Significance Testing Overestimates Effect Size When Power is Low

Nathan (Nat) Goodman

January 16, 2019

Replication researchers cite low power and inflated effect size estimates as major causes of replication failure. It turns out these are two sides of the same coin linked by significance testing. Requiring significant p-values forces underpowered studies to overestimate effect size. The overstimate is large, more than 2x, under conditions typical in social science research. The bias is inherent in the math; it's not due to p-hacking or other malfeasance. One solution is to increase power. A cheaper and easier solution is to abandon significance testing.

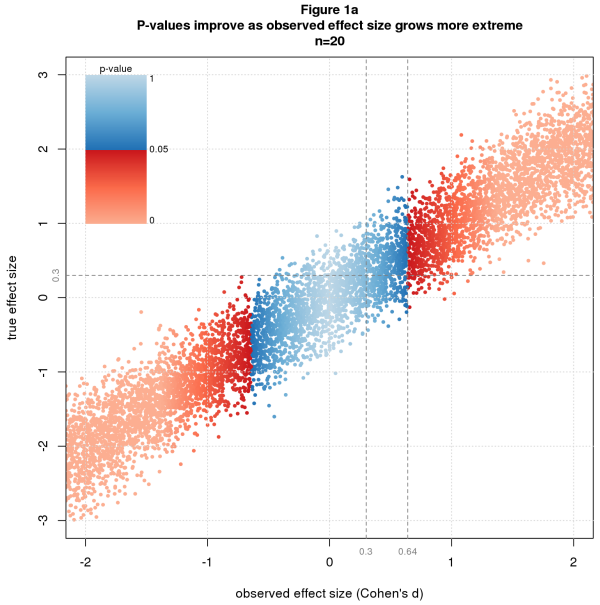
The basic argument is simple (the Supplement has the gory details). (1) The p-value you get when you do a study depends on the effect size you observe; the true effect size plays no role. (2) Bigger[1] observed effect sizes give better p-values. (3) There's a smallest observed effect size with significant p-value. (4) This minimum significant effect size is a *critical* value that cleanly separates nonsignificant and significant observed effect sizes. (5) When the critical effect size exceeds the true effect size, significance testing must overestimate the effect size, since to get a significant p-value, the observed effect size must be even larger.

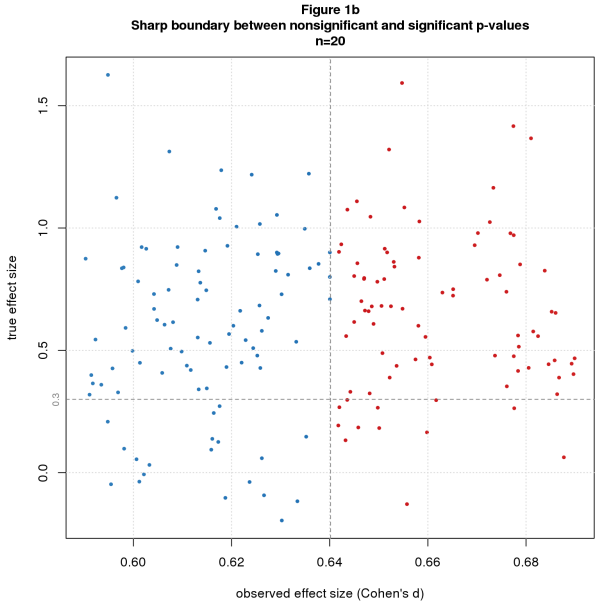
Observed effect size determines p-value. There's no magic: a simple mathematical formula converts effect size into p-value. This is common knowledge among statisticians yet seems to be forgotten in the replication crisis literature and is rarely explained to statistics users. I hope this post will give readers a simple, visual understanding of this point. Shravan Vasishth offers another good explanation in his excellent TRN post and related paper.

For a sample size of 20, the critical effect size is about 0.64. From what I read in the blogosphere, the typical true effect size in social science research is 0.3. For this effect size, the overestimate is at least 2.13x (0.64/0.3).

Figures 1a-b plot true vs. observed effect size for simulated data with points color-coded by p-value. The first plot shows the big picture: bigger effect sizes have better p-values. The second plot zooms into the region around the critical value to show the sharp boundary between nonsignificant and significant effect sizes.

[1] In the text, I assume effect sizes *>0* so I can use terms like "bigger" and "smaller" rather than "more extreme" and "less extreme".





Power is very low for this scenario, about 15%. I used to think this meant there was a 15% chance of observing a significant effect of 0.3 but now know this is incorrect. With *n=20*, 0.3 is always nonsignificant (*p=0.35*). The 15% is the probability of observing an effect size greater than 0.64.

What sample size is enough? At a bare minimum, it should be big enough for the expected true effect size to be significant: 87 in this case. This raises power to about 50%. That's still not good enough for my taste. If I'm going to do all the work of running a study, I want better than 50:50 odds of success. I'd set power to 95%. This requires a sample of 290.

Another way to increase power is to increase the true effect size. Good empiricists excel at finding study conditions that amplify the phenomenon of interest. This is an essential part of empirical research in many scientific disciplines. For example, in biology, basic researchers work with "model systems" such as mice, rather than people, to more readily study phenomena of interest. A modest increase in effect size would help the sample size a lot: an effect of 0.5 (still "medium" in Cohen's d vernacular) would need a sample of 105 instead of 290.

I'm not recommending that social science researchers reshape their studies this way. Bigger studies are more expensive. They're also harder to run and may require more study personnel and study days, which may increase variability and indirectly reduce the effect size. Increasing effect size through artifical study scenarios may further reduce the ability to generalize from lab to real world. All in all, it's not clear that the net effect is positive.

A cheaper and easier solution is to abandon significance testing. The entire problem is a consequence of this timeworn statistical method. Looking back at Figure 1a, observed effect size tracks true effect size pretty well. There's a lot uncertainty, of course, but that seems an acceptable tradeoff for gaining unbiased effect size estimates at reasonable cost.

Comments Please!

Please post comments on Twitter or Facebook, or contact me by email natg@shore.net.