Covered in P: The consequences of null hypothesis significance testing on point and interval estimates

Raphael T. Gerraty[1], Matti Vuorre[1]

[1] Columbia University, Department of Psychology

Author note

The authors declare no conflicts of interest.

Correspondence concerning this article should be addressed to Raphael T. Gerraty, Columbia University, Psychology Department, 406 Schermerhorn, 1190 Amsterdam Avenue, New York, NY 10027. E-mail: https://github.com/neurostorm

Abstract

Confidence intervals (CI) do not allow post-data inference on parameter values.

Keywords: confidence interval, NHST

Word count: Short and sweet.

Covered in P: The consequences of null hypothesis significance testing on point and interval estimates

Introduction

Scientists, psychological or otherwise, routinely use null hypothesis significance testing procedures (NHSTP) to move from data to conclusions—a practice thats applicability has been debated since its inception. Recently, concerns about the replicability and reliability of empirical findings (Collaboration, 2015) have underlined the concerns about NHSTP as *the* valid form of statistical inference (Cumming, 2014; Gelman & Loken, 2014).

One response to the growing concerns regarding the reliability of NHSTP has been an appeal to effect size and interval estimation in addition—or as replacement—to NHSTP test statistics (Cumming, 2014). For example, many journals in psychology and neuroscience now ask authors to include *confidence intervals* (CI) with their test statistics. These confidence intervals are intended, by practitioners, to communicate unbiased estimates of likely ranges of parameter values, although the common computations simply result in ranges of parameter values that would not be rejected by the very statistical test the CI is supposed to supplement. This approach is problematic from at least two perspectives:

- CIs do not support probability statements about parameters
- CIs are designed (Neyman, 1957) to be a procedure of generating a set of intervals that, as a whole set, contain the true parameter value X% of the time
  - This property is severely compromised by the current practice of using CIs as a post-data inferential tool, as we show in this paper

A Confidence Interval

While the first of these interpretations is inaccurately held by many researchs in psychology and other fields (citing stuff), In this paper, we report an unappealing property of confidence intervals. Because the claim of confidence intervals is to have a coverage proportion of the true parameter value equal to the nominal value (usually 95%), it is crucial that this claim is substantiated in its long-run property for a CI to be what it claims to be (not a *confidence* interval.) We show that using confidence intervals *in addition* to P values leads to an undesirable distortion of the coverage proportion. This is a direct result of the more general problem with interpreting any particular *obtained* CI in terms of frequency coverage, but we feel the specific case of significance thresholding on this interpretation is worth describing, given the per

P stains the nominal coverage proportion

Methods

We performed a simulation study…

Results

```
[1] 0.6457245
```

Discussion

Here we show a pervasive bias in the paramaters and intervals passing a null hypothesis significance threshold. We don't know if this result is well known to statisticians, but from the perspective of practitioners, we found it suprising. This paper was motivated in part by the discussions with colleagues who were equally suprised by the biases induced by hypothesis testing, especially on interval estimation. The "significance filter" has been discussed previously (Gelman, 2011), but to our knowledge there have been no discussions of the effect of this filter on the frequency properties of confidence intervals.

We note that, while the issues discussed in this paper are related to questionable research practices as well as known issues in null hypothesis testing such as alpha inflation due to multiple comparisons, the biased point estimates and interval coverage for significant results we discuss here are present in expectation even for single tests. Thus this bias will be more severe for significant results which have been filtered through such processes, but…

References

Collaboration, O. S. (2015). Estimating the reproducibility of psychological. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Gelman, A. (2011, October 9). *The statistical significance filter. Statistical modeling, causal inference, and social science.* Retrieved December 9, 2015, from http://andrewgelman.com/2011/09/10/the-statistical-significance-filter/

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. doi:10.1511/2014.111.460

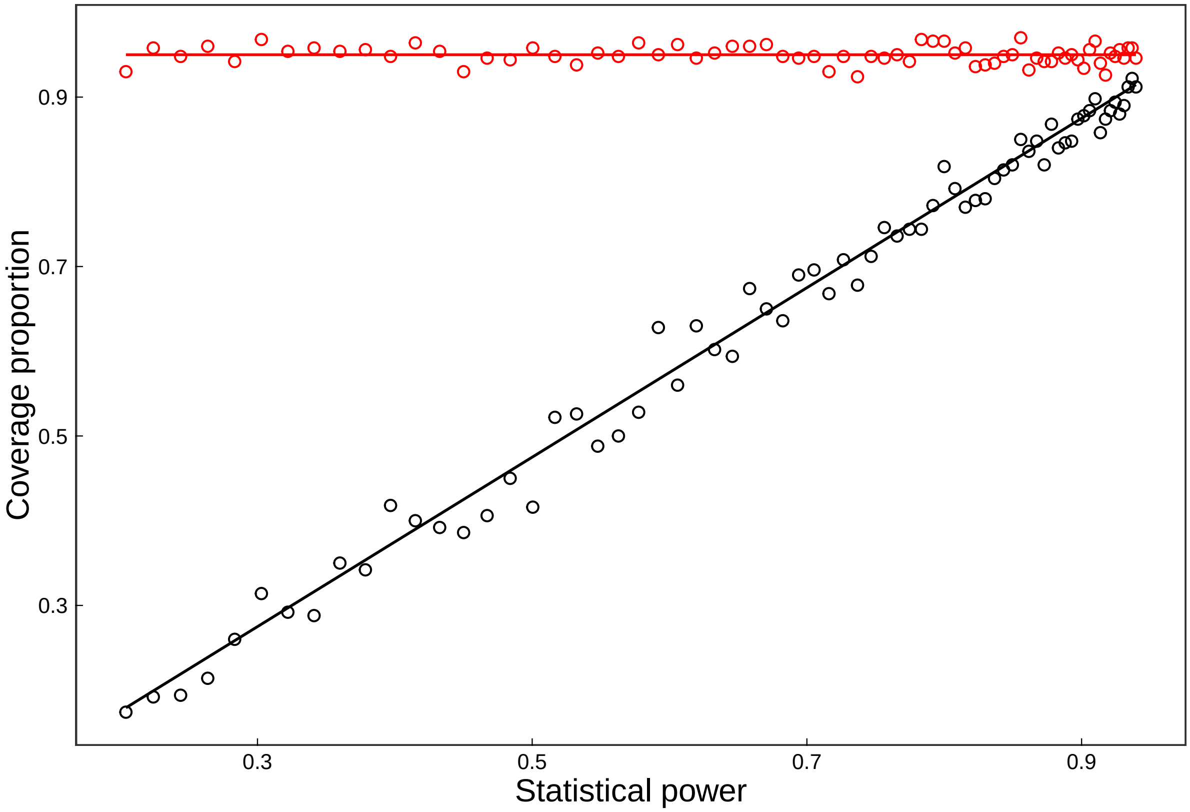

Figure 1. Coverage proportion of confidence intervals versus statistical power. The red line represents the nominal coverage proportion (95%), black line is the actual coverage proportion when CIs are conditioned on the result being significant.