Todo list

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 size are overestimated when ignoring the null when the null is a plausible description of the data. Next, we illustrate how this overestimation can be avoided using Bayesian model averaging.

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., Cumming, 2014, NEJM editorial!) you convince your colleague that it is better to replace the p-value with an estimate of 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 $p < .05$, 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 $d = 0.35$, CI $= [0.01, 0.69]$, in line with guidelines for reporting statistics (TODO ref). Given the results shown in Figure 1, what is a reasonable point estimate of effect size? An straightforward answer is "0.35" 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.[1] Suddenly, an effect size of "0" also appears plausible.[2]

ref NEJM editorial

ref apa guidelines, guidlines PBR, Psych Science

*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]Specifically, imagine your colleague took 100 plants and measured their growth during three 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 & Delampady, 1987).

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

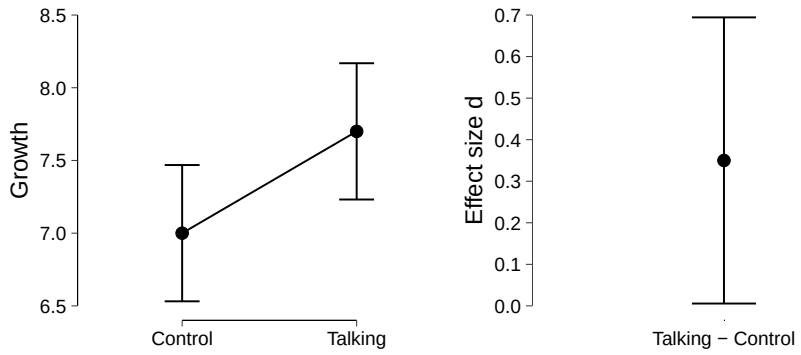

Figure 1: The left panel shows a descriptives plot with the mean and 95% confidence interval of the simulated plant growth. The right panel shows an estimate of the effect size, Cohen's d, and a 95% confidence interval.

When is Effect Size Overestimated?

Point estimates and confidence intervals of effect size based on only the alternative hypothesis tend to overestimate effect size. This overestimation is caused by the strong assumption that the null or perinull hypothesis is irrelevant. However, the null or perinull (ref Tukey) can have high plausibility after seeing the data. This happens when the prior odds in favor of a null effect are large (e.g., for a null hypothesis like "Talking to plants has no effect on their growth."), or when the data are so uninformative that after seeing the data there is substantial uncertainty about which model best describes the data. If the null hypothesis has a high posterior plausibility, 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.

ref Tukey

A Bayesian Model-Averaged Perspective

Here, we illustrate the overestimation and a remedy against it by reanalyzing the simulated data from Figure 1.[3] We consider two hypotheses: The null hypothesis ($\mathcal{H}_0$), speaking to plants does not make them grow faster or slower ($d = 0$), and the alternative hypothesis ($\mathcal{H}_1$), speaking to plants makes them grow faster or slower ($d \neq 0$). We consider both these hypotheses using a paired-samples t-test. Typically, an estimate of effect size is based on solely the alternative hypothesis, which yields a point estimate and an uncertainty interval (for frequentists, $d = 0.35$, 95% CI: [0.01, 0.69]; for Bayesians $d = 0.25$, 95% CRI: [0.05, 0.44]). However, it has been shown repeatedly that averaging over the models considered provides the best predictive performance (Zellner & Vandaele, 1975, pp. 640–641, as described in Zellner & Siow, 1980, p. 600–601; Haldane, 1932, p. 57, Iverson, Wagenmakers, & Lee, 2010, Rouder, Haaf, & Vandekerckhove, 2018), and conceptually similar ideas date back much further (Wrinch & Jeffreys, 1921, p. 387, Jevons, 1874/1913). Accordingly, our pro-

(Comment is too long, see the first page). Something is off with the Bayesian posterior mean in comparison to the frequentist ones. Both Jeff & Julia and Bayesfactor seem

[3]Code for the reanalysis is available at `https://osf.io/uq8st/`.

posed remedy is to average across the null model and the alternative model, weighted by their posterior model probabilities.[4] Figure 2 contrasts inference based on the alternative hypothesis with inference based on the averaged model by showing intervals and posterior means. The model averaged posterior mean

