DATA LITERACY IS STATISTICAL LITERACY

ROBERT GOULD
University of California, Los Angeles
rgould@stat.ucla.edu

ABSTRACT

Past definitions of statistical literacy should be updated in order to account for the greatly amplified role that data now play in our lives. Experience working with high-school students in an innovative data science curriculum has shown that teaching statistical literacy, augmented by data literacy, can begin early.

Keywords: *Statistical literacy, Data literacy, Data science.*

STATISTICAL LITERACY

Many traditional notions of statistical literacy (SL) are about understanding the data representations and statistical arguments of others. The set of knowledge and understanding required to be statistically literate is often defined by differentiating the needs of consumers of statistics from those of producers of statistics, a dichotomy that goes back at least as far as Hotelling (1940). For example, Gal, focusing on adults, defines statistical literacy in consumer terms as people's "ability to interpret and critically evaluate" statistical products, as well as their ability to "discuss or communicate their reactions" to statistical products (Gal 2002, p. 2–3). Citizenship is often cited as a reason for advancing statistical literacy, since democracies require informed debate, and almost all policy discussions require some statistical understanding.

I feel that the basic SL goal of developing critical consumers of statistics produced by others is worthy (and necessary), but falls far short not only of what is required for life in modern democracies, but also in terms of what is possible for today's students to achieve. The needs of modern students has grown: all students should be educated to perform the dual role of statistical producer and consumer.

More specifically, I argue for an augmented definition of SL that includes, at a minimum:

- understanding who collects data about us, why they collect it, how they collect it;
- knowing how to analyze and interpret data from random and non-random samples;
- understanding issues of data privacy and ownership;
- knowing how to create basic descriptive representations of data to answer questions about real-life processes;
- understanding the importance of the provenance of data;
- understanding how data are stored;
- understanding how representations in computers can vary and why data must sometimes be altered before analysis; and
- understanding some aspects of predictive modeling.

This augmented notion of SL is necessary because the role and nature of data have changed. Data are now ubiquitous in all aspects of life for all people, not just policy. Also, the tools needed to access and analyze data are inexpensive. Research and discussion should therefore focus on what minimal level of understanding of the topics in the above list is necessary for citizenship.

In 2015 I participated in a workshop on data literacy, hosted by the Oceans of Data Institute (ODI). One result of that meeting was a definition of a data-literate individual as a person who

> "understands, explains, and documents the utility and limitations of data by becoming a critical consumer of data, controlling his/her personal data trail, finding meaning in data, and taking action based on data. The data-literate individual can identify, collect, evaluate, analyze, interpret, present, and protect data." (ODI 2015).

I found this definition striking in that it seems to include statistical literacy and go far beyond (perhaps a function of the composition of the participants). One frustrating aspect of the meeting was the extent to which this definition was driven not by statisticians or statistics educators, but by professional data scientists who did not have statistics degrees. Quite plausibly, this is because notions of SL that have arisen from the statistics education community are perceived by those who work in data science as falling short of what is required.

PARTICIPATORY SENSING

My understanding of statistical literacy has been greatly influenced by my participation over the last few years in the Mobilize project. Funded by the National Science Foundation, Mobilize is a partnership between the Los Angeles Unified School District and the University of California, Los Angeles (UCLA), Department of Computer Science, Department of Statistics, and the Graduate School of Education and Information Sciences. Mobilize originated in the Computer Science department as a project designed to enhance computational thinking skills in secondary-level mathematics and science courses. From its inception, Mobilize had a substantial data analysis component that grew over time and ultimately became the core concept behind the Introduction to Data Science (IDS) curriculum. IDS is a yearlong course that teaches secondary-school students the computational and statistical thinking skills necessary to become both consumers *and* producers of statistics. In the last three years, it has been taught in over 30 classrooms in Los Angeles, and in 2017 it will expand to school districts beyond the original project partner.

IDS makes use of a data collection paradigm called *participatory sensing* Participatory sensing (PS) enables individuals to use mobile devices to collect data, and strives to create a community that shares data and analyses in order to understand common concerns (Burke et al., 2006). I like to think of participatory sensing as a democratic version of citizen science. In citizen science, data tends to flow from the many collectors to the few scientists/analysts. In participatory sensing, data flows between all participants equally, and all are invited and encouraged to analyze the data. There is an exception; each participant must take explicit action to share her data. Otherwise, her data remain private and are visible only to her.

What sort of data can be collected with current mobile devices? In one Mobilize participatory sensing campaign, students collected data whenever they discarded an item. At that moment, they opened a survey tool on their phone to collect data about that action. These data varied in type, and included

- Categorical variables related to questions such as *Was the item recyclable or compostable or meant for a landfill? Was it placed in a recycling bin or a trash bin or a compost bin?*;
- Numerical variables, e.g., *How many recycling bins were visible when you discarded this item? How many trash bins?*; and
- Text variables related to questions such as *Describe the item. What activity generated the need to discard it?*

In addition, the students took a picture of the item and the phone automatically recorded time, data, and location.

