A Cautionary Note on Estimating Effect Sizes

Don van den Bergh*[1], Julia M. Haaf[1], Alexander Ly[1,2], Jeffrey N. Rouder[3], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam
[2]Centrum Wiskunde & Informatica
[3]University of California Irvine

Abstract

An increasingly popular approach to statistics is to focus on estimation and to forgo hypothesis testing altogether. Through an example, we show that estimates and confidence of effect sizes are overestimated the null is ignored, and when it is a plausible description of the data. We illustrate how this overestimation can be avoided by incorporating the plausibility of the null into the estimation process.

Consider the following scenario: Your colleague has just conducted an experiment for a Registered Report. The analysis yields $p < 0.05$ and your colleague believes that the null hypothesis can be rejected. In line with recommendations both old (e.g., Grant, 1962; Loftus, 1996) and new (e.g., Harrington et al., 2019; Cumming, 2014) you convince your colleague that it is better to replace the p-value with an estimate of the effect size and a 95% confidence interval (but see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). You also manage to convince your colleague to plot the data. Instead of simply reporting $p < 0.05$, the statistical analysis in the report is now more informative. The result is shown in Figure 1. In the text of the paper, the result is summarized as Cohen's $d = 0.30$, CI $= [0.02, 0.58]$, in line with guidelines for reporting statistics (e.g., the guidelines of the Psychonomic Society[1], or those of Psychological Science[2]). Given the results shown in Figure 1, what is a reasonable point estimate of the effect size? A straightforward answer is "0.30" which makes intuitive sense from an estimation perspective. However, your colleague now tells you about the nature of the experiment: plants grow faster when you talk to them.[3] Suddenly, a population effect size of "0" also appears plausible. Any

*Correspondence concerning this article should be addressed to: Don van den Bergh, University of Amsterdam, Department of Psychological Methods, Nieuwe Achtergracht 129B, 1018VZ Amsterdam, The Netherlands. E-Mail should be sent to: donvdbergh@hotmail.com.

[1]https://www.springer.com/psychology?SGWID=0-10126-6-1390050-0

[2]https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STAT

[3]Specifically, imagine your colleague took 100 plants and measured their growth three times during two weeks. The first week 50 plants were randomly selected and spoken to while the other served as control. The next week, the roles reversed and the previously spoken to plants served as controls while the control plants were now talked to. The quantity of interest is the difference in growth between the weeks. This example is inspired by Berger and Delampady (1987).

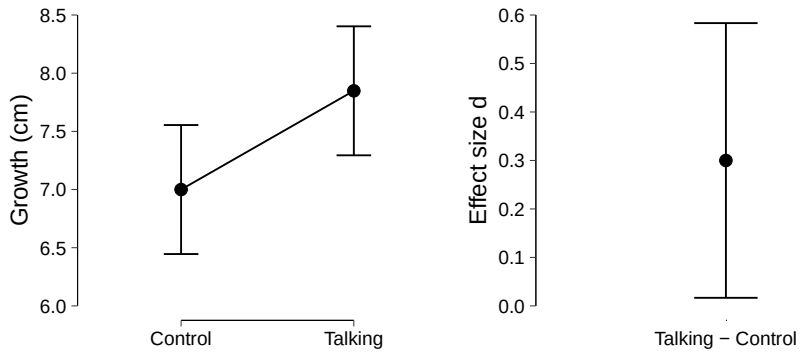

Figure 1: The left panel shows a descriptives plot with the mean and 95% confidence interval of the simulated plant growth. The right panel shows an estimate of the effect size, Cohen's d, and a 95% confidence interval.

observed difference may merely be due to the individual differences between the plants.[4]

When are Effect Sizes Overestimated?

Standard point estimates and confidence intervals are based solely on the alternative hypothesis and therefore tend to overestimate effect sizes. This overestimation is caused by the strong assumption that the null hypothesis is irrelevant. However, the null can have high plausibility after seeing the data. This happens if the alternative is a-priori unlikely (e.g., for a null hypothesis like "Talking to plants does not affect their growth."), or when the data are so uninformative that after seeing the data there is substantial uncertainty about which model is best. If the null hypothesis still has a high plausibility after seeing the data, it is obvious that it cannot be ignored, but that is exactly what is done when estimates are only based on the alternative hypothesis. As a result, the estimates are overconfident and, because the null hypothesis would shrink the estimates towards zero, overestimated.

