests for the absence of a difference within a certain effect size, while the "probability of compatibility" seems to have a less clear (and uncertain from an inferential point of view) practical meaning.

The second point, regarding the fallacy of the p-value due to the inflation of statistical significance, is obvious. That significance cannot depend on sample size is self-evident. However, in this criticism, people often forget that hypothesis testing, or the Neyman-Pearson paradigm, was developed with an inferential intent during a time when experiments involved small samples, few trials, and manual calculations. It’s worth noting that Fisher's randomized approach to the same problems Pearson addressed with the chi-square test generates a p-value (from the exact test) that expresses a probability related to the randomness of sample formation. Therefore, it has a very different meaning. Perhaps this aspect

could have been included in the criticisms raised in this work to clarify the author’s intentions more immediately.

In conclusion, I don’t believe the statement "the p-value measures the difference between two groups (or two populations) at the sample-mean level. Therefore, the p-value is not helpful for extracting evidence from the data or exploring properties of the data" is acceptable, as it disregards the inferential origin and intent of statistical tests.

However, I agree with the idea that a measure of significance cannot depend on sample size (but it can depend on the number of trials, hence the value of randomization tests). In this regard, the answer to the concerns raised by the author does not lie in the use of a “new” significance probability, such as the exceedance probability. I don’t believe this can be considered exhaustive if its asymptotic form does not account for the non-centrality of the distribution of a parameter that assumes the presence of a difference.

So, apart from some repetitions in the text and a certain verbosity in its presentation, I do not find this work to bring innovations worthy of publication.