Review of: "Intersections of Statistical Significance and Substantive Significance: Pearson's Correlation Coefficients Under a Known True Null Hypothesis"

Elena Kulinskaya[1]

1 London School of Hygiene & Tropical Medicine

Potential competing interests: No potential competing interests to declare.

About 10 years ago, a substantial paradigm shift, both in some statistical circles and, even more so, in social sciences and psychology, led to a strong discouragement or even a full prohibition of the use of statistical testing by a number of applied journals (Wasserstein and Lazar, 2016; Wasserstein et al., 2019). Even though the ASA statement itself was rather balanced and encouraged the use of effect sizes with their confidence intervals, statistical inference itself, and therefore techniques such as confidence intervals or Bayesian methods, was also forbidden in some journals, and only descriptive statistics were permitted (Trafimow and Marks, 2015).

To oppose this prohibition, the author demonstrates that, for small to medium sample sizes, the probability of obtaining non-zero effect sizes (ES) is rather high. Thus, it is proposed to use the p-values for the test of $\rho = 0$ based on the Fisher's z-transformed z_r scores to filter the obtained effect sizes.

The article is aimed at applied researchers who lack sophisticated statistical knowledge. To demonstrate how this works, the author simulated Pearson correlation coefficients under the null hypothesis of no correlation $\rho = 0$ for sample sizes $n = 4,\ 30,\ 100,\ 1000$ and 2000. For each simulation, he calculated the number of sample correlation coefficients $r \geq 0.1$ (J. Cohen's threshold for small ES). These values of $r \geq 0.1$ are termed to be of 'Substantive Significance'.

These simulations demonstrate that the number of ES above0.1 that also reach statistical significance at the 5%-level is not very large even for $n = 4$ and quickly decreases for larger sample sizes, as does the number of effect sizes above 0.1 itself, whereas the empirical significance level of the test of $\rho = 0$ does not depend on sample size.

I strongly agree with the author that the idea of cancelling statistical inference is hugely detrimental to the development of science. However, I suggest some corrections and improvements to the content of this article.

Fisher's z-transform transforms Pearson's correlation coefficient r to an approximately normal z-score $z(r) = \ln((1 + r)/(1 - r))/2$. For large enough sample sizes, $z(r)$ has an approximately normal distribution with the mean $\ln((1 + \rho)/(1 - \rho))/2$ and the variance $1/(n - 3)$. These variances, for the sample sizes used in simulations, are provided in row 2 of Table 1. Thus, when $n = 4$, the variance is rather large at 1, explaining the large scatter of the r values, including their large proportion above 0.1. However, the variance quickly decreases with n, reaching 0.01 for $n = 100$ and 0.001 for $n = 1000$. Therefore, the distribution of z-scores becomes tightly concentrated around its mean value of zero, and the

values above 0.1 practically disappear. This is the intuition behind the findings from the simulations.

In the discussion, the author claims that 'no statistical theory predicts the percentage of effect size errors to expect under a true null hypothesis'. However, it is rather straightforward to calculate the probabilities of finding the r values above 0.1 for z-scores: $P(|r| \geq 0.1) = P(|z(r)| \geq |z(0.1)|) = 2\Phi(-z(0.1) * \sqrt{n-3})$, where $\Phi(\cdot)$ is the standard normal distribution function and $z(0.1) = 0.1003353$. I calculated these probabilities for the sample sizes in question. These probabilities and their empirical counterparts from simulations by the author are provided in the lowest two rows of Table 1 and demonstrate good agreement.

Variances, probabilities and empirical fractions for simulated r values.

Sample size	4	30	100	1000	2000
Frequency * ($\|r\| \geq 0.1$)	4428	2983	1556	8	0
Variance of $z(r)$	1.0000	0.0370	0.0103	0.0010	0.0005
Probability * * $P(\|r\| \geq 0.1)$	0.9201	0.6021	0.3231	0.0015	$7.33e-06$
Empirical fraction * ($\|r\| \geq 0.1$)	0.8945	0.6026	0.3143	0.0016	0

* from simulations by E. Komaroff, 4950 repetitions each; * * using normal approximation.

Overall, the author's suggestion of using the p-values to filter out non-significant values of the effect sizes makes perfect sense. However, there is an equivalent, well-established, and more informative option. This option is to report any effect sizes with accompanying confidence intervals. An equivalence between a p-value of a test at a significance level α and a corresponding confidence interval at $1-\alpha$ confidence level is well known. Non-significant p-values $p > 0.05$ correspond to 95%-confidence intervals which cross zero, and the p-values below 0.05 to those which do not. A confidence interval will also provide additional information on its width.

To summarise, I urge all applied researchers to report any effect sizes with their confidence intervals, and I also urge all Editorial Teams to include the reporting of confidence intervals in their editorial guidelines.