NOW IS THE TIME FOR CAUSAL INFERENCE IN INTRODUCTORY STATISTICS

Karsten Lübke and Matthias Gehrke
Leimkugelstraße 6
45141 Essen, Germany
karsten.luebke@fom.de

In times of ubiquitous measurement and data, more and more decisions are based on multivariate observational data. Conclusions based on such data are easily flawed, as the rather simple paradoxes of Simpson and Berkson show. Therefore, it seems necessary to integrate elements of causal inference into the curriculum to help students to think beyond data and beyond the mantra of "correlation is not causation". Based on qualitative assumptions about the data generating process, potential omitted variables or selection bias can be made visible by means of directed acyclic graphs. The same holds true for the difference in the joint distribution depending on whether data are observational or from a randomized experiment. In introductory statistics courses, students should learn to distinguish the different levels of causal inference as well as to decide whether to adjust or not to adjust for a covariable.

MOTIVATION

For some time now there has been a call to focus on data (rather than mathematics) in introductory statistics courses (e.g. Cobb & Moore, 1997). GAISE (2016) and ProCivicStat (2018) motivate the use of real, complex, and multivariate data as well as technology to explore such data. We agree with these claims! But data is not just "there". *Data are not just numbers with a context* (Cobb & Moore, 1997), moreover data have a generating process and this process enables and limits the conclusions (and actions) we can and should take from the analysis of data. In the univariate world this thinking was most often limited to some kind of distributional assumptions. Of course, topics like the importance of random sampling and/or random allocation in experiments, usually also covered in introductory courses, are also part of the data generating process – and closely connected to the validity of statistical inference. But due to e.g. digitization there are more and more multivariate observational data, most often generated neither by random sampling nor random allocation, and with that vulnerable to several kinds of biases and confounding. So, within the emerging role of multivariate thinking in inferential reasoning, we think that statistical education should enable students to articulate (and formalize) qualitative assumptions in the framework of causal modeling.

The basic ideas of causal inference, like Directed Acyclic Graphs, the difference between observing and manipulating data, and counterfactual evaluation may foster a deeper understanding of what can and – maybe even more importantly – what cannot be deduced by (observational) data analysis and so may help to refrain from naive conclusions based on "Big-Data" analysis.

In a changing world, where statistical education aims at Data Literacy (e.g. Gould, 2017), now is the time (and need) for a fundamental rethinking of the curriculum (Cobb, 2015). The curriculum reform towards statistical thinking (e.g. Steel et al., 2019) and conceptual understanding by means of e.g. simulation based inference (e.g. Chance et al., 2016), data modeling (e.g. Stigler & Son, 2018), and integration of technology (e.g. Nolan & Temple Lang, 2010) offer a chance so that we concur with Hernán et al. (2019): "We now have a historic opportunity to redefine data analysis in such a way that it naturally accommodates a science-wide framework for causal inference from observational data." Cummiskey et al. (2020) also argue that causal inference fits well into the current guidelines for introductory courses. This means that the classical introductory statistics course is reshaped towards data literacy. I. e., some (mathematical) statistical topics are reduced and causal inference is added to the curriculum. It should be noticed, however, that many statistical methods (means, medians, boxplot, inference, etc.) are now covered in school.

A SHORT INTRODUCTION TO CAUSAL INFERENCE

There are different approaches to and of causality which we will not discuss here. For good introductions to the Potential Outcomes Framework or Directed Acyclic Graphs (DAGs) we refer to e.g. Angrist & Pischke (2014) or Pearl, Glymour & Jewell (2016).

Pearl (2019) distinguishes three levels of causal inference:

- Association: P(y|x): Seeing: *what is?*, i.e., the probability of Y = y given that we observe X = x.
- Intervention: P(y|do(x)): Manipulation: *what if?*, i.e., the probability of Y = y given that we intervene and set the value of X to x.
- Counterfactuals: $P(y_x|x_0, y_0)$: Imagining: *what if I had acted differently?*, i.e., the probability of Y = y if X had been x given that we actually observed x_0, y_0.

In order to take the best action, i.e., maybe an intervention (do(x)), as requested by a data literate individual (e.g. Gould, 2017), she must estimate the effect on y. But without taking the causal model of the data

generating process into account, such a conclusion is easily flawed as the famous paradoxes by Simpson and Berkson show. DAGs can make the assumptions visible and help answer if and how such a causal effect can be estimated from data.