A Spike-and-Slab Perspective

Here, we illustrate the overestimation and a remedy against it by reanalyzing the fictitious data from Figure 1.[5] We consider the spike-and-slab model (Rouder, Haaf, & Vandekerckhove, 2018; Clyde, Desimone, & Parmigiani, 1996; Mitchell & Beauchamp, 1988), which consists of two components. The first component corresponds to the position that talking to plants does not make them grow faster or slower ($d = 0$). The second component corresponds to the position that speaking to plants does influence their growth ($d \neq 0$). A-priori both components are equally likely; the prior probability each component is $1/2$. Here we view the spike-and-slab as a single model, although it can also be viewed as a form of Bayesian model averaging (see Box 1 for more details). Typically, an estimate of the effect size is solely based on the second component, which yields

[4] Unless you talk out loud, with consumption, and the plant is near.

[5] R code for the analysis is available at https://osf.io/uq8st/.

a point estimate and an uncertainty interval (for frequentists, $d = 0.30$, 95% CI: [0.02, 0.58]; for Bayesians $d = 0.29$, 95% CRI: [0.02, 0.57]). The spike-and-slab model, however, also considers the possibility that an effect can be exactly zero and thus the estimate is an average of the two components and thus shrunken towards zero.[6] Figure 2 contrasts inference based on one component with inference based on both components by showing their 95% credible intervals and posterior means. The posterior mean and central 95% credible interval of the

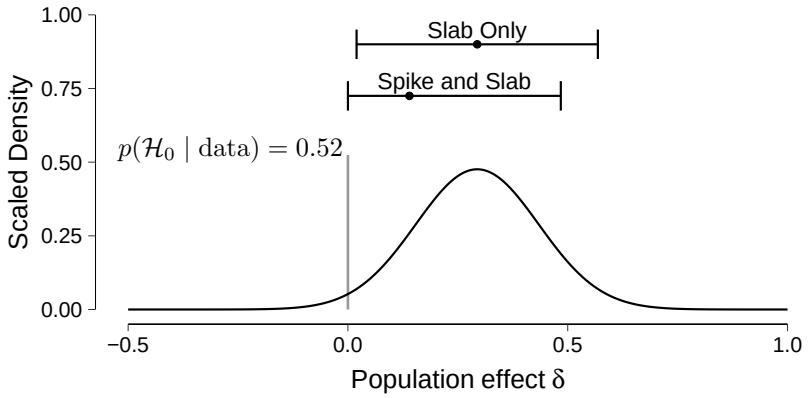

Figure 2: Visualization of the results from the spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\delta = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\mathcal{H}_1 \mid \text{data}) = 0.48$). The grey line represents the posterior probability of the spike (i.e., the effect is exactly 0). The error bars and dots above the density show 95% credible intervals and the posterior mean for a posterior based on only the slab and for a posterior based on the spike-and slab.

spike-and-slab model are shrunken towards 0 compared to the results based on only the slab (i.e., 0.14 (95% CRI: [0.00, 0.48]) vs. 0.29 (95% CRI: [0.02, 0.57])). This shrinkage is caused by the non-negligible probability that the effect is absent. Although 0 appears to be included in the credible interval, this cannot be interpreted directly as evidence against an effect. Instead, the spike-and-slab posterior simultaneously contains evidence for the presence of the effect and an estimate of the magnitude of the effect across the two components. Note that as the posterior probability of the null decreases, the spike-and-slab results approach those of the alternative model.

Discussion

We argue that estimates of effect sizes based only on the alternative hypothesis tend to be overconfident, in particular when a null hypothesis may also describe the data well. Consequently, point estimates and confidence intervals

[6]For the spike-and-slab model, the posterior distribution is constructed in the following manner: $p(\delta \mid \text{data}) = 1\{\delta = 0\}\Pr(\mathcal{M}_0 \mid \text{data}) + p(\delta \mid \text{data}, \mathcal{M}_1)\Pr(\mathcal{M}_1 \mid \text{data})$. Here, $1\{\delta = 0\}$ is the Dirac delta function which represents the spike under the null, Pr denotes probability of a model, and p denotes density related to the magnitude of the effect.

based solely on the alternative overestimate effect sizes. A solution for this overestimation is the spike-and-slab model which explicitly considers the possibility that an effect is exactly 0 (as is also advocated for by Kiers & Tendeiro, 2019; Rouder et al., 2018). Although this idea is not new, the influence of the null is still too often ignored in practice.