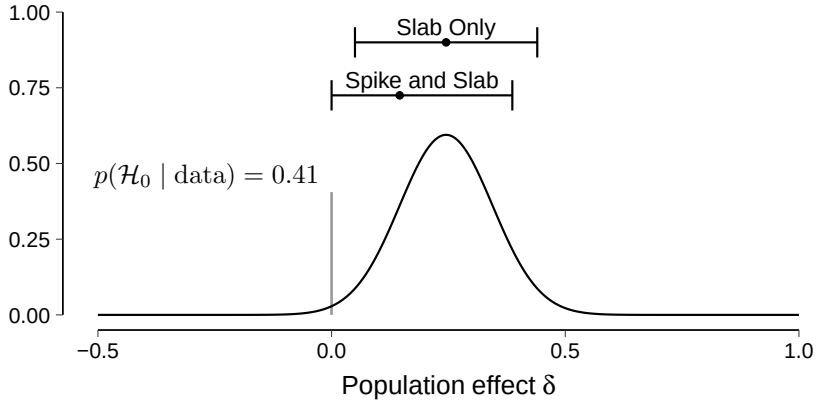

Figure 2: A visualization of model averaging. The black line represents the posterior distribution of effect size given the alternative model (i.e., the slab). The posterior is scaled so that its mode equals the posterior probability of the alternative model. The gray line represent the posterior probability of the null model (i.e., the spike). The error bars and dots above the density show a 95% credible intervals and the posterior mean for both the the slab and the model averaged posterior.

and credible interval are shrunken towards 0 compared to the posterior mean conditional on the alternative hypothesis (0.15 (95% CRI: [0.00, 0.39]) vs. 0.25 (95% CRI: [0.05, 0.44])). This makes sense as the posterior probability of the null hypothesis (0.41) is non-negligible.

Discussion

Here, we argued that estimates of effect size based on only the alternative hypothesis tend to be overconfident, in particular when a null or perinull hypothesis could also describe the data well. Consequentially, point estimates and confidence intervals based solely on the alternative overestimate effect size. A solution for this overestimation is averaging over the null and alternative hypothesis. Although this idea is not new, the influence of the null is still too often ignored in practice.

This approach contrasts 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; Cumming, 2014). Some may argue that all null hypotheses are false (Cohen, 1994). However, there are statistical motivations to consider a point null (Berger & Delampady, 1987) and

Double check ref Cohen

[4]For effect size, we obtain the following model-averaged posterior distribution: $p(\delta|\text{data}) = s(\delta)pr(\mathcal{M}_0|\text{data}) + p(\delta|\text{data}, \mathcal{M}_1)pr(\mathcal{M}_1|\text{data})$. Here, s is the Dirac delta function which represents the spike under the null, pr denotes probability, and p denotes density.

several large-scale replications studies have demonstrated that a near-zero effect size is reasonable in practice (e.g., see the meta-analyses conducted by Klein et al., 2018; Camerer et al., 2018; Nosek & Lakens, 2014). The argument is not affected if the point null is replaced by a perinull.

A key aspect of model averaging is that it does not require model selection; there is no need to commit to a single model to obtain parameter estimates, although multiple models are considered. Therefore, model-averaged predictions and parameter estimates do fit the philosophy behind focusing on estimation. A skeptic might remark that in order to model average, it is necessary to obtain some form of model evidence and transform this into posterior model probabilities (e.g., Bayes factors, information criteria) which reintroduces the importance of testing. This could reintroduce the importance of testing as the method for obtaining model evidence has a large influence on the results. However, the same can be said for traditional estimation based inference, since confidence intervals can be constructed in multiple ways (e.g., bootstrapping, normal approximation).

When not to model average

When the sole aim is prediction rather than estimating and interpreting parameters, model averaging will likely be beneficial. However, if the goal is to interpret or base decisions on parameter estimates, there are several reasons to forgo model averaging. First, if the models are theoretical opposites it makes little sense to model average. The model-averaged parameters will be uninterpretable and thus meaningless. Second, the nature of the problem may be ill-suited for model averaging. For example, imagine we want to maximize patients' quality of life. A new experimental treatment, living at a high altitude for 2 years, will improve patients' quality of life but only if they stay for the entire 2 years. Providing patients with treatment essentially boils down to subsidizing their stay there for two years. For each patient, it is unknown whether they will complete the treatment or not. Here, one model represent that a patient finishes the treatment while another model represents that they do not. Using some background variables as predictors, we obtain for each patient the posterior probability that they complete the treatment. However, in this scenario it is meaningless to average the subsidy spent on a patient by the posterior probability of them completing the treatment, as such an average would always provide patients with less than the required amount to complete the treatment, essentially wasting the subsidy. More generally, whenever some form of thresholding will be applied to the model outcomes, results of model averaging may not be useful.

In sum, we argue that descriptions of effect size based on only the alternative hypothesis are overconfident and as a consequence overestimate effect size. A remedy for this overestimation is model averaging. Although this idea is not new, model averaging remains underutilized in practice and we hope this paper brings more attention to model averaging.

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.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

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.

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.

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

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.

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

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.

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.