How analysis strategy affects analysis results

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

Master's thesis - DRAFT

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

Contents

. . .

Abstract–in progress.

. . .

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 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 effects to publish. This example shows how easy it is to engineer these 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 effects of skin colour on the odds of being sent off the football field 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 researchers' prior believes and experience as well as peer rated analysis quality, none of them accounted for the variation of results. Other crowdsourced analysis projects have made similar observations (Botvinik-Nezer et al., 2020). These studies are interesting for two reasons (a) from a statistical perspective they highlight the RDF effects, and (b) from a content perspective they provide the most robust insights into the scrutinised concepts. Nonetheless, crowdsourcing approaches are impractical as they require extensive personnel (N number of researchers) and time resources (years) and are therefore often not feasible analytical strategies (Silberzahn et al., 2018).

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 all plausible analyses, aggregate the results and conduct an inferential test across all possibilities to come to a conclusion. "Plausible" refers to statistically valid, non-redundant, tests

appropriate to the research question (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 model specification was conducted by Patel et al. (2015). They specified three modelling approaches (gaussian, cox and binomial) and 13 covariates. Their algorithm ran every possible covariate combination for each modelling approach which allowed the authors to assess the "vibration of effect" (VoE). The VoE is another term for quantifying the extend of variation in the results. The measure allows to assess the effect's robustness. The smaller the VoE, the less the effect depends on the analytical choices and vice versa. The authors therefore called for more caution when interpreting observational associations when finding large VoE's.

Taken together, statistically significant effects are easy to engineer through adjusting analysis parameters like sample size. Crowdsourcing analysis presents a suitable strategy to understand the RDF induced variation but is impractical due to its substantial personnel and time requirements. Alternatives like multiverse analysis present a promising, low-personnel, but computationally intense alternative to identify RDF induced variation. The present research project will build upon the crowdsourcing findings of Silberzahn et al. (2018). Specifically, I will take the median analysis and rerun it with all possible covariate combinations. The objective is to define and understand the results space of the analysis i.e., assess the robustness of the median effect (VoE). Based on these insights I aim at contributing to further understanding the value, limitations and mechanisms of the multiverse approach.

Analysis

Analysis plan

The project's objective is to define the results space of the Silberzahn et al. (2018) in order to draw inferences about the relationship between crowdsourcing and multiverse strategies. The results space refers to the numerical interval between the lowest and highest possible outcome. It is created by running every possible analytical strategy. Patel et al. (2015) refer to a related concept as the "Vibration of Effect" (VoE) and have developed a standardised approach to identifying this space. Essentially there are two factors: the analytical approach (i.e., the statistical model) and the covariates (or control variables). For Silberzahn et al. (2018) this means running every combination of 29 different analytical approaches and 15 different covariates used across all teams. These are $29 * 2^{15} = 950,272$ possibilities, which exceeds the scope of this research project. To get a more manageable number of possibilities of I will, therefore, focus on one analytical approach. In particular, on the approach that produced the median outcome of all analyses. The median outcome, being the middle number of any set of values, is a reasonable starting point in order to assess the VoE. Focusing on one analytical approach still leaves $1 * 2^{15} = 32,768$ possibilities. Due to computational

limitations of my computer I will focus on a random sample of 1,000 thereof.

In Silberzahn et al. (2018) the median outcome was produced by Stafford et al. (2014) which will hereinafter be referred to as *team 23* due to it's designated team number in the original study. *Team 23* first transformed the data and then conducted a mixed-model logistic regression. To recreate the identical starting point for the analyses I will, first, apply the same transformations using the scripts provided by *team 23*. (The transformation process will be described in more detain the "data" section.) Second, I will create a matrix of all possible covariate combinations whereof a random sample of 1,000 is drawn. Third, because the original study investigated the relationship between football players' skin colour and the odds of being sent off the field, there will be a set of core variables included in all models to be able answer the original research question. Following on *team 23*, I will define two interaction terms as the model's core: "skin tone X implicit bias" and "skin tone X explicit bias." (The data set's variables will be described in more detail in the "data" section.) Forth, I will combine the core variables and the randomly drawn covariates as formulas to run the models. Fifth, after running the models I will visualise the outcomes as the conventional specification curve plot as well as a rain-cloud plot (Allen et al., 2021) and scatter plot. The aim here is develop easy to understand graphs that succinctly give insight into the results. Finally, I will discuss the implications for multiverse analysis.

