DRAWING (CAUSAL) CONCLUSIONS FROM DATA – SOME EVIDENCE

Karsten Lübke, Bianca Krol and Sandra Sülzenbrück
FOM University of Applied Sciences, ifes Institute for Empirical Research & Statistics, iwp Institute for Business Psychology, Leimkugelstr. 6, 45141 Essen, Germany
karsten.luebke@fom.de

To be data literate, one should be able to draw conclusions from multivariable observational data. But this is tricky. E.g., to investigate the gender pay gap, it must be decided whether the effect should be calculated adjusted or unadjusted for the job. The correct conclusion depends on the qualitative assumptions about the data generating process. To investigate the conclusions drawn by students, a randomized experiment is conducted. The same data is presented in two different contexts with (possible) different structural causal models so once the adjusted and once the unadjusted effect might be appropriate. Also, it is varied whether a directed acyclic graph is presented before or after the data table with the estimated effect. Results indicate that conclusions drawn from the same data differ by context but may also be incoherent with the assumed data generating process.

INTRODUCTION

As Simpson's paradox shows, the same data can lead to opposite conclusions being adjusted for a third variable or not. But whether adjusting or not adjusting is appropriate depends on the qualitative assumptions about the data generating process. One possibility to encode these assumptions and to derive correct conclusions based on these is the use of directed acyclic graphs (DAG). An illustration and discussion of resolving Simpson's paradox by DAGs is given in Pearl et al. (2016).

Integration of Causality into data literacy education may help students to connect contextual information and computation (Greenland, 2021). However, assessing tri-variable relationships as in Simpson's paradox is a complex higher-order cognitive skill (e.g. Fiedler et al., 2003), requiring more cognitive resources to overcome attributional biases like the so-called fundamental attribution error (Ross, 1977).

This paper presents data about student's conclusions for the effect (adjusted or unadjusted mean) of an exposure on an outcome as well as assumptions about the underlying data generating process encoded in a DAG in different contexts regarding the pay gap.

STUDY DESIGN

An online experiment was conducted to assess the conclusions of students in such a three-variable (x, y, z) setting like Simpson's paradox. In a between-person design the two different *exposures* ("*x*"), the context of the analysis (*gender* or *lifestyle*), was randomized as the first factor. The outcome ("*y*") in both cases was *salary*, the covariable ("z") for both was *management position*. Also, the *presentation order* was randomized: In condition *DAG first*, participants initially had to choose the appropriate DAG before the numerical summary (data table and mean calculation) was presented, and the appropriate effect had to be selected. In condition *Table first*, participants were first shown the numerical summary and then the DAG.

The short online survey introduced a fictitious company whose employees were split into two groups (A, B) according to the *exposure* (*gender* (female, male) or *lifestyle* (healthy, unhealthy)). The average salary per (sub-)group was presented as shown in Table 1 combined with the calculation for the adjusted (conditional on management) and unadjusted mean. For the exposure *gender*, A represents female and B male, for the exposure *lifestyle*, A corresponds to a healthy lifestyle and B to an unhealthy lifestyle.

Table 1. Salary data for both exposures and both management positions

	Exposure (Group A)	Exposure (Group B)
Non-Management	3100 € (n=80)	3000 € (n=60)
Management	5850 € (n=20)	5500 € (n=40)

The *Table task* was constructed so that unadjusted for management group, group A (female or healthy lifestyle, respectively) earned on average 350 € less than group B (male or unhealthy). However, since in both conditions (management and non-management) members of group A earned more than members of group B, the average pay adjusted for management for group A is 175 € higher than for group B. Therefore, as described in the Simpson's paradox, the sign of the effect changes. Participants were asked to indicate the effect on the salary of being in group A as compared to group B (average -350 € or conditionally adjusted for job +175 €).

In the *DAG task*, participants were asked to choose which one of the two graphs presented in Figure 1 they thought was appropriate for the data generating process. Whereas in model A (left side of Figure 1), management is in the middle of a *chain* between exposure (*gender* or *lifestyle*) and outcome (salary), it is in the middle of a *fork* in model B (right side of Figure 1). So, in model A, being in management depends on the value of the exposure, whereas in model B the exposure depends on the value of management. Graph theory implies that one should not adjust for a mediator *Z* in a chain to determine the total causal effect (model A) of *X* on *Y*, whereas one should adjust for a confounder *Z* (middle of fork) like in model B, see e.g. Pearl et al. (2016).

It should be noted that the *true* model cannot be derived or verified by data alone. However, for the exposure *gender* only model A (management as a descendant of gender) is appropriate whereas for the exposure *lifestyle* both models might be regarded as true.

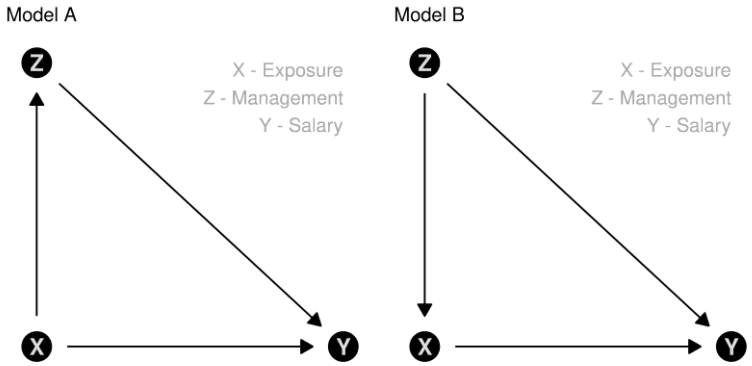

Figure 1. Directed Acyclic Graph for Salary Data, with the chain model A on the left and the fork model B on the right side.

