CAUSAL DIAGRAMS FOR DESCRIPTIVE STATISTICS

Karsten Lübke and Matthias Gehrke
FOM University of Applied Sciences
karsten.luebke@fom.de

Without random sampling and/or random allocation, even descriptive statistics like simple means or proportions can be quite misleading. Therefore, causal diagrams were added to the existing course materials to address this topic to illustrate the differences between random and convenience samples and between observational and experimental studies. We assessed student understanding in different courses with a pre-/post-survey. Additionally, we asked students to evaluate the helpfulness of the diagrams for their understanding. There is a statistically discernible positive effect with more than seven different courses and 280 students from pre- to post-knowledge. Also, most of the students agreed with the statement that the causal diagrams helped in their understanding.

MOTIVATION

Nowadays, data is everywhere. Statistics and data-science education does not only aim to help students analyze the data but also to learn from data for a problem at hand. Hernán et al. (2019) distinguish three different tasks of data science: Description, prediction, and causal inference. Many introductory courses may only cover the first task, description, but as Greenland (2022) pointed out, causality might be even central for description. In her list for promoting statistical literacy, Utts (2021) also emphasizes the topic of observational studies, confounding, and causation. Statisticians are aware of the magic of randomness. As pointed out by, e.g., Cobb (2007): Randomize data to protect against bias. Draw random samples to generalize to populations and use random assignment to support conclusions about cause and effect. The protection against bias by randomization is easy to depict with causal diagrams. Both random sampling and random assignment are erasing arrows pointing into the sampling or treatment variable.

For some time now, there has been a call to include causality in statistics and data-science curricula (e.g., Cummiskey et al., 2020; Greenland, 2022; Kaplan, 2018; Lübke et al., 2020; Schield, 2018). In the current study, we investigate if causal diagrams, even presented in a very informal way, could help students to draw appropriate conclusions. Therefore, we try to add to the available empirical evidence given by e.g., Ellison (2021) about classifying covariates or Reinhart et al. (2022) who explore students reasoning about correlation and causation.

METHOD

Two instructors conducted the study in seven different statistics-related courses, like introduction to quantitative research methods (Bachelor and Master). The students are majoring in business-related subjects. The voluntary, anonymous web-based survey took place during the second lecture of the course within the regularly used classroom response system (https://tweedback.de). The first previous lecture mainly covers organizational and general science topics, so no statistic-specific topics like, e.g., sample and population are covered so far. The pre-assessment takes place at the beginning of the lecture. We confronted the students with the following single question:

> On an internet platform, 10,000 people report a positive effect of a particular shampoo on gray hair (Study A). An experiment with 100 randomly selected people finds no positive effect of the shampoo (Study B). With the information given, the result of which study is more credible?
> A: The result of study A
> B: The result of study B
> C: Both studies are equally credible

At this point in time, we did not show results or the correct answer (B) to the students. During the now following lecture, topics like measurement, random sampling, and random assignment are introduced. The introduction of random sampling and random assignment was supported by presenting causal diagrams, without formally discussing graph elements like nodes and edges. The sampling example is embedded in a fictitious study where a teacher tries to analyze the learning time of her students by a voluntary survey, i.e., a convenience sample. Both learning time and participation in the survey may, e.g., depend on conscientiousness. With a random sample, participation no longer depends on conscientiousness. The arrow from there pointing into the sample is erased and replaced by the researcher's study design, as shown in Fig. 1.