Basic elements are chains (X → C → Y), forks (X ← C → Y), and colliders (X → C ← Y): The expected change in Y resulting of a change of X is unconditional of C for a chain or a collider and it is conditional of C in case of a fork. So, by adjustment (e.g. by integration of a covariable C in a linear model) or not-adjustment (e.g. exclusion of C) one wants to block non-causal paths, not to open spurious paths, and to open causal paths.

For example, consider the data set discussed by Dick DeVeaux in his Stat 101 case study "How Much is a Fireplace Worth?" (https://community.amstat.org/stats101/home): Unconditionally, i.e., not adjusted for house size, a fireplace is worth $65000. But as the existence of a fireplace is associated with house size, the marginal effect goes down to $5560. Assuming that the price depends on house size and fireplace and that the existence of a fireplace depends on the house size, i.e. a model like in Figure 1, students can see that the path from fireplace (X) to price (Y) via house size (C) is non-causal (a "backdoor") so in order to estimate the effect based on this model one should use the effect adjusted for house size. Drawing causal diagrams like Figure 1 helps thinking about possibly omitted variables or selection bias. Note however that in real-life the DAG should include all possible observable or not-observable factors.

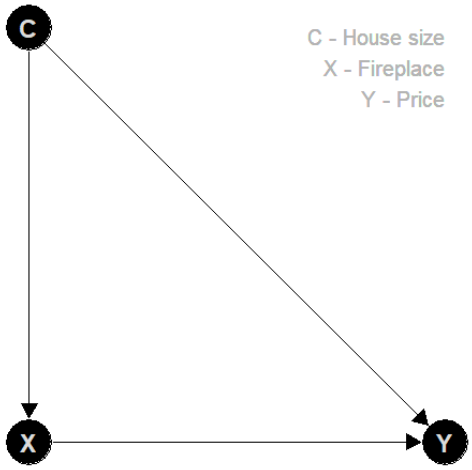

Figure 1. Simplified Directed Acyclic Graph of "How Much is a Fireplace Worth?"

Simulated as well as real-life examples suitable for introductory courses are available (Lübke et al., 2020). Here, in all examples the causal effect of X on Y is investigated. For a chain, we use the simulated and simplified example that learning (X) effects knowing (C) and knowing causes understanding (Y). In the first example, adjusting for C (using it as covariable in a linear regression model explaining Y) would result in a small effect for X, only. But this is not true as Y is only conditionally independent given C. One must leave out C to get the real effect of X on Y.

Another simulated example uses intelligence (C) which reduces learning time (X) and increases test score (Y). If one wants to reveal the true effect of X on Y then one must adjust for C. C is a fork as given in the example above ("How Much is a Fireplace Worth?"), both are examples for the Simpson's paradox.

The last element, a collider, is simulated via ability to network (X), competence (Y), and promotion (C). X and Y are independent, but promotion depends on both. Adjusting for C (adding it to the linear model) would result in a bias (spurious regression between X and Y) and, therefore, would give incorrect results. This is an example of Berkson's paradox.

All three simulated examples can easily be demonstrated via linear models (regression) and are straightforward to be followed by the students. As such, this helps students to understand causal effects, (linear) modelling, and how much this depends on the data generating process. Additionally, the effect of an experimental setup to study causal effects can be demonstrated with the simulated examples as well.

For real-life examples we use e.g. the Saratoga House data set (the data set also used in the Stat 101 case study mentioned above). The price of a house could be modelled by number of bedrooms. But this would not result in the true price as the size of the house is missing. The size has a direct effect on the price and an indirect effect via number of bedrooms, therefore it has to be considered as a fork and needs to be adjusted for.

Another real-life example is the case presented by Cummiskey (2019) (see also Cummiskey et al., 2020) which investigates the effect of teenage smoking on lung function. The height of the teenager is on the

causal path between smoking and forced expiratory volume, constituting a chain. To find the correct effect of smoking on lung function one must not adjust for height.

