More confidently uncertain? Teaching learners to apply Bayesian methods to make sense of scientific phenomena

Joshua M. Rosenberg, University of Tennessee, Knoxville

Abstract: While analyzing data is an important learning in science domains, existing methods and tools for those learning to work with data have key limitations, particularly concerning scientific modeling. This early-stage research is intended to begin a line of work on students' data analysis that is not yet widely used in K-12 learning environments, Bayesian statistical methods, with implications for how learners use evidence in science education.

Goals of the Research

The goals of the research are to support science students to make sense of their world through Bayesian methods. This involves middle and secondary classroom-based design research, teacher professional development, and curriculum and statistical software development. If successful, this work has the potential to resolve difficulties with respect to how learners at the K-12 level can work with data in ways that are meaningful and empowering to them and to better prepare learners for subsequent coursework or areas of study that involve quantitative or computational dimensions. This work also has the potential to broaden conceptions of the courses in which complex work with data takes place, from mathematics and statistics to science classes.

Background

Working with data and data literacy are important capabilities individuals' everyday lives and in many domains, particularly scientific and engineering (Lovett & Shah, 2007). There have been a number of efforts to make data literacy more accessible to more individuals, including pedagogical approaches (Lee & Wilkerson, 2018; Lehrer & English, 2018). One set of strategies focuses on curricula (Lehrer & Schauble, 2004), while others focus on developing *tools* to make it easier for students to work with data (Konold et al., 2017).

While these tools that have been developed have strengths, none allows students to fully engage in the kind of data analysis activities that professionals do. A commonly-used tool, the Common Online Data Analysis Platform (CODAP) is well-suited to exploring data, but does not allow for more sophisticated data processing pipelines, like analyzing data from two sources. The SAGE Modeler tool allows for pipelines but is designed to be used primarily for computational models of complex systems, and therefore is not designed for the most widely-used statistical analyses. In the context of these tools available for learners to work with data and *technical* opportunities for enhancements, there are also *cognitive* challenges that learners face. Modeling data, and doing so to explain variation, is a core capability, but is also challenging for learners (Lehrer & English, 2018), who may tend to represent *all* of the data they collect, instead of summarizing the data with a statistic or model (Lehrer & Schauble, 2004), or may find hypothesis-testing to be counter-intuitive.

One solution to the technical and learning-related challenges is an approach that has not been widely used outside of the professional practice of statisticians, data scientists, and engineers, Bayesian methods. Bayesian methods have become popular because they allow scientists and engineers to understand and make statements about how likely phenomena in the world are--even highly complex phenomena, like voting in elections, when births occur within the year and over time, and where radon in the soil beneath homes may be a hazard to individuals' health (Gelman et al. 2013). While developmental psychologists have explored Bayesian thinking on the part of young children (see Gopnik, 2012, for a review), this work has often taken place outside of K-12 learning environments. Yet, Bayesian methods provide a helpful way to approach work with data and data literacy in terms of modeling, one which emphasizes making inferences about phenomena, in addition to (or instead of) carrying out hypothesis tests as a part of more commonly-used statistical approaches.

Methodology and Preliminary Findings

During the 2018-2019 academic year, I engaged in preliminary research: developing curricula and related to involving students in analyzing data sources from scientific studies. I focused on using CODAP and helping students to work toward developing data-based scientific explanations for phenomena. This work was encapsulated in a practitioner-focused article that described (for teachers) how students can use CODAP to analyze complex (and even messy) data in secondary science classrooms (Rosenberg et al., 2020).

In the 2019-2020 year, we continued this work, with an emphasis on not only analyzing already-existing data from scientific studies but also supporting students to analyze data that they collect (and to record and

structure this data). I also have used the R and R Markdown software with students in the classroom of a collaborating teacher. This work has presented new opportunities (such as for students to report that the work was akin to their prior experiences in programming) as well as challenges (such as how to remain focused on being able to answer scientific questions, rather than emphasizing how to write code in R and how to use R.

At the same time, during the 2019-2020 year, I have been involved in projects that are complementary to this work and which suggest new directions for it. In particular, I began a National Science Foundation-funded project focused on providing greater opportunities for students to be involved in computer science (in light of new K-8 computer science and computational thinking standards in the state of Tennessee). I have also begun to explore the role of data science in education through publications (Dai, Rosenberg, & Lawson, in press; Rosenberg et al., 2020). This work has collectively emphasized the potential utility to students of more accessible and useful statistical methods in science education contexts.

This work over the past two years informs the method for the proposed line of work in support of supporting science students to make sense of their world through Bayesian methods. The proposed work will involve working with more than the one collaborating teacher with whom I have worked; I am planning to work together with four area teachers. Design-based research is essential for developing curricula and pedagogical strategies, but a more systematic research strategy is also needed. The research plan will involve designing curricula for an approximately eight-week unit. The work will also involve modifying CODAP to address the main limitation I found from its use in prior research projects: making it easier to use as a part of a data processing and analysis workflow, through adding functionality to write code which can be run and reproduced.

Expected Contributions

Particularly in science education, an approach to modeling that emphasizes understanding how likely different outcomes or values has the potential to resolve challenges. A key expected contribution has to do with Bayesian methods being more accessible to learners. Another contribution concerns how existing tools may need to be changed to support a Bayesian approach to modeling and inference may be designed in line with what is known about the development of statistical software that is useable by learners (McNamara, 2019). As a technique that depends on computation, this work has some implications for research on computational thinking, especially in science education (Sengupta et al., 2013). Last, this work may contribute to data science education, especially within science education contexts, but also as a set of capabilities that may useful to learners across the curriculum.

References

Dai, T., Rosenberg, J. M., & Lawson, M. A. (in press). *Data representations and visualizations in educational research*. In T. L. Good & M. McCaslin (Eds.), Educational Psychology Section; D. Fisher (Ed.), Routledge Encyclopedia of Education (Online). Taylor & Francis.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal, 16*(2).

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects*. Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, D.C.

Lehrer, R., & English, L. (2018). *Introducing children to modeling variability.* In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), International handbook of research in statistics education (pp. 229-260). Springer.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635-679.

Lovett, M., & Shah, P. (Eds.). (2007). *Thinking with data*. Psychology Press.

McNamara, A. (2019). Key attributes of a modern statistical computing tool. *The American Statistician, 73*(4), 375-384.

Rosenberg, J. M., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher, 87*(5).

Rosenberg, J. M., Lawson, M. A., Anderson, D. J., Jones, R. S., & Rutherford, T. (2020). Making data science count in and for education. In E. Romero-Hall (Ed.), *Research Methods in Learning Design & Technology*. Routledge.

Sengupta, P., Kinnebrew, J. S., Basu, S., Biswas, G., & Clark, D. (2013). Integrating computational thinking with K-12 science education using agent-based computation: A theoretical framework. *Education and Information Technologies*, *18*(2), 351-380.