Outline

Introduction

- Introduce an example about reporting effect size + confidence interval (with data, preferably a t-test).
- Reference some editorials + papers on the topic.
- Pop the question: Given the data, what is the best estimate of effect size?

Ignoring Model Uncertainty Leads to Overestimating Effect Size

- There are many reasons why effect size is overestimated. Mention some reasons that will not be discussed (e.g., publication bias) and some that will be discussed (ignoring model uncertainty).
- Make the argument that since model uncertainty is considered initially, this uncertainty should propagate to parameter estimates.
- Relate this argument to model averaging and its successes.
- Reanalyze the data example from the introduction but model average over H0 and H1.

Discussion

- When not to model average:
  1. when models are theoretical opposites.
  2. statistically different (e.g., uninterpretable)
  3. don't have to model average but instead use spike and slab model
- say you can only report one number? Then we recommend the model averaged estimate + Bayes factor

Todo list

A Cautionary Note on Estimating Effect Sizes with Confidence Intervals

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

content...

*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.

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 $d = 0.54$, CI $= [0.01, 0.85]$.

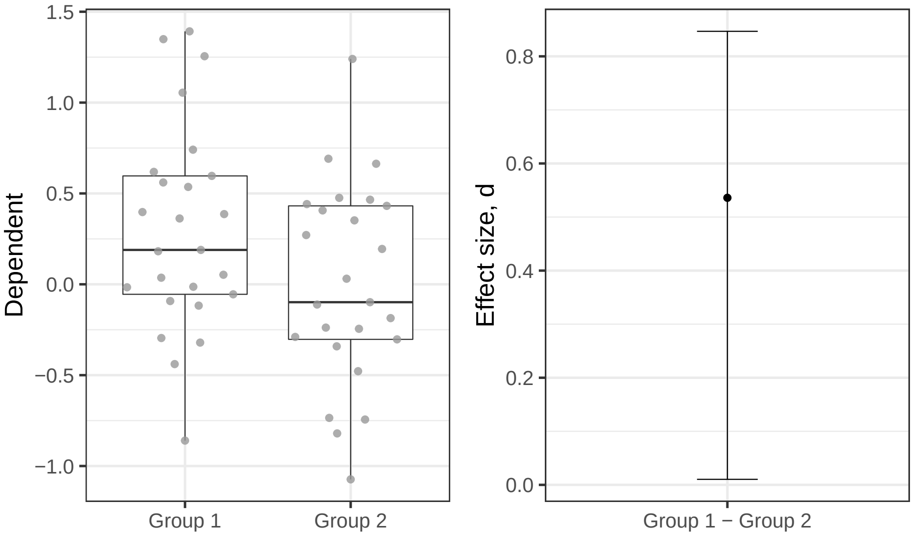

Figure 1: The left panel shows a box plot with the observations as gray dots. The right panel shows an estimate of the effect size, Cohen's d, and a 95% confidence interval. The data were simulated.

Given the results shown in Figure 1, what is the best point estimate of effect size? An obvious answer is "0.54"; since the confidence interval does not contain 0, the sample effect size seems to be a reasonable choice. However, your colleague now tells you about the nature of the experiment: plants grow faster when you talk to them. Suddenly, an effect size of "0" also appears plausible.[1]

Although some may argue that all null hypothesis are false (Cohen, 1994, ?), several large-scale replications studies have demonstrated that an effect size of zero is reasonable (e.g., see the meta analyses conducted by Klein et al., 2018; Camerer et al., 2018; Nosek & Lakens, 2014).

When Effect Sizes are Overestimated

Given a limited sample size, a point estimate of effect size and a confidence interval tend to overestimate the true effect size. To illustrate, we reanalyze the simulated data from Figure 1. There are two hypotheses considered: The

[1] Unless you pray out loud, with consumption, and the plant is near.

null hypothesis which states that the two groups have the same population mean and the alternative hypothesis that states the two groups differ in their population mean. For simplicity, we test these hypothesis using a two-sample t-test. Traditionally, one decides upon a single model using some form of significance testing. Assuming the the alternative model comes out victorious, it is then used to obtain estimates of effect size and predictions for future observations. However, it has been shown repeatedly that model averaging 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, an estimate of the model averaged effect size can be obtained by averaging the effect size of the null model and that of the alternative model by the posterior probabilities of both models. This yields $p(\delta \mid \mathcal{D}) = 0pr(\mathcal{M}_0 \mid \mathcal{D}) + p(\delta \mid \mathcal{D}, \mathcal{M}_1)pr(\mathcal{M}_1 \mid \mathcal{D})$. Figure 2 shows both the model averaged posterior of effect size and the posterior effect size under $\mathcal{M}_1$.

Let d be the effect size of the model averaged posterior predictive distribution. Consider the limit of d as the number of posterior predictive observations goes to infinity. Do we retrieve the model averaged posterior effect size? Intuitively yes, which could be a nice argument that the model averaged effect size can definitely be interpreted for this model.