INTEGRATION INTO THE SYLLABUS

With the beginning of winter term 2019/2020 the topics covered in our courses called "Scientific Research Methods – Quantitative Data Analysis" for students studying a bachelor's degree in business administration with 8 ECTS (European Credit Transfer and Accumulation System, which allows students to transfer study results within Europe) points. There are no specific prerequisites, just the normal higher education entrance qualification is needed. The course is organized as follows:

- Background: Science
- Foundations of Quantitative Data Analysis
- Introduction to R
- Exploratory Data Analysis
- Linear Regression
- Introduction to Statistical Inference
- Introduction to Causal Inference (pilot test)

It should be noted that our students are studying while working and that the ability to produce statistical results (Gould, 2010) is a learning outcome according to the curriculum. The technology used to teach these topics, integrated in the statistical-inquiry cycle (Wild & Pfannkuch, 1999) as well as Data Science pipeline (e. g. Wickham & Grolemund, 2017) are R mosaic (Pruim et al., 2017), interactive shiny apps (e.g. Doi et al., 2016) as well as learnr tutorials (Schloerke et al., 2019). Case studies (e.g. Neumann et al., 2013) as well as quizzes (e.g. McGowan & Gunderson, 2010) are also integrated as well as reproducible analysis via R Markdown (Baumer et al., 2014).

In the lecture "Background: Science" students are introduced to concepts like internal and external validity as these are fundamental for drawing conclusions based on data. Here also the data modelling theme *Data = Model + Rest* is introduced. These topics are repeated and connected to (random) sampling and (randomized) experiments in the "Foundations of Quantitative Data Analysis". The data modelling theme is iterated via the R formula Y ~ X in the "Introduction to R", "Exploratory Data Analysis", and "Linear Regression" lectures. Statistical Inference, i.e., estimation and testing is introduced informally via bootstrapping and permutation in the "Linear Regression" lectures but formally defined – and connected to sampling, experiments as well as internal and external validity in the "Introduction to Statistical Inference" lectures.

All these fundamental topics are re-iterated and rounded up in the last, new block about "Introduction to Causal Inference": Using DAGs and simple simulated examples with (multiple) linear regression, we show how convenience sampling can introduce bias and how randomized experiments avoid confounding.

By a framework to think about the (multivariate) data generating process using DAGs we hope to develop key ideas and skills related to multivariate thinking. Mainly, these are knowing how confounding and bias arise so that paradoxes like Simpson's and Berkson's are better understood.

EVALUATION

The desired learning outcomes are two-fold: improved knowledge about conclusions based on data and attitude towards statistics and data.

We have already started a survey (also among data science practitioners) where participants are asked about their conclusion about the effect of an intervention based on the usual regression output, with and without the corresponding DAG. Here, we are interested in finding out how many respondents erroneously draw causal conclusions from regression results, and how this number changes when the DAG that enables such a conclusion is given. The structure of our university with 29 study centers allows us to compare results obtained from students who have already learnt about causal inference with those to whom causal inference has not yet been taught.

Currently we are designing a study to evaluate the effect on post course value of statistics, see e.g. Schield (2018) for current problems. As in the survey referred to earlier, this should be done by means of quasi-experimental results for those who learn about causal inference in comparison with those who do not.

CONCLUSION

The Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2016) recommend to "Teach statistical thinking" by "Teach[ing] statistics as an investigative process of problem-solving and decision-making" and to "Give students experience with multivariate thinking" and also "Integrate real data with a context and purpose". Similar ideas include the aim to focus on modelling (e.g. Stigler & Son, 2018) and to support students "to think with data" (Pruim et al., 2017).

However, multivariate modelling can be misled through the presence of confounding. Given the ubiquitous presence of data in current times and the fact that the field of applied statistics has changed a lot since Fisher's days, we think that today's students should learn to think even more thoroughly about the data generating process in order to draw conclusions from data.

Through the "Focus on conceptual understanding" (GAISE, 2016) by teaching techniques such as simulation-based inference (bootstrapping, permutation test) and by de-emphasizing more traditional inference techniques, some free time is gained during the curriculum. Therefore, our proposed curriculum places data and context in the main focus. We agree with Wild et al. (2018): “The mission of statistics education is to provide conceptual frameworks (structured ways of thinking) and practical skills to better equip our students for their future lives in a fast-changing world.“ We think that DAGs can help developing such frameworks in the emerging role of multivariate thinking in inferential reasoning. With a focus on the statistical inquiry cycle in a holistic manner, data modelling, simulation-based inference, and the usefulness of technology, a consistent and integrated curriculum as well as active learning is possible. The syllabus has been reshaped towards the aim of multivariate thinking within the statistical-inquiry cycle.

