Journal of Memory and Language 103 (2018) 151–175

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The statistical significance filter leads to overoptimistic expectations of replicability

Shravan Vasishth[a,*], Daniela Mertzen[a], Lena A. Jäger[a], Andrew Gelman[b]

[a] Department of Linguistics, University of Potsdam, Potsdam, Germany
[b] Department of Statistics, Columbia University, New York, USA

ARTICLE INFO

Keywords:
Type M error
Replicability
Surprisal
Expectation
Locality
Bayesian data analysis
Parameter estimation

ABSTRACT

It is well-known in statistics (e.g., Gelman & Carlin, 2014) that treating a result as publishable just because the p-value is less than 0.05 leads to overoptimistic expectations of replicability. These effects get published, leading to an overconfident belief in replicability. We demonstrate the adverse consequences of this statistical significance filter by conducting seven direct replication attempts (268 participants in total) of a recent paper (Levy & Keller, 2013). We show that the published claims are so noisy that even non-significant results are fully compatible with them. We also demonstrate the contrast between such small-sample studies and a larger-sample study; the latter generally yields a less noisy estimate but also a smaller effect magnitude, which looks less compelling but is more realistic. We reiterate several suggestions from the methodology literature for improving current practices.

Introduction

Imagine that a reading study shows a difference between two means that has an estimate of 77 ms, with standard error 30, that is, with $p = 0.01$. Now suppose instead that the same study had shown an estimate of 40 ms, also with a standard error of 30; this time $p = 0.18$. The usual reporting of these two types of results—either as significant and therefore "reliable" and publishable, or not significant and therefore either not publishable, or seen as showing that the null hypothesis is true—is misleading because it implies an inappropriate level of certainty in rejecting or accepting the null. Indeed, it has been argued that this routine attribution of certainty to noisy data is a major contributor to the current replication crisis in psychology and other sciences (Amrhein, Korner-Nievergelt, & Roth, 2017; Open Science Collaboration, 2015). For recent examples from psycholinguistics of replication difficulties, see Nieuwland et al. (2018) and Kochari and Flecken (2018). The issue is not just the high frequency of failed replications, but also that these failed replications arise in an environment where routine success (defined as $p < 0.05$) is expected. We will refer to this $p < 0.05$ decision criterion for publication-worthiness as the *statistical significance filter*. We will demonstrate through direct replication attempts one well-known adverse consequence of the statistical significance filter (Gelman, 2018; Lane & Dunlap, 1978), that it leads to findings that are positively biased. We want to stress that none of the statistical points made in this paper are new (for similar arguments, see Button et al., 2013; Dumas-Mallet, Button, Boraud, Gonon, & Munafò, 2017; Frank et al., 2017; Goodman, 1992; Hedges, 1984; Ioannidis, 2008, among others). However, we feel it is necessary to demonstrate through direct replication attempts why significance yields no useful information when statistical power is low. The fact that underpowered studies continue to be treated as informative suggests that such a demonstration is needed.

We assume here that the reader is familiar with the null hypothesis significance testing (NHST) procedure as it is used in psychology today. NHST can work well when power is relatively high. But when power is low, published studies that show statistical significance will have exaggerated estimates (see Appendix A for a formal argument). The effect of low power is demonstrated in Fig. 1 using simulated data: for a low-power scenario, the estimates from repeated samples fluctuate wildly around the true value, and can also have the wrong sign. Whenever an effect is significant, it is necessarily an overestimate. Gelman and Carlin (2014) refer to these overestimates as Type M(agnitude) errors (when the sign of the effect is incorrect, Gelman and Carlin call this Type S(ign) error). These overestimates occur because the standard error is relatively large in low-power situations; the wider the sampling distribution of the mean, the greater the probability of obtaining extreme values. By contrast, when power is high, the estimates under repeated sampling tend to be close to the true value because the standard error is relatively small.

Fig. 1 illustrates another important point: when power is high, the estimates have much narrower 95% confidence intervals. We will express this by saying that high-powered studies have higher *precision* than low-powered studies. We borrow the term precision from Bayesian

* Corresponding author.
E-mail address: vasishth@uni-potsdam.de (S. Vasishth).

https://doi.org/10.1016/j.jml.2018.07.004
Received 17 January 2018; Received in revised form 10 July 2018
Available online 30 August 2018