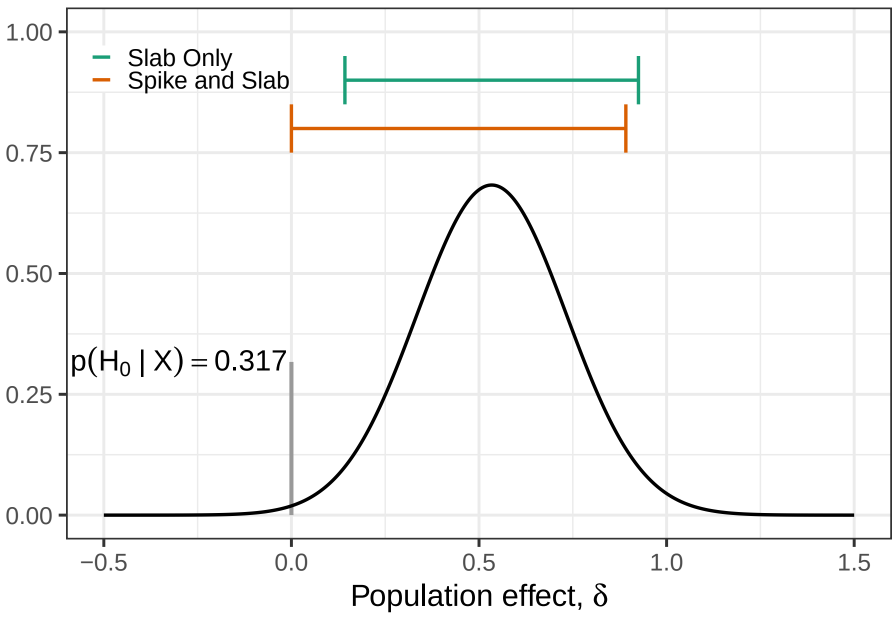

Figure 2: A visualization of model averaging. The black line represents the posterior of the effect size given the alternative model (i.e., the slab). This posterior distribution integrates to the posterior probability of the model given uniform prior model probabilties. The grey line represent the posterior probability of the null model (i.e., the spike). The yellow error bar shows a 95% central credible interval if inference is only based on the posterior under the slab. The purple error bar shows a 95% central credible interval if inference is based on the model averaged posterior distribution.

Do we want a plot of the prior as well? That would be a bit close to Johnny's figure though.

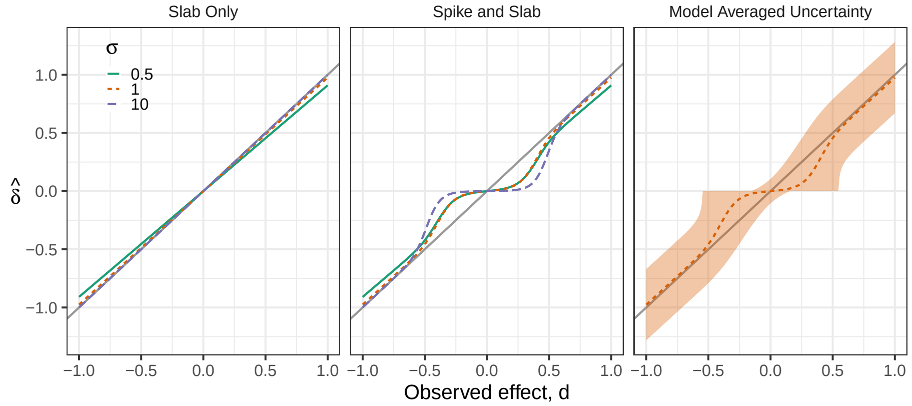

Figure 3: Posterior mean for effect size (y-axis) as a function of the observed effect size (x-axis). The left panel shows inference conditional on the alternative model (i.e., the slab). The right panel shows the model averaged posterior mean, which shrinks towards 0 as the sample effect size approaches 0 and the null model becomes more plausible. The colors and line types represent different variances of the prior distribution.

Show prior-posterior plot from JASP: With sparse data, the prior distribution on effect size will shrink the estimate towards 0. Happens with the default settings, but even more when the width is smaller.

Also, show spike at zero (maybe we need better JASP plot, will ask Don to create it):

The impact of H0 will shrink the estimate toward 0. This is most clear when H0 is a priori very likely (”plants do not grow faster when you pray for them”) or when the sample effect is very close to zero, so that it becomes clear that H0 might provide a more reasonable explanation.

Mention earlier literature: Model averaging effect size: (Zellner & Vandaele, 1975, pp. 640-641), as described in (Zellner & Siow, 1980, p. 600-601); also (Haldane, 1932, p. 57), (Iverson et al., 2010)

Early ideas that are conceptually similar can be found in (Wrinch & Jeffreys, 1921, p. 387) (show BMA between $\mathcal{H}_1$ and $\mathcal{H}_0$ – but for prediction, not estimating effect size!; see also (Jevons, 1874/1913)).

See also (Rouder et al., 2018):

Key is Figure 5. The spike-and-slab model shows shrinkage towards zero for small observed effect sizes because the spike has increased influence.

“There are alternative interpretations that we find somewhat cumbersome. One is that the spike-and-slab can be viewed not as a model but as a model-averaging device. Here, the goal is not so much to define categories of effect and no-effect, but to average across both of them, and this averaging results in regularization. If one uses this interpretation, the prior odds settings are important as they influence the posterior weight given to each model component in the averaging. Another alternative interpretation comes from Kruschke and Liddell (Kruschke, 2011; Kruschke & Liddell, 2017; this issue). Here, the spike and slab are seen as separate components in a hierarchical model. Accordingly, a focus on Bayes factors denotes a focus on the choice between components; a focus on posterior estimation entails parameter estimation after choosing the slab. We find this view difficult inasmuch as there is no a priori reason to choose the slab to focus on estimation. If one admits the possibility of the spike, then assuredly it should affect posterior estimation as well."

Discussion

Bayesian estimators more likely to underestimate the effect size, as prior mean is typically centered around the null which pulls the posterior mean towards 0.

When not to model average

Which models to average? When to average? Also consider alternatively one model that has all relevant components for estimation instead of separate models.

References

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.

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.