All Null Hypotheses are False

The spike-and-slab approach may contrast with the popular estimation mindset, where it is argued that statistical significance should be abandoned in favor of estimation (McShane, Gal, Gelman, Robert, & Tackett, 2019; Valentine, Aloe, & Lau, 2015; Cumming, 2014). An argument to forgo hypothesis testing is that all null hypotheses are false (Cohen, 1990; Meehl, 1978) and therefore there is no need to consider a component that states that an effect is exactly zero. However, a statistical argument to consider point nulls is that they accurately approximate intervals near zero (Berger & Delampady, 1987). Thus, it is irrelevant whether null hypotheses are exactly true. Furthermore, several large-scale replications studies have demonstrated that a near-zero effect size is reasonable in practice, i.e., the null can be a plausible description of the data (e.g., see the meta-analyses conducted by Camerer et al., 2018; Klein et al., 2018; Nosek & Lakens, 2014). Moreover, the spike need not always represent a null of exactly zero; it could also represent a perinull —a region near zero that represents "practically irrelevant" effect sizes. Typically, the perinull spike is a distribution centered around zero with a small variance. Key is that the argument remains unchanged. Since the perinull has most of its mass near zero, estimates will still be shrunk towards zero. Thus the null is not ignored and overestimation is avoided. For a comparison of null and perinull spikes, see Malsiner-Walli and Wagner (2018).

When (not) to Consider the Spike

The main reason to ignore the spike is when overestimation is implausible. This happens in two scenarios: First, when prior research indicates that the prior probability of the spike is negligible, the spike can safely be ignored. For example, if we estimate the popularity of two politicians, a spike that represents that the two politicians are equally popular is so implausible that it can be ignored. Second, when the data provide overwhelming evidence that an effect is non-zero, the results from a spike-and-slab model become virtually identical to those of a slab only model and thus the spike can be ignored. For instance, when the data show that one politician is 10 times as popular as another, the posterior probability of the spike that states the politicians are equally popular is nearly zero. A practical recommendation is to ignore the spike whenever the sample size is larger than 50 and the point estimate (posterior mean or maximum likelihood estimate) is more than three standard deviations away from the spike (Jeffreys, 1939, p. 193–194; Jeffreys, 1980, p. 75). Yet, the spike does not hurt in either of these scenarios.

Conclusion

In sum, we argue that descriptions of effect sizes based on only the alternative hypothesis are overconfident and as a consequence overestimate effect sizes. A remedy for this overestimation is the spike-and-slab model which explicitly accounts for the possibility of the null of no effect. Although this idea is not new, it remains underutilized in practice, and we hope this paper brings more attention to the spike-and-slab approach.

References

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science*. *Nature Human Behaviour*, *2*, 637–644.

Clyde, M., Desimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, *91*(435), 1197–1208.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304–1312.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*, 55–61.

Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.-L. T., ... Hamel, M. B. (2019). *New guidelines for statistical reporting in the journal*. Mass Medical Soc.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 451–453). Amsterdam, The Netherlands: North-Holland Publishing Company.

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. London: MacMillan.

Kiers, H., & Tendeiro, J. (2019). With Bayesian estimation one can get all that Bayes factors offer, and more. *manuscript submitted for publication*. Retrieved from psyarxiv.com/zbpmy

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., ... Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Malsiner-Walli, G., & Wagner, H. (2018). Comparing spike and slab priors for Bayesian variable selection. *arXiv preprint arXiv:1812.07259*.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Atatistician*, *73*(sup1), 235–245.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press.

Zellner, A., & Vandaele, W. (1975). Bayes–Stein estimators for k–means, regression and simultaneous equation models. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* (pp. 627–653). Amsterdam, The Netherlands: North-Holland Publishing Company.

Box 1: The Spike-and-Slab Distribution as Bayesian Model Averaging

The spike-and-slab distribution can be viewed as a single model that consists of two components: the slab, which assumes that the effect is nonzero, and the spike, which assumes the effect is exactly zero. However, the spike-and-slab distribution can also be seen as a form of Bayesian model averaging. From that perspective, the spike and the slab are two individual models. The slab represents the unconstrained model that freely estimates effect size, and the spike represents the constrained model where the effect size is fixed to zero. Next, the results for each model are weighted by the posterior model probabilities and averaged, so that inference can be made using results from both models simultaneously. Such averaging over models provides the best predictive performance (Zellner & Vandaele, 1975, p. 640–641, as described in Zellner & Siow, 1980, p. 600–601; Haldane, 1932, p. 57; Iverson, Wagenmakers, & Lee, 2010; Rouder et al., 2018), and conceptually similar ideas date back much further (Wrinch & Jeffreys, 1921, p. 387; Jevons, 1874/1913). Note that these two perspectives —a two-component model or averaging of two models— differ in semantics but are mathematically equivalent.