Covered with 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 with 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 whose applicability has been debated since its inception. Recently, concerns about the replicability and reliability of empirical findings (Collaboration, 2015) have reignited questions about the validity of NHSTP as they are implemented and interpreted in practice (Gelman & Loken, 2014; Krantz, 1999).

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: "The problems that pervade NHST are avoided by the new statistics—effect sizes, confidence intervals[. . .]" (Eich, 2014, p. 3). "Reports of values of r must, like reports of means, be accompanied by appropriate confidence intervals." (Lindsay, 2015, p. 2) We agree on the benefits of interval estimation over computing p-values and embrace these recommendations, but fear that turning to CIs—of the particular variety that are most commonly computed—does not. . . (get rid of p in our intervals).

These confidence intervals are intended, by practitioners, to communicate 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. First, CIs do not support probability statements about parameters. Second, a single observed CI does not have coverage properties apart from either including or excluding the true parameter value. A (infinite) sequence of CIs, or the generating procedure, however, does have the property of containing the true parameter on X% of occasions a single CI is realized

from the sequence. This property is severely compromised by the current practice of using CIs as a post-data inferential tool, as we show in this paper.

While the first of these perspectives is misunderstood by many researchers in psychology and other fields (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), we focus here on the second, which is more subtle and even less appreciated among practitioners. A search of [searching for textbooks in a principled way] shows that ~X% [surely some depressing number] endorse the view that the frequency coverage or "confidence" of a CI refers to the specific values obtained for a *particular* interval. This is an incorrect interpretation to begin with, but given its widespread acceptance it is important to describe specific examples in which such an interpretation produces misleading results.

In this paper, we report an unappealing property of obtained confidence intervals for results that pass a significance test. 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 pervasive reporting and interpretation of such intervals following NHSTP in a wide range of research areas. We demonstrate analytically and with simulations that the proportion of confidence intervals that are significant *and* contain the true parameter in question is a function of the power of a statistical procedure. Given the low power of many psychological studies (cite), and the widespread belief that obtained confidence intervals' nominal coverage provides valid information about parameters of interest (see blank for recommendations along this line), this demonstration may be of use to practitioners who feel confused or even strongly about the benefits of confidence interval estimation in NHSTP.

Methods

We performed a simulation study...

Results

[1] 0.6457245

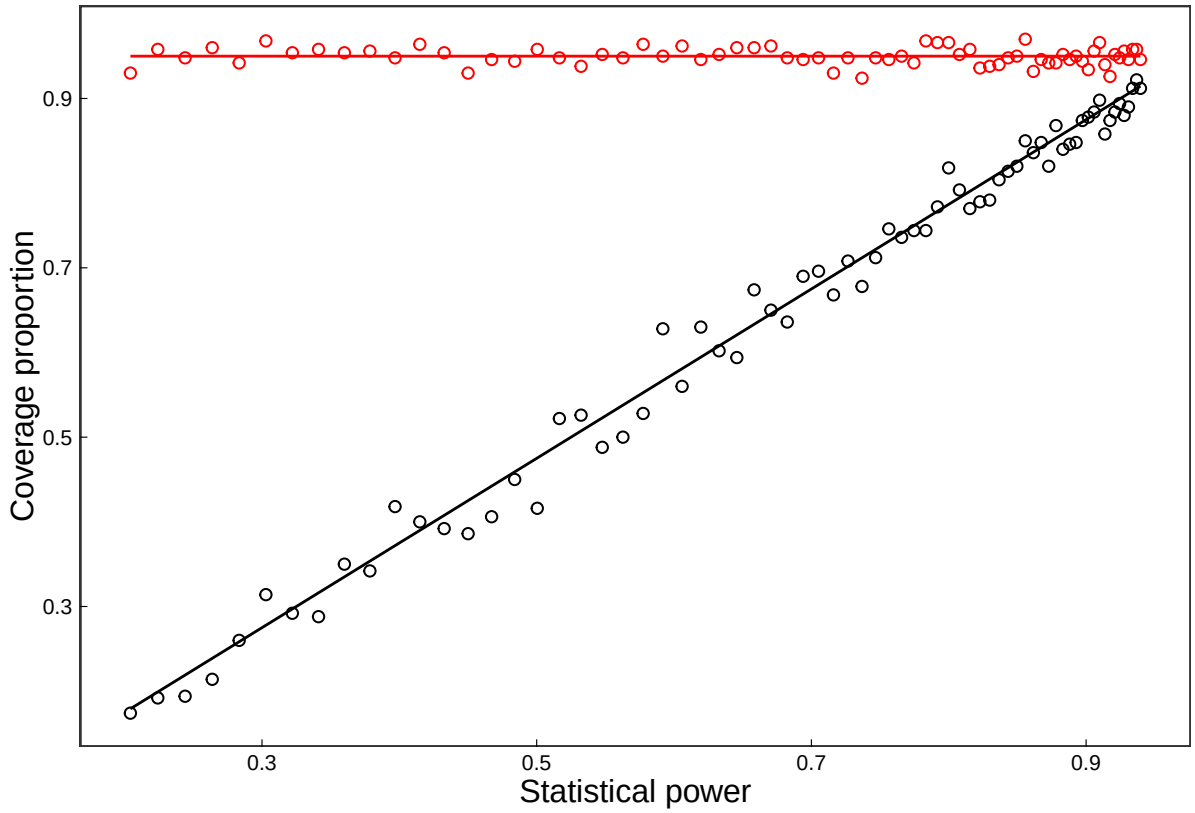

Figure 1. Coverage proportion of confidence intervals versus statistical power. The red line represents the nominal coverage proportion (95%), black line is the actual coverage proportion when CIs are conditioned on the result being significant.

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

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(1), 3–6. doi:10.1177/0956797613512465

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

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. doi:10.3758/s13423-013-0572-3

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*(448), 1372–1381. doi:10.2307/2669949

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. doi:10.1177/0956797615616374