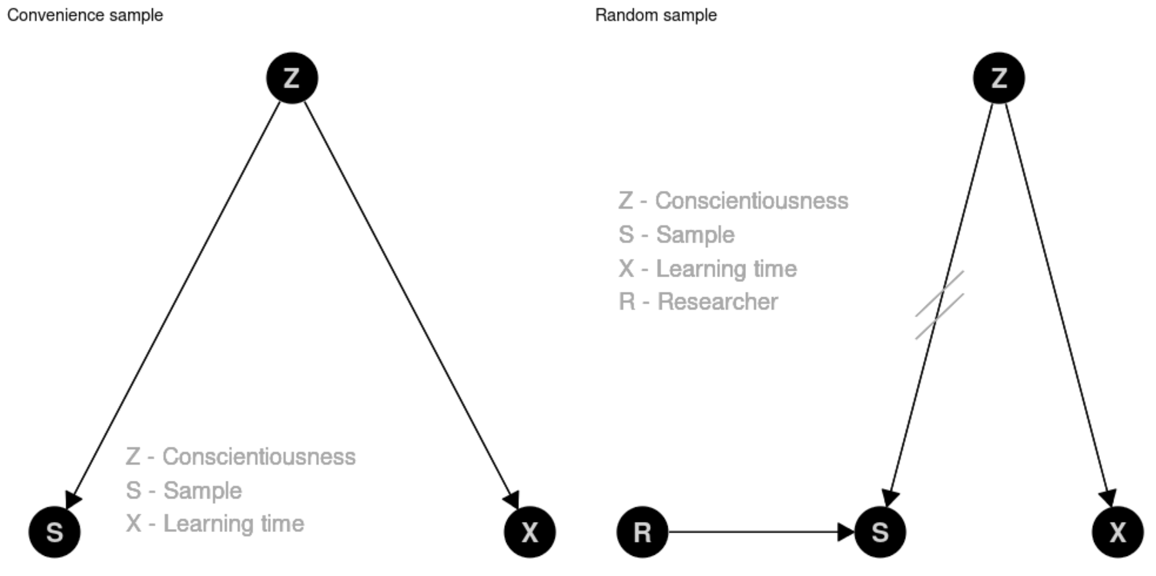

Figure 1. Causal diagram for an example on convenience sample vs. random sample

The example of random assignment within a randomized controlled trial is embedded in a fictitious study where the teacher tries to analyze the relationship between learning time and test score. Prior knowledge is one reasonable confounder here. This confounder may even give rise to Simpson's paradox, i.e., observing a negative correlation between learning time and test score, whereas the true (direct) causal effect is positive. Again, randomness erases the arrow pointing into the treatment (Fig. 2).

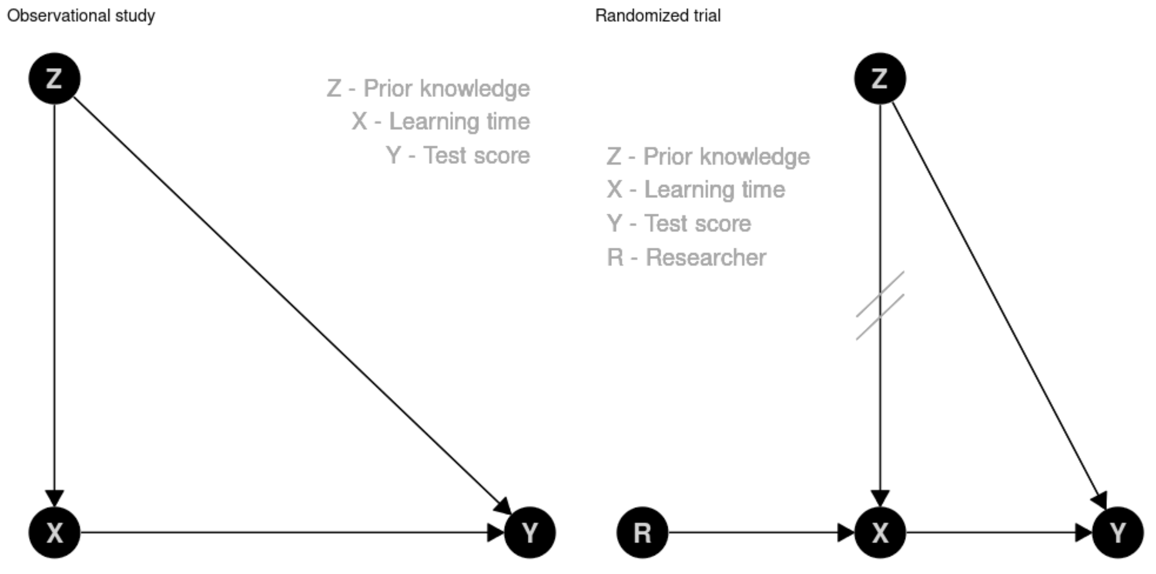

Figure 2. Causal diagram for an example on observational study vs. randomized trial

Both examples provide opportunities to discuss the practical and ethical challenges in conducting these fictitious studies in reality: How can we achieve a random sample? Is it ethical to randomize learning time? What about non-compliance?

