Significance Testing Overestimates Effect Size When Power is Low

Nathan (Nat) Goodman

January 1, 2019

Significance testing overstimates effect size when sample and true effect sizes are small, in other words, when power is low. The overstimate can be considerable: 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 investigator malfeasance. The only solutions are to increase sample size, switch to problems with bigger true effect size, or abandon significance testing.

The basic argument is simple. The p-value you get when you do a study depends on the observed effect size, not the true effect size. Limiting attention to non-negative effect sizes, bigger observed effect sizes give better p-values. With a dollop of math, it follows that there's a smallest observed effect size with a significant p-value. Further, this minimum significant effect size is a critical value that cleanly separates non-signifant and signifant observed effect sizes. For $n = 20$, the critical observed effect size is about 0.64. To get a significant p-value, your must observe an effect size bigger than (or equal to) this. The true effect size doesn't matter.

Comments Please!

Please post comments using the GitHub Issue Tracker.

Test footnotes

Here is a footnote reference,[1] and another.[2]

This paragraph won't be part of the note, because it isn't indented.

[1] Here is the footnote.

[2] Here's one with multiple blocks.

Subsequent paragraphs are indented to show that they belong to the previous footnote.

```
{ some.code }
```

The whole paragraph can be indented, or just the first line. In this way, multi-paragraph footnotes work like multi-paragraph list items.