Code

The code is based on Haessler et al. (2020) and Patel et al. (2015). The former provided the operational backbone of the code. The latter the idea for the algorithm. R's random number generator is first fixed to be able to reproduce the outcomes. Then, all the packages are detached and only the relevant ones are loaded to prevent other functions from interfering. If the required packages are not installed, the code will do so. Next, the prepared data is loaded using the "Here" package (Müller, 2020) for operating systems independence and the "data.table" package for faster importing of the data (Dowle & Srinivasan, 2021). (The data has been prepared by running the "disaggregate_v3.py" script of *team 23*; for more detail see the "data" section.) After having loaded the data the variables are defined as dependent, base (or core) variables and covariates. This includes defining the variables' types: numeric, factor or random effect. Next, a matrix is created established all possible combination of covariates (including base variables). Each row represents one combination, the columns represent the variables. The cells are set to TRUE or FALSE to indicate a variables presence or absence, respectively. Thus, the matrix structure looks as the following:

$$\begin{bmatrix} \text{dv} & \text{base var}_1 & \text{base var}_2 & \text{base var}_3 & \text{covar}_1 & \text{covar}_2 & \text{covar}_{...} & \text{covar}_{15} \\ \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{FALSE} & \text{FALSE} & ... & \text{FALSE} \\ \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{FALSE} & ... & \text{FALSE} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & \text{TRUE} & ... & \text{TRUE} \end{bmatrix}$$

Using the *apply* function, looping over each row, the variables are pasted together thereby creating the corresponding formulas which are then saved as a vector and are appended as column to the aforementioned matrix.

Team 23 ran a mixed-model logistic regression analysis using the "lme4" package more specifically, its "glmer" function (Bates et al., 2015). The function only works if a random structure is included in the formula. As all possible covariate combinations are explored, a random structure is not included in every formula. Therefore, before running the models, the formulas need to separated based on including a random structure or not. The random structures were defined at the beginning of the code using the following syntax: "(1|*covariate*)." Hence, this pattern is also used to identify random structures through the "grepl" function. For the non-random structure formulas the regular "glm" function is run. Both models, mixed-effect and "regular," are estimated using Maximum Likelihood estimation to ensure the outcomes are as comparable as possible. *Team 23* had also set the "nAGQ" parameter to zero (for the mixed-model) this sacrificed parameter estimation accuracy for speed. The same was done here to stay as close to the original analysis as possible. Finally, 1,000 formula samples were drawn. 75% of all formulas include at least one random structure this proportion is reflected in the sample. Furthermore, the samples are drawn without replacement i.e., every formula is unique.

Running the models yielded another challenge. Not only were the running times per model at times above 30 minutes, storing the model summaries also required significant memory. For instance, testing the code with a minimal sample of 10 (random structure) formulas yielded an analysis output larger than 400MB. This does not scale well, hence, a wrapper function was build. It's purpose was to extract the estimates, standard errors, test-statistics and p-values for the variable of interest (in this case: skin tone). This effect was substantial. The final outcome object for 1,000 models was reduced to 0.3MB(!).

The models were run using the "lapply" function more specifically, the paralellised version thereof including progress bars ("pbmcapply," Kuang et al. (2019)). Parallisation was necessary in order to reduce running time. The standard R is set to use one core, only. Parallelising the processes to six cores decreased the running time by six-fold. The results of each model iteration was appended to a list. Hence, the outcomes were two lists, one from the mixed-models, and one from the "regular" models. These lists were transformed

to data frames and merged. Moreover, the p-values was coded into significant and non-significant ($\alpha = 0.05$). The odds ratios as well as their confidence intervals were calculated (the same way *team 23* did it). The resulting data frame was save as a comma-separated values (CSV) file.

