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

Felix Bittmann[1]

1 Leibniz Institute for Educational Trajectories

Potential competing interests: No potential competing interests to declare.

The paper is interesting, and I see this as mainly for students and beginners in statistics who are learning about NHST and p-values. The graphs can be helpful to understand the relation between what can happen in random samples and data, yet the main benefit for more experienced researchers is small. In the conclusion, potentially wrong and misleading recommendations are given, which is a big problem and needs to be corrected. I think the author can highlight with this paper how p-values and effect sizes might be misleading, yet this is not really novel.

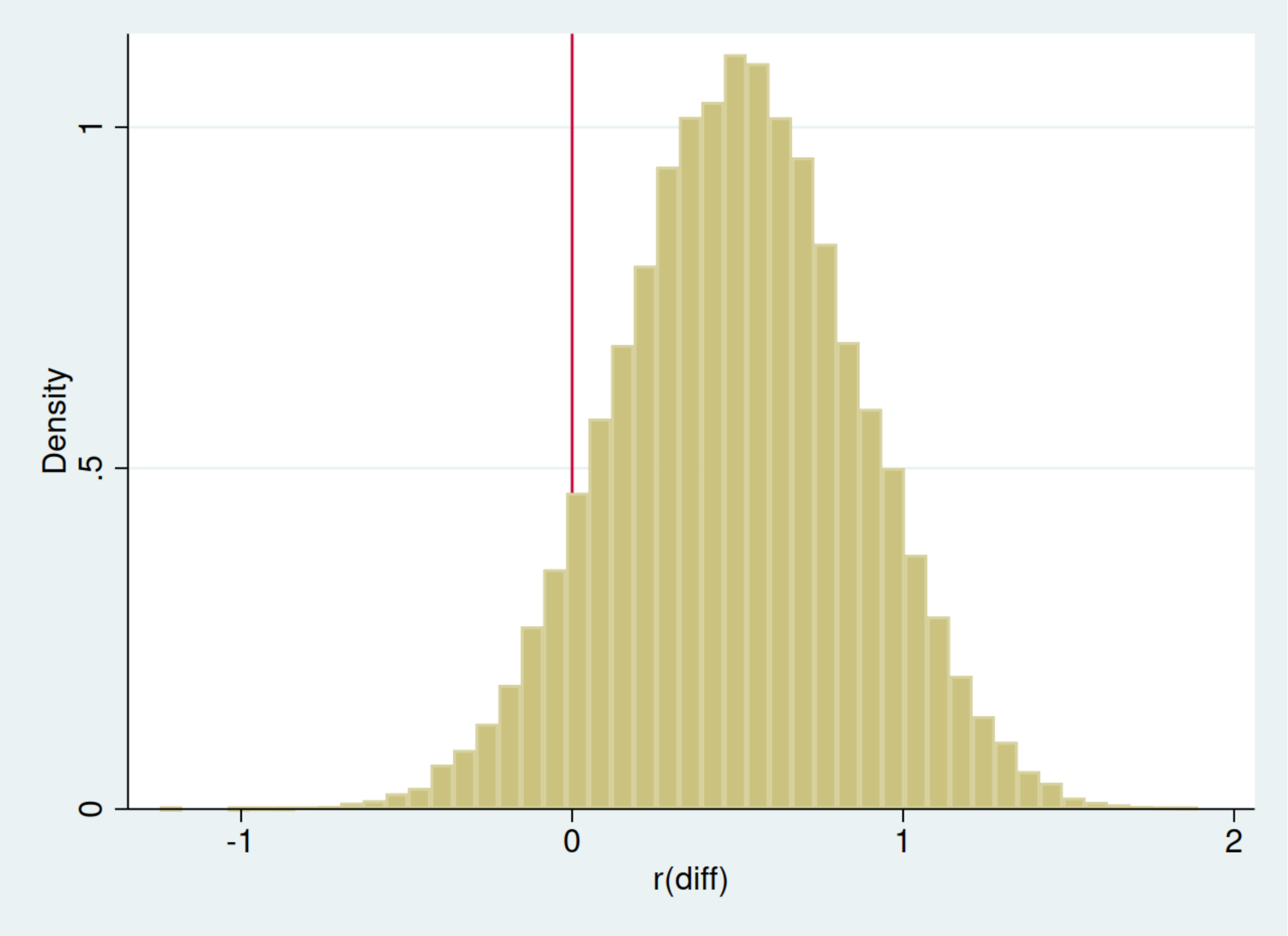
Minor points:

