STATISTICAL LITERACY FOR ACTIVE CITIZENSHIP: A CALL FOR DATA SCIENCE EDUCATION

JOACHIM ENGEL
Ludwigsburg University of Education
engel@ph-ludwigsburg.de

ABSTRACT

Data are abundant, quantitative information about the state of society and the wider world is around us more than ever. Paradoxically, recent trends in the public discourse point towards a post-factual world that seems content to ignore or misrepresent empirical evidence. As statistics educators we are challenged to promote understanding of statistics about society. In order to re-root public debate to be based on facts instead of emotions and to promote evidence-based policy decisions, statistics education needs to embrace two areas widely neglected in secondary and tertiary education: understanding of multivariate phenomena and the thinking with and learning from complex data.

Keywords: *Statistical literacy; Data science education; Multivariate reasoning; Open data*

INTRODUCTION

These are exciting times for active citizen participation in public decision processes, for influencing public policies and for social action. Digital media and the accessibility of data are changing our access to information and shaping the political discourse. More than 4000 people gathered in Paris in December 2016 for the Open Government Partnership Global Summit that was dedicated to making government decision-making more open, inclusive and responsive (see Open Government Partnership, n.d.). The World Wide Web is a powerful influence on the public access to open data and government documents, the tactics and content of political campaigns, the behavior of voters and the efforts of activists to circulate their messages. The explosion of social media is fuelling new and unanticipated directions in e-democracy and e-participation – from increased pressure for direct democracy and new ideas to engage citizens in service to direct democracy.

Completely contrary to these new opportunities, a growing disdain for factual knowledge in the public discourse has been observed lately, accelerated through the proliferation of fake news in social media, which are disseminated without journalistic fact-checking filters. “Post-truth” has been chosen as “word of the year” by the Oxford Dictionaries to denote circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. In a climate of anxiety, confusion, and nostalgia of an idealized past, as the former president of the European Research Council Helga Nowotny (Nowotny, 2016) notes, raw emotions prevail implying a poor way to cope with uncertainty, which stands in striking contrast to the methods of science and free inquiry. Notably during the UK Brexit referendum and the US presidential election campaign false utterances and outright wrong claims based

Statistics Education Research Journal, 16(1), 44-49, http://iase-web.org/Publications.php?p=SERJ

on misrepresented statistics reached a disturbing level. Fortunately, there are groups dedicated to checking the accuracy of claims made by elected officials, who produce commentaries based on careful analyses of data sources and interpretations (Ridgway, Arnold, Moy, & Ridgway, 2016). One such group, PolitiFact (Tampa Bay Times, n.d.) provides more than 1200 accounts relating to US politics in 2016 alone. As an illustrative sample, look at the following data-related statements

- "Inner city crime is reaching record levels" (Donald Trump, August 30, 2016).
- "The unemployment rate for African-American youth is 59 percent" (Donald Trump, June 20, 2016).
- "Two million more Latinos are in poverty today than when President Obama took his oath of office less than eight years ago" (Donald Trump, August 25, 2016).
- "We are now, for the first time ever, energy independent" (Hillary Clinton, August 10, 2016).
- "If you look worldwide, the number of terrorist incidents have not substantially increased." (Barack Obama, October 13, 2016).
- "Illegal immigration is lower than it's been in 40 years." (Barack Obama, March 17, 2016).

It is not the purpose on this short note to discuss the rights or wrongs of any of these claims. The interested reader is encouraged to explore the details on the above website which includes some statistical arguments and reference to open data sources. Instead, we discuss the statistical knowledge and data-related skills required for a typical fact check on statements like these.