Algorithm 1: Pseudocode overview–in progress

Data: Prepared data based on team 23 script
$variables \leftarrow$ Define dependent, base variables and covariates
$specifications \leftarrow$ Use $variables$ to create matrix containing all possible covariate combinations
$formula \leftarrow$ Paste $specifications$ by row and append as column to $specifications$
$formula.ranef; formula.ef \leftarrow$ Separate formulas based on including a random-structure
$samlple.ranef; sample.ef \leftarrow$ Random, without replacement, sample from $formula.ranef; formula.ef$
initialization
while *not at end of this document* **do**
  read current
  if *understand* **then**
    go to next section
    current section becomes this one
  else
    go back to the beginning of current section
  end
end
Result: List containing model statistics

...

Explaining why not adjusting for multiple comparison–in progress

...

To better understand the specified effects of the covariates I ran an ANOVA where each covariate was coded into included yes/no. The above retrieved effects where used outcome measure. The results were stored and saved as a separate data frame. As a final step the graphs were created using the "tidyverse" (specifically, "ggplot," Wickham et al. (2019)), "PupillometryR" (Forbes, 2020) and "cowplot" (Wilke, 2020) packages.

...

Explaining how the graphs were created including necessary data transformation needed–in progress.

...

The code is written in R (Version 1.4.1103) on macOS Big Sur (Version 11.4) and can be retrieved from my *GitHub repository*.

Data origin and preparation

Outline:

- The same dataset as in Silberzahn et al. (2018) is used.
- What is the data about
- Data origin i.e., how did silberzahn et al. (2018) come up with the data set
- how was the data processed by team 23
- summary statistics of the data set

	N	Missing	Mean	SD	Min	Q1	Median	Q3	Max
height_cm	335509	28	181.98	6.81	161.00	178.00	183.00	187.00	203.00
weight_kg	334727	810	76.33	7.12	55.00	71.00	76.00	81.00	100.00
age_yrs	335537	0	37.11	4.11	25.60	34.20	37.10	40.20	50.20
games	335537	0	7.61	6.64	1.00	2.00	6.00	11.00	47.00
goals	335537	0	0.88	1.79	0.00	0.00	0.00	1.00	23.00
victories	335537	0	3.43	3.61	0.00	1.00	2.00	5.00	29.00
ties	335537	0	1.83	1.96	0.00	0.00	1.00	3.00	14.00
all_reds	335537	0	0.01	0.09	0.00	0.00	0.00	0.00	1.00
player_cards_received	335537	0	413898.00	0.00	413898.00	413898.00	413898.00	413898.00	413898.00
ref_cards_assigned	335537	0	413898.00	0.00	413898.00	413898.00	413898.00	413898.00	413898.00
skin_tone_num	335537	0	0.28	0.28	0.00	0.00	0.25	0.38	1.00
imp_bias	335537	0	0.00	0.02	-0.39	-0.01	-0.01	0.02	0.23
exp_bias	335537	0	0.00	0.16	-1.81	-0.10	-0.08	0.15	1.36

Outcome I: vibration of effect

Outcome II: specified covariate effects

Outcome III: specification curve

Discussion

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T., Langen, J. van, & Kievit, R. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*, *4*(63). https://doi.org/10.12688/wellcomeopenres.15191.2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Forbes, S. (2020). *PupillometryR: A unified pipeline for pupillometry data*. https://CRAN.R-project.org/package=PupillometryR

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*.

Haessler, T., Ullrich, J., Bernardino, M., Shnabel, N., Van Laar, C., Valdenegro, D., Sebben, S., Tropp, L. R., Visintin, E. P., Gonzalez, R., Ditlmann, R. K., Abrams, D., Selvanathan, H. P., Brankovic, M., Wright, S., von Zimmermann, J., Pasek, M., Aydin, A. L., Zezelj, I., ... Ugarte, L. M. (2020). A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour*, *4*(4), 380–386. https://doi.org/10.1038/s41562-019-0815-z

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

Kuang, K., Kong, Q., & Napolitano, F. (2019). *Pbmcapply: Tracking the progress of mc*pply with progress bar*. https://CRAN.R-project.org/package=pbmcapply

Müller, K. (2020). *Here: A simpler way to find your files*. https://CRAN.R-project.org/package=here

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

Stafford, T., H.Evans, M., J.Heaton, T., & Bannard, C. (2014). *Crowdstorming team 23: There is definite racial bias in which soccer players are sent off, but its locus is unclear"*. https://osf.io/akqt4/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot