Assessing the median effect robustness in Silberzahn et al. (2018) through model specification

How analysis strategy affects analysis results

Sebastian Ploner, sploner1@sheffield.ac.uk, University of Sheffield

Introduction

The Covid-19 pandemic has reaffirmed the need for rigours scientific research to inform decision-making on a scientific and societal level. However, particularly the social sciences have a history of inflated positive findings (Ionanidis) leading to the reproducibility crisis (Pashler & Wagenmakers, 2012). Simmons et al. (2011) demonstrated how the so-called *researcher degree of freedom* (RDF) affect study results. This concept refers to the analytical choices researchers make and their induced variation in results. For instance, sample size can be highly deterministic of finding statistically significant effects. John et al. (2012) found 70% of researchers have at least once stopped their data collection based on a preliminary analysis. P-value simulations as a function of sample size showed an increased likelihood of obtaining false-positives for small sample sizes. Specifically, a sample size of ten had a 22.1% chance of being a false positive. Adding 20 observations per condition decreased the probability to 12.7% (Simmons et al., 2011). The authors did not insinuate any malicious intent per-se, but rather attributed the issue to an ambiguity of a "right way" to analyse data, and needing statistically significant results to publish. This example shows how easy it is to engineer statistically significant effects. To prevent such bad practices Simmons et al. (2011) suggested requirements and guidelines for authors and reviewers.

Among these requirements and guidelines are a minimum sample size, reporting all collected variables and presenting results with and without covariates (Simmons et al., 2011). These are important suggestions, but they struggle to cope with the full spectrum of the analytical possibilities. Silberzahn et al. (2018) further scrutinised the effects of RDF. Specifically, they had 29 teams of researchers investigate the same research question based on the same dataset. Other than being part of the project there was no incentive for researchers to participate i.e., there was no ulterior motive for researcher to manipulate outcomes (e.g. wanting/needing to publish). The variation between the analytic approaches was substantial. There were 29 different analyses with 21 different combinations of covariates. Twenty teams found a statistically significant effect; the odds ratios ranged from 0.89 to 2.93 (median = 1.31). The authors also controlled for prior believes, experience and peer rated analysis quality, none of them accounted for the variation of results. Similar studies support these

RDF effects (Botvinik-Nezer et al., 2020). These studies show that even extended efforts do not necessarily yield conclusive results. Moreover, this crowdsourcing approach is not practical.

Silberzahn et al. (2018) and Botvinik-Nezer et al. (2020) suggest using multiverse analysis, or specification-curve analysis, as an alternative to crowdsourcing approaches. This approach requires a researcher, ideally a more than one, to conduct with every plausible analysis, aggregate and report it, and conduct an inferential test across all possibilities to come to a conclusion. "Plausible" refers to statistically valid tests appropriate to the research question and non-redundancy (Simonsohn et al., 2020). This approach is statistically more complex and computationally more intense than conventional analyses, but makes inherent strategy bias and noise transparent, and forces researchers to look at the entire "garden of forking paths" (Gelman & Loken, 2013). A good example of quantifying the extend of variation due to analytical choices was conducted by Patel et al. (2015). They tested the association of clinical, environmental and physiological variables with all-cause mortality. Specifically, they specified three model approaches (gaussian, cox and binomial) and 13 different covariates. Their algorithm ran every possible covariate combination (on top of the base variables) for each model which allowed the authors to assess the "vibration of effect" (VoE). They concluded that a larger VoE calls for more caution when interpreting observational associations.

Taken together, statistically effects are easy to engineer through adjusting analysis parameters like sample size. Crowdsourcing analysis is a good strategy to understand the RDF induced variation but is impractical due to its substantial personnel, time and effort requirement. Alternatives like multiverse analysis present a promising, low-personnel, but computationally intense alternative to identify RDF induced variation. The present study will build upon Silberzahn et al. (2018), Botvinik-Nezer et al. (2020) and Patel et al. (2015). The objective is to conduct a multiverse analysis based on the data used in Silberzahn et al. (2018) and identify the VoE i.e., the results space of the outcome variable. This will allow for a better understanding of how robust crowdsourced identified effects are. Due to time constraints this study will function as a prove-of-concept. Specifically, it will take the analytical approach which produced the median outcome and run it with every possible covariate combination. Similarly to Patel et al. (2015) this produces a distribution of the outcome effects and its variation provides insight into the Silberzahn et al. (2018) median effect robustness. The smaller the variation, the greater is the effect's robustness.

References

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/https://doi.org/10.1016/j.jclinepi.2015.05.029

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, S., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z