UNDERSTANDING MULTIVARIATE PHENOMENA ABOUT SOCIETY

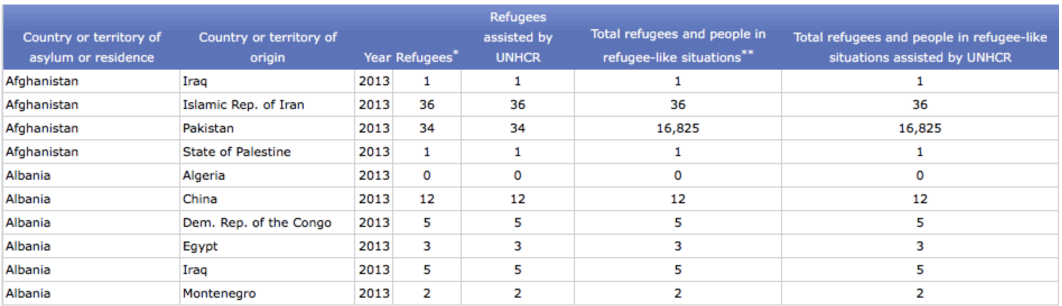
An enlightened citizenry that is empowered to study evidence-based facts and that has the capacity to manage, analyze and think critically about data is the best remedy for a world that is guided by fake news or oblivious towards facts. Nowadays massive amounts of data on important societal topics are increasingly accessible to the general public and to individual citizens or social action groups, on a huge range of topics including migration, employment, social (in)equality, demographic changes, crime, poverty, access to services, energy usage, living conditions, health and nutrition, education, human rights, and many others. To understand these topics is essential for civic engagement in modern societies, but achieving understanding involves statistics that often are open, official, and multivariate in nature; patterns in data can change over short time periods – such data are usually not part of regular statistics instruction, neither at the secondary nor at the tertiary level. Making sense of these statistics requires the ability to explore, understand, and reason about complex multivariate data, because social phenomena do not happen in a vacuum, and their understanding requires awareness of how variables co-vary, or affect each other, or are situated in a network of causal factors that may change over time in manifold ways.

Making sense of multivariate data does not necessarily involve advanced sophisticated multivariate statistical procedures as often applied in social science research (e.g., factor analysis or logistic regression). Rather, it involves understanding multivariate *phenomena* and is based on developing sound heuristics, including awareness of biases and fallacies such as Simpson's paradox, conditional probabilities, or misinterpreting correlations. Also, contextual knowledge is indispensable. Exploring authentic data on topics about society provokes discussion and reflection about central statistical issues like operationalization of relevant variables and their measurement, methods of data collection, the choice of relevant covariates and the role of potential confounders. How can one develop concrete measurable definitions for the concepts involved? How does one define deprivation of cultural rights, economic discrimination or restricted access to public services? Even commonly used concepts like the unemployment rate are far from trivial, and official measures and definitions do not always correspond to the everyday perception of the concepts. Reasoning about large-scale multivariate datasets and understanding their features via graphical representations requires different skills than the analysis of small samples that dominate today's curricula. It involves understanding various multivariate data displays such as scatter-plot matrices, classification and regression trees, heat maps and time-animated or color-coded (Peng, 2008) multivariate time series. Making sense of multivariate data necessitates an awareness of other constructs that are seldom included in the introductory statistics materials at the high-school level, such as interactions, confounders, and non-linear relations.

THINKING WITH AND LEARNING FROM COMPLEX DATA

For all of the six sample statements introduced above, official data are available from government sources (search for the data via Tampa Bay Times, n.d.). Imported into suitable user-friendly software, there are still serious obstacles to overcome to check the claims made. Relevant data from the web rarely come in the tidy format that we are used to in our traditional statistics class. Textbook data usually consist of a (often not even very large) number of rows (the cases) and have only a few columns (the variables or attributes), but real data about society are often more complex and messy. While traditional data are "numbers in context" (Moore, 1990), data from the web may consist of a variety of formats: images, text, sounds, dates, coded symbols, and locations. Data about society may have a hierarchical structure that has to be heeded for analysis.

