DRAWING (CAUSAL) CONCLUSIONS FROM DATA – SOME EVIDENCE

Bianca Krol, Karsten Lübke, and Sandra Sülzenbrück
FOM University of Applied Sciences, Leimkugelstr. 6, 45141 Essen, Germany
karsten.luebke@fom.de

In a data-driven company stakeholders should agree on the evidence the data provides. Data literacy includes that one should be able to draw conclusions also 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 job. As the Simpson paradox shows, the same data can lead to opposite conclusions being adjusted or not. The correct conclusion depends on the qualitative assumptions about the data generating process.

To investigate the conclusions drawn by young professionals, a 2x2 randomized experiment is conducted. The same numeric data is presented in two different contexts with different structural causal models so once the adjusted and once the unadjusted effect is appropriate. The other randomized factor varied is whether a directed acyclic graph is presented before or after the numerical summary.

Preliminary data of this ongoing research indicate that, as expected, the conclusions drawn from the same data differ by context and presentation. Also, the assumed causal model and whether the adjusted or not adjusted difference is the true (causal) effect are far too often not consistent. Results seem to indicate that within a data science project it is advisable to discuss the assumptions of the data generating process first and then present the numerical summary.

A consequence for data literacy education is that the process of data modeling should be a central part of the curriculum. More emphasis should be put on the mapping and link between subject matter knowledge and statistical modeling. We should provide a framework to discuss this science with data with all stakeholders. Further research is needed on how such a “bridge” can be built.