Additionally, we checked if the conclusions of participants were consistent, i.e., if those who chose the adjusted mean for the presented table also chose model B (management as a confounder) and those who chose the unadjusted mean were consistent by choosing model A (management as mediator).

The online survey was presented to employees of a broad range of industries, additionally being part-time or full-time students of business economics and business psychology. They received course credit for participating in the study.

After having received general information about the study and given consent to participate, participants were randomly allocated to the four experimental conditions, which were presented on different landing pages. These four experimental conditions resulted from combining the two randomized experimental factors *exposure* (*gender* or *lifestyle*) and *order of task presentation* (*DAG first* or *Table first*).

RESULTS

For the four different experimental conditions, we collected data from $n = 187$ students who finished the survey. No response was excluded from the analysis. Data and R code to reproduce the analysis is available from https://github.com/luebby/IASE-Satelite. Due to the implemented randomization approach, sample sizes for the different groups varied between 44 to 50. Of the respondents, 69 % regarded themselves as female (31 % as male, no other). Participants were between 19 to 45 years old, with a mean age of 26.6 years (sd = 4.6). Altogether, 12 % held a management position, 88 % indicated that they follow a healthy lifestyle.

As expected, there was a statistically discernible difference for the estimated effect regarding the *exposure*. For *gender,* 70 % chose the unadjusted mean, i.e. less payment for female employees, whereas for *lifestyle,* 64 % regarded the adjusted mean, i.e. the higher salary for a healthy lifestyle, as the correct effect. Not aggregated results are given in Table 2. Note that for both exposures the arguably correct answer (unadjusted mean for *gender,* adjusted mean for *lifestyle*) is chosen more often if the DAG is presented first.

Table 2. Proportions of responses for the true effect

Exposure	Order of task	Adjusted	Unadjusted
Gender	Table First	0.32	0.68
	DAG First	0.28	0.72
Lifestyle	Table First	0.57	0.43
	DAG First	0.70	0.30

For the chosen model (chain (A) or fork (B)) the results are given in Table 3. Aggregated across order of task, the proportion for chain is 72 % for *gender* and 58 % for *lifestyle*. Remember that for the *gender* setting, only a chain is appropriate, whereas, for *lifestyle,* both might be true with arguably B being more plausible.

Table 3. Proportions of responses for model choice

Exposure	Presentation	Fork	Chain
Gender	Table first	0.32	0.68
	DAG first	0.24	0.76
Lifestyle	Table first	0.45	0.55
	DAG first	0.39	0.61

The consistency of both choices (adjusted or unadjusted mean as estimated effect and covariable management in the middle of a fork or a chain) is rather low as Table 4 reveals.

Table 4. Consistency of choices

Gender		Lifestyle	
Table first	DAG first	Table first	DAG first
0.52	0.57	0.40	0.41

DISCUSSION

The Covid-19 pandemic shows the importance of drawing correct (causal) conclusions from observational data in a multivariate world. “Thinking clearly about correlations and causation” (Rohrer, 2018) seems to be more important than ever. As expected, in our study the conclusions drawn from the same data differ by context as well as the graph of the assumed data generating process. However, the low internal consistency might be alarming (compare Table 4) but the first results regarding DAGs look promising without being statistically discernible. In the interpretation of the results, it should be noted that participants did not receive any formal training in causal inference yet.

The current study has several limitations that should be addressed in future studies and analyses. The target population analyzed is very narrow, so far only data from employees additionally being business psychology students is available. So far, we did not analyze if the choices are self-serving, for example, if men choose the adjusted mean more often than women as this would imply that men earn more than women. Also, the preferred conclusions based on the fictitious data follow some kind of folk

theory and common stereotypical biases that can be explained by the fundamental attribution error (Ross, 1977). Our results confirm the findings by Fiedler, Walther, Freytag and Nickel (2003), that confirmation bias and fundamental attribution error play a crucial role in Simpson's paradox. In a recent study, Yanai and Lercher (2021) found that conclusion from data is biased by participants' own expectations. Women earn less and a healthy lifestyle is also beneficial for salary. Different fictitious data might show different results. Also, it would be interesting to investigate the within-subject effect of the randomized factors.

In conclusion, we concur with the title of Miguel Hernán's edX course: "Causal Diagrams: Draw Your Assumptions Before Your Conclusion". Data by itself is not enough, and we should try to make our assumptions as transparent and discussible as possible. Integrating DAGs may be a step in that direction. In data literacy education, the process of data modeling should be a central part of the curriculum. More emphasis should be put on the mapping and link between scientific subject matter knowledge and statistical modeling, see e.g. Pfannkuch et al. (2108). Therefore, teachers should provide a framework to discuss this science with data. Our results indicate that it is not enough just to provide graphical and numerical summaries of the data and to teach tools how to do so. Thinking beyond data – critical thinking about the story of the data and how it was generated – is needed to draw (causal) conclusions. Further research is needed on how this can be achieved appropriately. A first step might be to supplement the use of real data with simulations (Lübke et al., 2020).

ACKNOWLEDGMENTS

The authors thank Paul Hünermund for the inspiration for this study with his presentation on "Causal Inference with Directed Acyclic Graphs", see also https://youtu.be/6ZwarKVgAzQ.

REFERENCES

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin*, *29*(1), 14-27.

Greenland, S. (2021). The causal foundations of applied probability and statistics. *arXiv preprint arXiv:2011.02677*.

Lübke, K., Gehrke, M., Horst, J., & Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, *28*(2), 133-139.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM*, *50*(7), 1113-1123.

Pearl, J., Glymour, M. & Jewell, N.P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27-42.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology*, (Vol. 10, pp. 173-220). Academic Press.

Yanai, I. and Lercher, M. (2021). Novel predictions arise from contradictions. *Genome Biology*, *22*(153).