Review of: "Intersections of Statistical and Substantive Significance Under a True and False Null Hypothesis"

Hening Huang

Potential competing interests: No potential competing interests to declare.

The author examined the distribution of the*p*-values produced by simulated two-sample *t*-tests. "*The author hopes the empirical sampling distributions in this paper help students, applied researchers, and science writers to properly understand and appreciate the value of statistical significance for scientific inference and decision-making with small sample sizes (n < 1,000) in the face of uncertainty*." While the author has good intentions, the methodology used and results presented do not provide a complete picture of the *p*-values. In particular, an important behavior of*p*-values is missing from the analysis. As a result, the results and analysis are misleading and do not contribute to a proper understanding of *p*-values.

The author's simulations under the true null hypothesis (i.e. μ*T* – μ*C*=0) show that the *p*-values are uniformly distributed regardless of the sample sizes *n*=15, 64, and 500. This is expected and consistent with Wang et al. (2019). However, what is missing from the analysis is how the *p*-values are distributed as a function of the sample size when the null hypothesis is not true, i.e., a true effect size exists. Notably, the author only conducted simulations at a medium effect size (0.50) at *n*=64. This simulation cannot help explore important behavior of the*p*-value when only tiny true effect sizes are present. If the author had conducted simulations with tiny effect sizes (e.g., 0.1, which the author defines as no effect) at *n*=15, 64, 500, results would show that the percentage of significant*p*-values (i.e., false significance rates) would increase with increasing sample size. This is a well-known behavior of *p*-values, called "*N*-chasing" (Stansbury 2020). "*N*-chasing" is a very effective way of *p*-hacking. In principle, *p*-values decrease monotonically as sample size increases. So, a large sample will always produce a smaller *p*-value even if the effect size is very small and meaningless. However, in reality, there is nothing to stop scientists from using large samples, and in fact, large samples are preferred whenever possible in any study. Therefore, the problem of *p*-hacking via "*N*-chasing" cannot be solved unless the*p* value-based hypothesis testing is abandoned.

In summary, the author should provide complete information about*p*-values, including their practical meaning, flaws, and problems. In addition, the author is referred to two extensive discussions on *t*-tests and *p*-values on ResearchGate: https://www.researchgate.net/post/Does_the_two-sample_t-test_provide_a_valid_solution_to_practical_problems

and https://www.researchgate.net/post/Why_do_many_scientists_want_to_abandon_NHST_and_p-values_but_many_statisticians_dont

Stansbury D 2020 p-Hacking 101: N Chasing *The Clever Machine* https://dustinstansbury.github.io/theclevermachine/p-

hacking-n-chasing