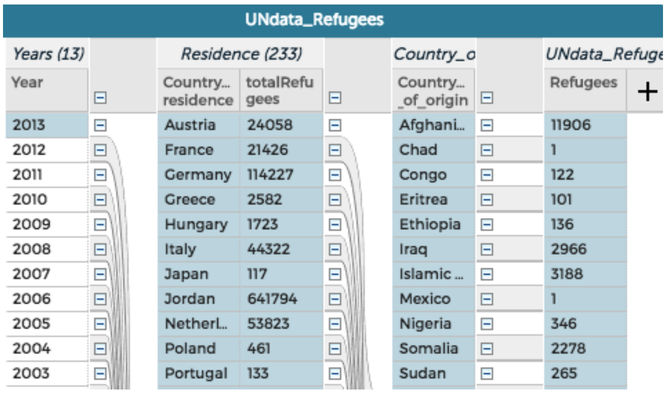
Figure 1 gives a screenshot of the top rows of a data table obtained from the United Nations data base on worldwide refugee movements (see United Nations Statistics Division, n.d.). The first row refers to the number of refugees ($n = 1$) who left Iraq to go to Afghanistan in 2013, the second row lists Iranians coming to Afghanistan, etc. Depending on the question of interest, the data have to be rearranged by country of origin or by country of residence before being graphed. When trends over time are of interest, the same data have to be restructured again in a timely order. In other situations data may have to be transformed or aggregated in order to be useful for an illustrative representation or the desired analysis. Cleaning, transforming, and structuring data are necessary skills, but these skills are not taught in the traditional classroom with its focus on inference-based statistics problems with tidy data.



Country or territory of asylum or residence	Country or territory of origin	Year	Refugees*	Refugees assisted by UNHCR	Total refugees and people in refugee-like situations**	Total refugees and people in refugee-like situations assisted by UNHCR
Afghanistan	Iraq	2013	1	1	1	1
Afghanistan	Islamic Rep. of Iran	2013	36	36	36	36
Afghanistan	Pakistan	2013	34	34	16,825	16,825
Afghanistan	State of Palestine	2013	1	1	1	1
Albania	Algeria	2013	0	0	0	0
Albania	China	2013	12	12	12	12
Albania	Dem. Rep. of the Congo	2013	5	5	5	5
Albania	Egypt	2013	3	3	3	3
Albania	Iraq	2013	5	5	5	5
Albania	Montenegro	2013	2	2	2	2

Figure 1. First rows of a table from the UN data base on worldwide refugee movement

CODAP (Finzer, 2016) is an educational web-based data analysis platform that supports the required transformation and restructuring of the data. Figure 2 shows a display of the first rows of the refugee data, ordered by year (upper level) aggregated at the level of country of residence. Now, the highlighted data of 2013 can be graphed displaying the number of refugees for each residence country in 2013.



UNdata_Refugees

Years (13)	Residence (233)		Country_o	UNdata_Refuge
Year	Country... residence	totalRefu gees	Country... _of_origin	Refugees
2013	Austria	24058	Afghani...	11906
2012	France	21426	Chad	1
2011	Germany	114227	Congo	122
2010	Greece	2582	Eritrea	101
2009	Hungary	1723	Ethiopia	136
2008	Italy	44322	Iraq	2966
2007	Japan	117	Islamic ...	3188
2006	Jordan	641794	Mexico	1
2005	Netherl...	53823	Nigeria	346
2004	Poland	461	Somalia	2278
2003	Portugal	133	Sudan	265

Figure 2. First rows of refugee data restructured with CODAP to explore the yearly distribution of refugees across various host countries

Handling and making sense of data, as available from the web, requires some data science skills. While various stakeholders hold different views on the question of whether data science constitutes an academic field on its own separate from statistics or rather is a subdomain of statistics (Donoho, 2015), a commonly accepted notion by Cleveland (2001) sees the field as the science of learning from data. As a highly interdisciplinary field data science refers to a set of skills and techniques that include statistics, data mining, computer science (coding, visualizing, computing with data, etc.), domain expertise, and communication (see Figure 3).

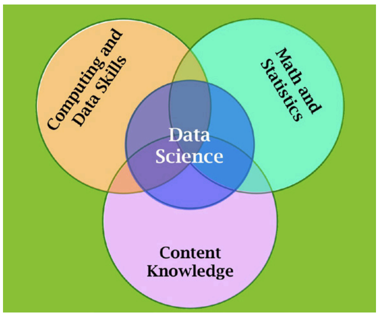


Figure 3. Data science as interdisciplinary field

If we expect our students to become active citizens who can understand the statistics published in the public sphere (by the media, official statistics producers, etc.) and fact-check on their own, they need a sound knowledge base that includes context knowledge, basic mathematics, familiarity with (at least) elementary statistics and familiarity with appropriate graphical and numerical tools for data representation, along with a capacity for critical thinking and a disposition to engage with evidence.