What understandings and knowledge do students need to make sense of data such as these? These data are not a random sample from any population, and yet they contain rich information. These data are complexly structured, with multiple observations nested within the individual student. Some students might (and do) collect many observations, and some might (and do) collect no observations. Some students might adhere to the data collection algorithm, while others might not. Much of the data are not only "numbers in context" (Cobb & Moore, 1997), but also images, texts, times, and places.

Participatory sensing data have important advantages when used in the classroom, despite the challenges they pose. First, issues of data quality and the ability to draw conclusions from data that might not originate from rigorous collection schemes are not abstract, as is the case when students work with data collected by professional researchers. These issues play a lively role in classroom discussions. Second, issues of data privacy and ownership are brought to the forefront. We were pleased to see many students demonstrate a fairly sophisticated understanding of the notion of data ownership, such as how ownership is affected by where the data were stored and what agreements existed between those collecting the data and those storing the data. Third, PS tools, such as the technology suite developed for the Mobilize project (Tangmunarunkit et al., 2015) provide classrooms the ability to carry out their own data collection campaigns on topics of their own choosing.

Finally, PS data are useful because the very characteristics that make them complicated to use in the classroom (non-numerical data types, non-random sampling schemes, complex structures) are characteristics shared by much of big data. Big data are ubiquitous in our society, and developing SL in the context of big data is equally important as developing SL with more traditional data types.

ALGORITHMIC CULTURE

While students must still learn to read and evaluate tables and graphics presented in newspapers and other forums, much of the statistics they encounter in their daily lives belong to what Breiman (2001) called the "algorithmic culture", which stands in contrast to the traditional inference culture. Algorithms that aggregate our news feeds, recommend products to buy, select advertisements for us to view, or determine which news items are fake or real, rely on statistical methods that, though rarely taught at the secondary level, are important for statistically literate consumers to understand.

For example, the IDS curriculum teaches a basic understanding of classification and regression trees (CART). Classification problems are arguably more easily understood than regression problems and are inherently multivariate. Classification: Can we predict which patients will die within 72 hours in an intensive-care ward? Regression: Which factors are linearly associated with death within 72 hours in an intensive-care ward, and what do their parameter values tell us about the multivariate relationship?

The CART algorithm itself is accessible to high-school students, arguably more so than univariate linear models with least squares, and definitely more so than multivariate linear models. Admittedly, interpreting the output of CART – which is usually visualized as a tree – does require some sophistication. High-school mathematics curricula often use statistical tables develop proportional reasoning, but one could do the same with CART diagrams and, in doing so, also enhance algorithmic statistical reasoning.

To conclude, our understanding of what is the minimal level of statistical knowledge needed by all people, regardless of their profession, professional aspirations, or social class, must be greatly augmented in recognition of the fact that the role played by data in our daily lives is changing dramatically.. The concept of data literacy goes a great distance towards providing the statistically literate individual with the skills and understanding she needs in order to participate in a society that frequently collects data about her and uses it to make predictions about her consumption and social patterns. Enhancing the notion of statistical literacy with that of data literacy allows for the development of citizens who can access and analyze data from government or from their own personal sensors in order to answer their own questions, giving them a powerful voice in a democratic society.

ACKNOWLEDGEMENTS

I thank the Mobilize team for its support. The Mobilize project is based upon work supported by the National Science Foundation under grant number 0962919.

REFERENCES

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.
[Online: projecteuclid.org/euclid.ss/1009213726]

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M. B. (2006). *Participatory sensing*. Los Angeles: Center for Embedded Network Sensing (UCLA). [Online: escholarship.org/uc/item/19h777qd]

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

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–25.

Hotelling, H. (1940). The teaching of statistics. *Annals of Mathematical Statistics, 11*(4), 457–470. [Online: projecteuclid.org/euclid.aoms/1177731833]

Oceans of Data Institute (2015). *Building global interest in data literacy: a dialogue*. Waltham, MA: Educational Development Center. [Online: oceansofdata.org/our-work/building-global-interest-data-literacy-dialogue-workshop-report]

Tangmunarunkit, H., Hsieh, C. K., Longstaff, B., Nolen, S., Jenkins, J., Ketcham, C., Selsky, J., Alquaddoomi, F., George, D., Kang, J., Khalapyan, Z., Ooms, J., Ramanathan, N., Estrin, D. (2015). Ohmage: a general and extensible end-to-end participatory sensing platform. *ACM Transactions on Intelligent Systems and Technology, 6*(3), Article 38, 21 p.

ROBERT GOULD
Dept. of Statistics, MC 951554
UCLA
Los Angeles, CA 90095-1554, USA