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 sound scientific research to inform decision-making on a scientific and societal level (Collins, 2021). Rigorous research builds upon a systematic and well-reasoned approach to solving a research problem. Based on available literature, a falsifiable research question is defined and a corresponding hypothesis is developed. An appropriate study is then designed and conducted. To draw inferences from the gathered data, statistical models are developed and the effect of each variable is assessed. Every one of these steps is influenced by the decisions researchers make, which are known as *researcher degrees of freedom* (RDF, Simmons et al., 2011; Wicherts et al., 2016). Usually there are many feasible analytical strategies to answering a research question and none of them are inherently right or wrong (Carp, 2012). This often creates uncertainty, for example, about what covariates to include and how to model them, which in turn leads to inconsistent findings (Ioannidis, 2008; Patel et al., 2015). Recently, efforts have been made to better understand the scope of variation induced by different analytical strategies.

These efforts include a crowdsourcing approach to data analysis (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). Its premise is to have a large number of researchers team up into to smaller, independent groups to investigate the same research question based on the same dataset. For instance, Silberzahn et al. (2018) had 29 teams investigate the effects of a football player's skin colour on the odds of being sent off the field. The variation between the analytical strategies was substantial. There were 29 different analyses with 21 different combinations of covariates. Twenty teams found a statistically significant effect. The authors also controlled for researchers' prior believes and experience as well as peer-rated analysis quality, none of which accounted for the variation of results. Botvinik-Nezer et al. (2020) made similar observations. Crowdsourcing data analysis excels at emphasising the substantial impact of different analytical strategies, but has a significant drawback. It is extremely time and personnel intensive, the two mentioned studies lasted between 2 to 3+ years, and included 61 and 180 analysts, respectively. Not to mention the organisational effort.

Silberzahn et al. (2018) and Botvinik-Nezer et al. (2020), therefore, suggest using multiverse analysis (also known as specification-curve analysis) as an alternative to crowdsourcing. This approach requires identifying and running all plausible analytical strategies. Plausible strategies are defined as statistically valid, non-redundant, tests appropriate to the research question. Their aggregated results are then used to make inferences about the research question. Despite being statistically more complex and computationally more intense, this approach has the advantages of only needing one, or ideally a couple, researchers. Moreover, a researcher's inherent strategy bias is neutralised and noise is made transparent (Simonsohn et al., 2020). Multiverse approaches are therefore well suited for assessing the scope of variation induced by different analytical strategies. For instance, Patel et al. (2015) assessed an effect's robustness by running every combination of covariates and modelling options. Their conclusion was the larger the variation of outcomes, the less robust is the effect, and the less it should therefore to be trusted. The authors termed this measure the *vibration of effect* (VoE).

Taken together, researchers make many analytical decisions throughout conduction a study which are know as *researcher degrees of freedom*. Crowdsourcing analysis presents a strategy to understand their induced variation, but is impractical due to its substantial time and personnel requirements. Multiverse analysis present a promising, low-personnel, but computationally intense alternative to assessing the scope this variation. Despite observing substantial variation in crowdsourced data analysis projects little is known about the extend to which these approaches cover the full range of possible results. In this project, using methods of multiverse analysis I will define the results space of Silberzahn et al. (2018) to estimate study's representativeness. 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) to the estimate study's representativeness. The results space refers to the numerical interval between the lowest and highest possible outcome. It is created by running every possible analytical strategy. Hence, it is closely related to assessing an effect's robustness. Patel et al. (2015) developed a standardised approach to assessing an effect's robustness which they termed assessing the *vibration of effect*. I will therefore use this standardised approach to define the results space. The approach essentially comprises two parameters: the statistical models and the covariates (or control variables). In Silberzahn et al. (2018) the analysts used numerous different statistical models like multiple linear regression, mixed-model logistic regression or Bayesian logistic regression. In total there were 29 different modelling approaches. Additionally, each team used a different set of covariates. Across all teams there 15 covariates used. This gives 2^n i.e., $2^{15} = 32,768$ possible combinations of covariates. Adding all modelling possibilities to the equation gives a total of $29 * 2^{15} = 950,272$ combinations to run. This number exceeds the project's scope, hence, it needs to be reduce to a more manageable count. I am, therefore, focusing on one modelling approach. In particular, on the approach that produced the median outcome of all analyses. The median, being the middle number of any set of values, is a reasonable starting point in order to estimate the results space. Nevertheless, even focusing on one analytical approach still leaves $1 * 2^{15} = 32,768$ possibilities. Due to computational limitations I will first run a sample of 200, followed by a sample of 1,000 combinations. If there's a substantial difference between the median and the spread I will run more models, if not, I will make the assumption that 1,000 reasonably estimates all ~33k combinations.

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. The first step of this project will be replicating the transformation and the analysis. Replicating other researchers analyses has the

benefit of checking their work and, if results are indeed replicated, increasing confidence in them. It also ensures that this project has the same starting point as *team 23* did. Given the team made all their scripts publicly available I expect this step to be straightforward. Next is drawing a random sample of covariates, without replacement. "Without replacement" ensures all covariate combinations are unique in the sample. The covariates will be appended to the base (or core) variables. Base variables are those variables that are primarily assessed to answer the research question. In this case whether a football player's skin colour effects the odds of being sent off the field. *Team 23* defined two interaction terms as the base: "skin tone X implicit bias" and "skin tone X explicit bias." (The variables are described in the "data" section.) Hence, all models will have the following structure:

$$RedCard = SkinTone * ImplicitBias + SkinTone * ExplicitBias + CovariateCombination_i$$

While running each model, the relevant statistics will be extracted. Those are the coefficient (or effect) of skin tone, its standard error, test-statistic and p-value. Just as *team 23* did I will then calculate the 95% confidence intervals (CI). The estimates and their CIs are then transformed to odds ratios (OR) through exponentiating them to the power of two. Odds ratios quantify the strength of association between two variables. If greater than one the dependent variable is more likely to occur given the independent variable, if lower than one it less likely. The odds ratios will be visualised to assessed through a conventional specification curve plot, a rain-cloud plot and scatter plot. The following describes them in more detail:

(i) The specification curve plot was developed to be a descriptive plot i.e., raw data is being plotted without any aggregation done. Its objective is describe the results space while allowing the reader to identify the specified covariate for each outcome (Simonsohn et al., 2020). It therefore has two vertically stacked components. The top part shows a sorted scatter plot. On the horizontal axis are the models specifications and on the vertical axis is the outcome measure. The points are sorted from lowest to highest outcome measure. This way the lowest outcome measure is on the bottom left corner and the highest in the top right one. The bottom part of the plot represents are table. Just like the top the horizontal axis hold the model specifications, the vertical axis lists the covariates. This arrangement allows each cell to specify the presence or absence for each covariate in a given specified model. For the top and bottom plot to work together it is imperative that the horizontal axes are identically arranged. If so, for every point in the top plot (i.e., for every outcome) the specified covariates can be seen in the bottom plot. For an example check the results section. The issue, however, is that if too many specifications are on the horizontal axis it becomes very hard to identify specific covariates. Hence, I also included the rain-cloud and scatter plot.

(ii) The rain-cloud plot was developed by Allen et al. (2021). Its objective is to give an unbiased, transparent view on the raw data. It therefore combines a density, box and scatter plot. If stacked vertically from top to bottom, respectively, they look like a cloud with rain drops, hence, its name. The advantage of

this plot is to be able to assess the raw data including key statistics (median, 25th and 75th quartiles and CIs) as well as the probability density all in one succinct plot. Given that this project seeks to define and describe a results space the rain-cloud plot the ideal graph to do this. However, the plot neither allows to identify the covariates giving rise to the outcome nor their specified effects. (For an example check the results section.) Therefore a third plot is developed which is to be assessed in tandem with this plot.

(iii) The third and final plot is a sorted scatter plot. While there is nothing special about the plot itself its content is important. Given the very high number of models that are run, it does not make sense to assess each model and its effects in isolation. Instead the effects for each covariates neeed to be aggregated. Hence, the goal of this plot is to viusalise the specified effect for each covariate. Prior to doing so, the coviates first need to be aggregated and their specified effects need to be calucualted. To this end, each covariate is recoded into a binary factor: included in the model yes/no. This newly defined factors are than used as an independent variables in an ANOVA. The previously calcualted odds rations are used as dependent variables. The estimates of the fitted model are the specified effect for each covariates. These are visualised similarly to the top part of the specification curve plot. The covariates are on the horizental axis, will the odds ratios are on the vertical axis. Here too, the lowest oucome is in the bottom left corner and the highest outcome in the top right one. Additionally the point are coloured indicating a statistically significant effect (or non-signifcant).

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

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6*, 149. https://doi.org/10.3389/fnins.2012.00149

Collins, F. S. (2021). COVID-19 lessons for research. *Science*, *371*(6534), 1081–1081. https://doi.org/10.1126/science.abh3996

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

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

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

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

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/

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

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