For post-assessment, the same question as at the beginning of the lecture is asked at the end of the lecture. As one lecture last two times 90 minutes the post-assessment took place roughly 3 hours after the pre-assessment. To lower the barrier for participation, there has been no attempt to link the results of the pre- and post-assessment.

We also included a short evaluation within the classroom response system to investigate students' perception of causal diagrams. The students were finally asked to rate their agreement on a 5-point Likert scale to the following statement:

> The diagrams (graphs) to describe the data generating process are helpful to understand concepts of data collection (randomized sampling and allocation).

RESULTS

Due to the Covid-19 pandemic, we conducted the survey within synchronous online lectures in the fall of 2021. Data and R code for the analysis are available from https://github.com/luebby/ICOTS-2022. n_{pre} = 282 students took part in the pre-assessment, n_{post} = 280 the post-assessment. n_{eval} = 230 students answered the short evaluation of the helpfulness of the graphs. Due to the study design, it is possible that some students, e.g., answered the post-assessment without answering the pre-assessment. Approximately $^2/_3$ of the attending students answered the survey.

In the pre-assessment only p_{pre} = 49.6 % provide the correct answer (B). In the post-assessment this proportion increases to p_{post} = 62.5 %. This increase of 12.85%-points is statistical discernible (p-value = 0.0012). To calculate the p-value, we permuted pre- and post-responses within each course, and the one-sided p-value for $\pi_{post} - \pi_{\text{pre}} \leq 0$ is calculated based on 10,000 permutations.

One should note that these results are based on an observational study within the classes. We neither did apply randomized controlled trial nor any, e.g., difference-in-difference methods, so we cannot identify a causal effect of including the diagrams versus not-including.

Fig. 3 shows the result of evaluating the helpfulness of the causal diagrams by the students. More than 71 % state, that they fully or strongly agree, and 10 % disagree.