However, teachers of such courses need to refocus and maybe minimize or omit other topics (GAISE, 2016). These losses must also be considered. In an ideal world for statistical education, we would have enough time to cover all our topics, but in the meantime we have to decide what to focus on – and to consider the opportunity costs associated with our decision (Cobb, 2007). As most of today's lecturers have not learned causal inference for themselves, training resources are required. Nevertheless, the notion of integrating causal modelling in introductory statistics is supported by many, e.g. Ridgway (2016), Angrist & Pischke (2017), Kaplan (2018), Hernán et al. (2019), Cummiskey et al. (2020), and the ASA “Causality in Statistics Education Award”.

ACKNOWLEDGMENTS

We thank Bianca Krol, Sebastian Sauer, Norman Markgraf, Jörg Horst, Christian Röver, Gero Szepannek, and numerous other colleagues for their contribution in the proposed change of the curriculum and for helpful comments in order to improve the teaching materials and this manuscript.

REFERENCES

Angrist, J.D. & Pischke, J.S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Angrist, J.D. & Pischke, J.S. (2017). *Undergraduate econometrics instruction: Through our classes, darkly. Journal of Economic Perspectives,* 31(2), 125–144.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L. & Horton, N.J. (2014). R markdown: Integrating a reproducible analysis tool into introductory statistics. *Technological Innovations in Statistics Education*, 8(1).

Chance, B., Wong, J. & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, 24(3), 114–126.

Cobb, G.W. & Moore, D.S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104(9), 801–823.

Cobb, G.W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, 1(1).

Cobb, G.W. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69(4), 266–282.

Cummiskey, K. (2019). Causal inference in introductory statistics courses. *https://github.com/kfcaby/causalLab*

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N. & Watts, K. (2020). Causal inference in introductory statistics courses, *Journal of Statistics Education,* DOI: 10.1080/10691898.2020.1713936

Doi, J., Potter, G., Wong, J., Alcaraz, I. & Chi, P. (2016). Web application teaching tools for statistics using R and shiny. *Technology Innovations in Statistics Education*, 9(1).

GAISE (2016): GAISE College Report ASA Revision Committee: Guidelines for assessment and instruction in statistics education (GAISE) College Report 2016.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, 78(2), 297–315.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal,* 16(1), 2–25

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.

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*, DOI: 10.1080/10691898.2020.1752859

Kaplan, D. (2018). Teaching stats for data science. *The American Statistician*, 72(1), 89–96.

McGowan, H.M. & Gunderson, B.K. (2010). A randomized experiment exploring how certain features of clicker use effect undergraduate students' engagement and learning in statistics. *Technology Innovations in Statistics Education*, 4(1).

Neumann, D.L., Hood, M., Neumann. M.M. (2013). Using real-life data when teaching statistics: student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, 12(2).

Nolan, D. & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64(2), 97–107.

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

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60.

ProCivicStat Partners (2018). Engaging civic statistics: A call for action and recommendations, *http://iase-web.org/islp/pcs/.*

Pruim, R., Kaplan, D.T. & Nicholas J Horton, N.J. (2017). The mosaic package: Helping students to 'think with data' using R. *The R Journal*, 9(1), 77–102.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review,* 84(3), 528–549.

Schield, M. (2018). Confounding and Cornfield: Back to the future. In M. A. Sorto et al. (eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics.*

Schloerke, B., Allaire, J.J. & Borges, B. (2019). learnr: Interactive tutorials for R. R package version 0.10.0. *https://CRAN.R-project.org/package=learnr*

Stigler, J.W. & Son, J.Y. (2018). Modeling first: A modeling approach to teaching introductory statistics. In M. A. Sorto et al. (eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics.*

Steel, E.A., Liermann, M. & Guttorp, P. (2019). Beyond calculations: A course in statistical thinking. *The American Statistician*, 73(sup1), 392–401.

Wickham, H. & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media, Inc.

Wild, C.J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–248.

Wild, C.J., Utts, J.M. & Horton, N.J. (2018). What is statistics? In: Ben-Zvi, D. et al. (eds.), *International Handbook of Research in Statistics Education*, Springer International Handbooks of Education, 5–36.