- The introduction would profit from a general (short) explanation of how Null hypothesis significance testing works in general. A definition of what a p-value is is also missing.
- Under „Statistical significance“, the sampling procedure for the simulation is described, yet here I think the main explanation for the approach is lacking. It should be emphasized that here, a simulation is done where a distribution in *accordance with the null* is enforced. This means that distributions are generated that show the distribution of differences / p-values under the assumption that no true effect is present, which might help researchers to build intuition about how strong effects can be in a true-null scenario.
- For the simulation details, no justification is given. Why are the sample sizes 15, 64, 500, and 1000? For the number of simulations, 1000, the same holds. Personally, I think that 1000 is too low. The distributions in some graphs should be more normal. I recommend giving a better justification. 10K or even 50k simulations are recommended in my view.
- Why is the p-value set to 5%? Why not 10 or 1?
- Why is a medium effect size of 0.50 specified in another scenario?
- I think it's trivial to show F1 and F2; the distributions are expected, and any deviations would be suspicious. Better combine them in a single Figure, potentially as a kernel-density plot.
- As we see in F4, the distribution of p-values is uniform if the null hypothesis is actually true. Again, more simulations would show this even clearer.
- I am not sure if the somewhat arbitrary classification of effect sizes in 4 categories (shown in F6) is even necessary. Why coarsen the original continuous values?
- I think F7 and T1 provide good arguments why Cohen’s classification can be misleading and how p-values actually work, exactly as intended. 5% are “significant”, despite the null being true.
- The author proceeds with the larger group sizes. The conclusions are exactly the same as before, but only Cohen’s

classification has fewer positives, as expected.

- From page 20 on, the “true” effect of 0.50 is introduced. Again, many graphs can be simplified or removed. Especially interesting is F19, where the distribution of p-values under a false-null is shown. The distribution is as expected.
- Later, the author shows a power computation (F22). Why? Before, statistical power was neither introduced nor discussed at all. This appears out of the blue. I would remove this or devote more information to the concept of statistical power. Probably, this paper has different goals in mind.
- I do not agree with some parts of the discussion. The author writes “Whatever alpha value is desired as the level of statistical significance, under a true null hypothesis, the probability of a statistically significant p-value does not increase with increasing sample size” — yup, that is how p-values are designed. The main question remains: what is the relation between p-values, measures of effect size, and “true” results. The author states that p-values become useless when sample sizes increase (“Researchers can ignore statistical significance…”). This might be true, but what is the main conclusion? As the author states, the problematic aspects of p-values have been recognized in the past, and measures of effect sizes are clearly helpful and relevant.
- “With graphs and a few numbers, this paper showed that statistical significance is a viable decision tool when working with small sample sizes (e.g., n < 1,000) and testing for differences in means with independent samples t-tests.” This sentence is highly problematic. No, as explained in the introduction before, p-values are not viable decision tools! P-values can be highly misleading. This sentence could be read as: “Well, if your sample size is small, just report and interpret the p-value and you are done.” This is a very bad idea.

Unfortunately, the author does not repeat the 0.50 effect size simulation with N=15 per group, so I have done this. The result of 50k simulations is as follows:



Note the red line, which marks 0. Despite the moderate effect size, about 4,300 of 50k samples have a*negative* mean difference! And 36 (0.072%) of all simulations have a negative sign *and* a p-value that is smaller than 0.05! If the sample size is small, both p-values and effect sizes can still be highly misleading if you are unlucky with your sample. The main message is, and this is beyond this paper: small samples suck and need to be avoided for almost all research projects as the power is low and very misleading findings can arise just due to random error! The current paper does not acknowledge this at all and gives potentially harmful advice. I think if your sample is very small, neither p-values nor CIs nor effect sizes measures can help you make confident assertions about reality. For some more discussion of the issue, refer to https://journals.sagepub.com/doi/10.1177/0193841X16655665

“With the total sample size of n = 1,000, the utility of statistical significance as a screening tool was lost. Researchers can ignore statistical significance and decide whether trivial (non-effect sizes according to Cohen’s criteria) are meaningful or substantively significant in their discipline.” This I can somewhat agree with. If sample sizes are indeed large, p-values are not a simple and binary tool to decide whether a result is “true” or not.