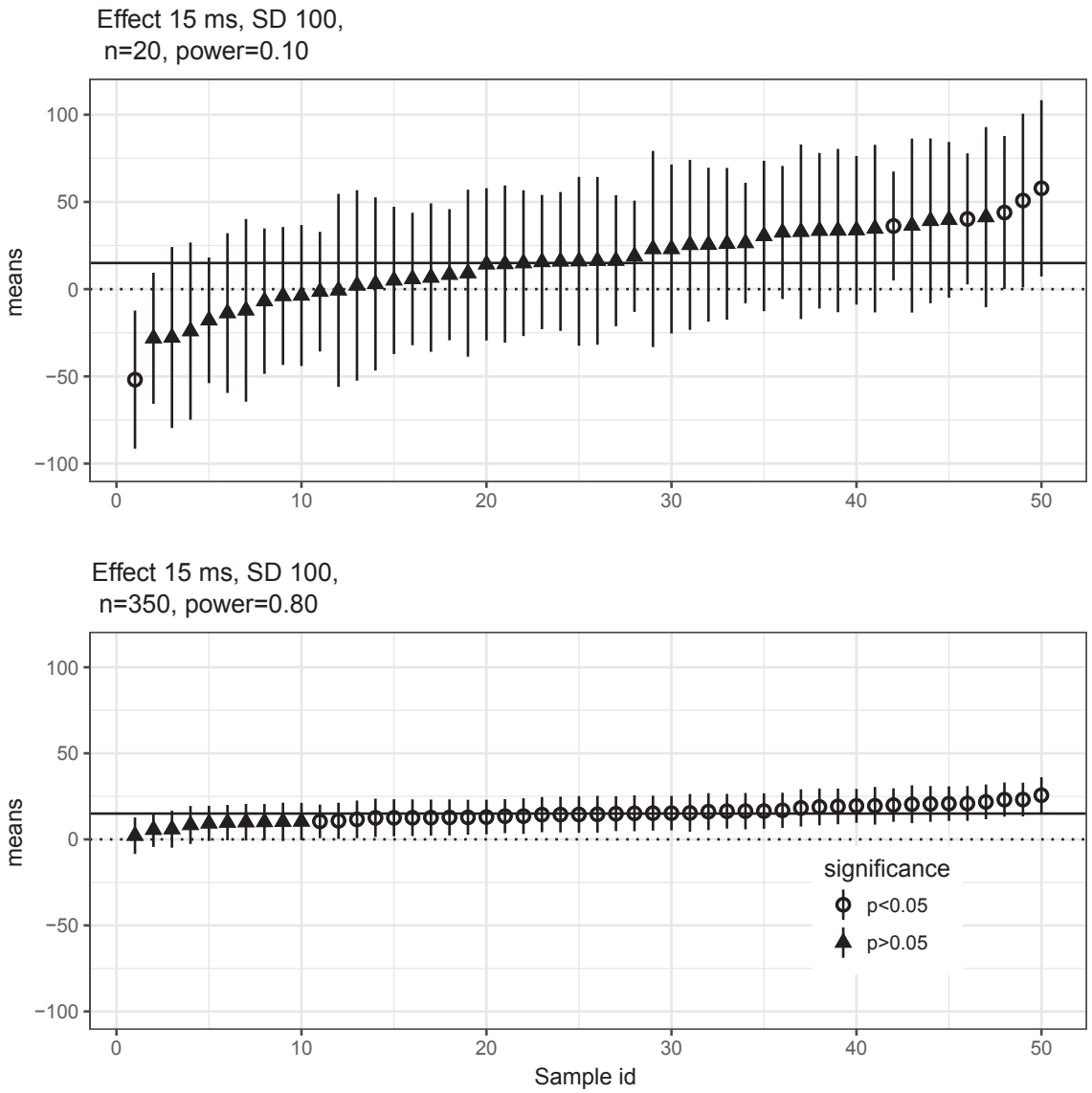

Fig. 1. A demonstration of Type M error using simulated data. We assume that the data are generated from a normal distribution with mean 15 ms and standard deviation 100 ms. The true mean is shown in each plot as a solid horizontal line. When power is low, under repeated sampling, whenever the estimates of an effect come out significant, the values are overestimates and can even have the wrong sign. When power is high, significant and non-significant effects will be tightly clustered near the true mean.

statistics, where it has a specific meaning: the inverse of the variance. Here, we are using the term precision to stand for the uncertainty about our estimate of interest (the sample mean, or a difference in sample means). This uncertainty is expressed in frequentist statistics in terms of the standard error of the sample mean. The standard error decreases as a function of the square root of the sample size; hence, if power is increased by increasing sample size, standard error will decrease.

Many researchers, such as Cohen (1962), and Gelman and Carlin (2014), have pointed out that a prospective power analysis should be conducted before we run a study; after all, why would one want to spend money and time running an experiment where the probability of detecting an effect is 30% or less? In medical statistics, prospective power analyses are quite common; not so in psycholinguistics. Suppose that we were to follow this practice from medical statistics and conduct a prospective power analysis based on the effect sizes reported in the literature. Gelman and Carlin (2014), and many others before them, have pointed out that this can lead to an interesting problem. Whenever an effect in an underpowered study comes out significant, it is *necessarily* an overestimate. In fields where power tends to be low, these overestimates will fill the literature. If we base the power analysis on the published literature, we would conclude that the effects are large. A formal power analysis based on such exaggerated estimates is bound to yield an overestimate of power, and we can incorrectly convince ourselves that we have an appropriately powered study.

In psycholinguistics, usually we do no power analyses at all. We just rely on the informal observation that most of the previously published results had a significant effect. From this we conclude that the effect must be "reliable," and therefore replicable.

Although the above observations about power and replications are well-known in statistics and psychology (see the discussion in Chambers, 2017; Wasserstein & Lazar, 2016), they are not widely appreciated in psycholinguistics. Our goal in this paper is to demonstrate—not via simulation but through actual replication attempts of a published empirical result—that relying exclusively on statistical significance to decide whether or not a result is newsworthy leads to misleading conclusions.

We show through a case study that small-sample experiments can easily deliver statistically significant results that overestimate the true effect and are non-replicable. For this case study, we chose a paper by Levy and Keller (2013) that investigated expectation and locality effects