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

Johan Lyhagen[1]

1 Uppsala University

Potential competing interests: No potential competing interests to declare.

There has been a long ongoing debate about significance testing, p-values, how to interpret confidence intervals, etc. The current paper contributes to this debate. My personal belief is that the criticism regarding significance testing, p-values, and so forth stems from two sources. The first is that these are complex theoretical constructs and many non-statisticians (but also some statisticians...) have big problems grasping how they should be used and how they should be interpreted, and the conclusion is that they, hence, should not be used. The second one, which mainly relates to significance testing and p-values, is that those criticizing often work in areas where there is no interest in significance testing. Hence, they conclude that significance testing is not useful and should not be used. From my point of view, both reasons are rather anti-scientific. We should not refrain from dealing with useful concepts that are difficult because some have problems understanding and using them correctly. Significance testing is important when developing scientific theories but perhaps not, e.g., when estimating causal effects. You should use the tool that suits your problem. Abuse is NOT a reason for banning significance testing but rather a reason for better statistical education and the realization amongst empirical researchers that they need the help of professional statisticians.

Now to comments directed to the paper. Figure 3 and the text below discuss the test of rho=0 and find that when the significance level is 0.05, 208 cases were rejected compared to an expected number of 247.5. From a pedagogical point of view, this is fine, but if you were more interested in testing if the sampling distribution actually is what we expect it to be, a Kolmogorov-Smirnov test or a chi2 test can be used as they would have higher power. It is also pedagogical of the author to show the difference between the empirical sampling distribution (which is nearly uniform) and the sampling distribution of the test statistic (which is very close to normal). The next exercise is to compare effect sizes based on the estimates and based on a pre-test. As the null is true, and the sample size is small, the effect size estimates are erroneously large, and using a pre-test estimator yields much better effect estimates. This effect is often neglected when discussing significance testing, and it is important to remind the scientific community of it. I do not think Table 2 adds much understanding and can be deleted. The results displayed in the table follow directly from the previous results. After this, everything is repeated for various sample sizes (n=30, 100, 1000, 2000). This is largely unnecessary. I would suggest keeping one of them and perhaps putting the rest in an Appendix, and just commenting upon the results worth mentioning in the body of the text.