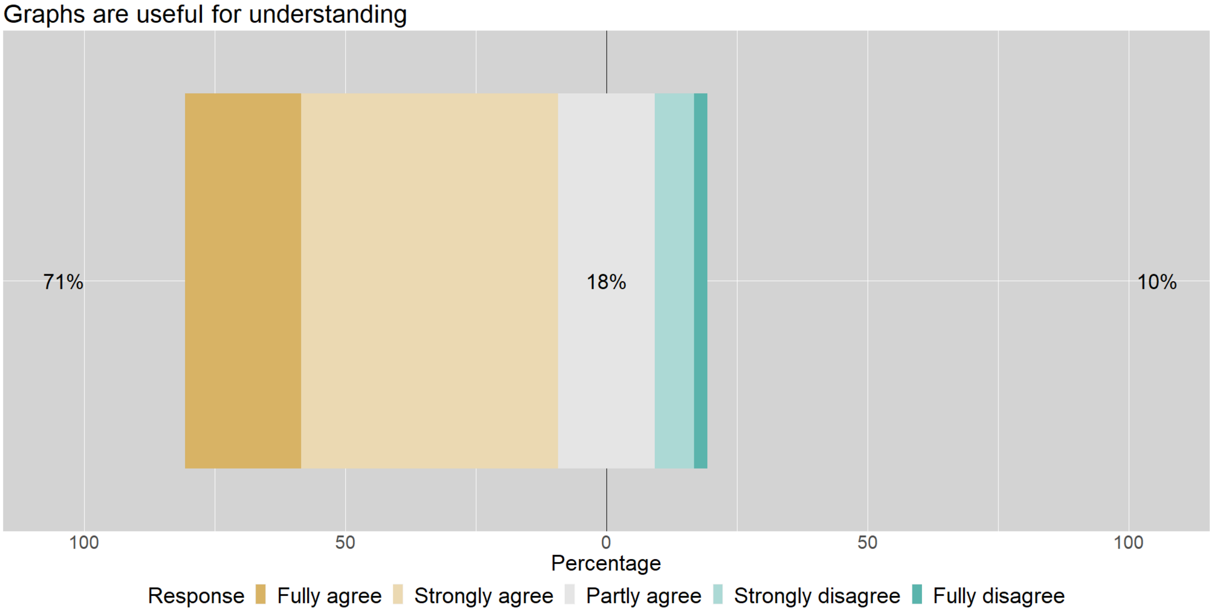

Figure 3. Evaluation of helpfulness

Our experiences as teachers in class are also positive. The causal diagrams of these elementary, fictitious examples provide an excellent opportunity to discuss the crucial topics of confounding and bias. In addition, we experienced lively debates about the merits of randomness in data generation and why this may be hard or even impossible to achieve.

CONCLUSION

In a world full of big data and many studies published based on the analysis of such data, we, as statistics and data-science educators, face the challenge of how we can help our students to draw correct conclusions. The latter is essential even on a purely descriptive level. Causal diagrams may help prepare students not to mess with data and make trustworthy conclusions based on data. We should sensitize them as early as possible that "compensating for quality with quantity is a doomed game" (Meng, 2018). Causal diagrams may provide an easy-to-grasp language to discuss the assumptions about the data-generating process. With these diagrams, instructors can formally assess confounding and bias and illustrate the important benefits of random sampling and/or random allocation. In more advanced courses, one can discuss methods of maybe recovering from bias and confounding in observational data.

The current study has many limitations which should be considered. For example, pre- and post-assessment took place in a single lecture, with only one question. No qualitative data was collected. Also, it is not possible to identify and estimate the causal effect of the intervention by including the diagrams. More and better designed studies are needed to investigate the effect of this teaching approach as well as student understanding and learning. E.g., an open question like "How did the causal diagram help you to come to the answer you choose?" could be added. So, we need more research on how statistics education can provide students with a conceptual framework to scrutinize the data generating process in a data-centric world.

REFERENCES

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1). http://dx.doi.org/10.5070/T511000028

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N. & Watts, K. (2020). Causal inference in introductory statistics courses, *Journal of Statistics Education*, 28(1), 2-8. https://doi.org/10.1080/10691898.2020.1713936

Ellison, G. T. (2021). Might temporal logic improve the specification of directed acyclic graphs (DAGs)?. *Journal of Statistics and Data Science Education*, 29(2), 202-213. https://doi.org/10.1080/26939169.2021.1936311

Greenland, S. (2022). The Causal Foundations of Applied Probability and Statistics. *Probabilistic and Causal Inference: The Works of Judea Pearl* (1st ed.). New York: Association for Computing Machinery, 605-624. https://doi.org/10.1145/3501714.3501747

Hernán, M.A., Hsu J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks, *CHANCE*, 32(1), 42-49. https://doi.org/10.1080/09332480.2019.1579578

Kaplan, D. (2018). Teaching stats for data science. *The American Statistician*, 72(1), 89-96. https://doi.org/10.1080/00031305.2017.1398107

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. https://doi.org/10.1080/10691898.2020.1752859

Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wieczorek, J., ... & Nugent, R. (2022). Think-aloud interviews: A tool for exploring student statistical reasoning. J*ournal of Statistics and Data Science Education*, 1-14. https://doi.org/10.1080/26939169.2022.2063209

Schield, M. (2018). Confounding and Cornfield: Back to the Future. In: *Looking Back, Looking Forward. Proceedings of the Tenth International Conference on Teaching Statistics*, eds. M. A. Sorto, A. White, and L. Guyot. http://www.statlit.org/pdf/2018-Schield-ICOTS.pdf

Meng, X.L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12 (2), 685 -726. https://doi.org/10.1214/18-AOAS1161SF

Utts, J. (2021). Enhancing Data Science Ethics Through Statistical Education and Practice. *International Statistical Review*, 89, 1-17. https://doi.org/10.1111/insr.12446