Along these lines, several recent publications (e.g., Horton, Baumer, & Wickham, 2014; Hardin et al., 2015) have argued for the need to bring data science concepts into the introductory college statistics curriculum. Beyond cognitive, technical and computational skills, successful data science requires particular habits of mind, i.e., ways of thinking, questioning and problem-solving. Finzer (2013) highlights important data habits of mind as a collection of attitudes and reflexive approaches to understanding the world through data:

- Acknowledge the need for data to gain insight.
- Look for the data: Ask “Which data could be helpful to reach conclusions, get insight or construct arguments?”
- Graph the data: Construct graphical representations that highlight potentially useful patterns in the data; patterns that are difficult to discern by staring at a table of numbers.
- Become immersed in the data: Use (and invent) measures. Look for and tell the story behind the data.

In dealing with data about society, an important additional habit of mind is:

- Develop a critical stance towards the quality and provenance of the data. Ask questions like: “How were the data collected? How are the variables defined and the constructs operationalized? Why, for what purpose, and in whose interest was the data collected in the first place?”

CONCLUSION

With the rise of a political culture in which public debate is framed by appeals to emotion disconnected from the details of policy, it is ever more important for citizens to be critical consumers of media reports, being aware of misuse of statistics and knowing effective ways to overcome them. Statistics students need to develop the capacity to make sense of the staggering amount of information collected in our increasingly data-centered world.

The call for redesigning our statistics curriculum with a focus on multivariate thinking is not new (e.g., Cobb, 2015; GAISE, 2016; Gould et al., 2016; Horton & Hardin, 2015; Ridgway, 2015). A special emphasis on promoting understanding of statistics about society has been the theme of the 2016 IASE Roundtable conference (Engel, 2016) whose proceedings give detailed account on the topic. In addition, the ProCivicStat project, a European program, involving 6 universities from 5 countries (for details, see ProCivicStat, n.d.) creates resources (theoretical framework, datasets, visualization tools, and concrete teaching materials) that engage students with multivariate data on a range of interesting topics relevant to their future roles as active citizens.

ACKNOWLEDGEMENTS

I am grateful to William Finzer for inspiring comments and discussions that helped to shape this paper.

REFERENCES

Cleveland, W. S. (2001). Data Science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*(1), 21–26.

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.

Donoho, D. (2015). *50 Years of data science*. Presentation at the Tukey Centennial workshop, Princeton NJ, Sept 18, 2015.
[Online: courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf]

Engel, J. (Ed.) (2016). *Promoting understanding of statistics about society. Proceedings of the IASE Roundtable*. [Online: www.iase-web.org/Conference_Proceedings.php]

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). [Online: escholarship.org/uc/uclastat_cts_tise]

Finzer, W. (2016). *Common online data analysis platform (CODAP)*. Emeryville, CA: Concord Consortium. [Online: concord.org/codap]

GAISE (2016). Guidelines for assessment and instruction in statistical education. *College report 2016*. Alexandria, VA: American Statistical Association. [Online: www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf]

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016). Teaching data science to secondary students – the mobilize introduction to data science curriculum. In: J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the IASE Roundtable Conference*. [Online: iase-web.org/Conference_Proceedings.php]

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward M. D. (2015). Data science in statistics curricula: preparing students to "think with data". *The American Statistician, 69*(4), 343–353.

Horton, N. J., Baumer, B., & Wickham, H. (2014). Teaching precursors to data science in introductory and second courses in statistics. In K. Makar, B. deSousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of ICOTS 9*. Voorburg, The Netherlands: International Statistical Institute.

Horton, N. J. & Hardin, J. S. (2015). Teaching the next generation of statistics students to "think with data". *The American Statistician, 69*(4), 259–265.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: The National Academies Press.

Nowotny, H. (2016). A scientific outlook for a post-factual world. *Project syndicate. The world's opinion page.* [Online: www.project-syndicate.org/commentary/science-uncertainty-post-factual-world-by-helga-nowotny-2016-11]

Open Government Partnership (n.d.). *Newsletter*. 2016 Global Summit Edition. [Online: www.opengovpartnership.org]

Peng, R. (2008). A method for visualizing multivariate time series data, *Journal of Statistical Software, 25*. [Online: www.jstatsoft.org/article/view/v025c01/v25c01.pdf]

ProCivicStat (n.d.). *Promoting civic engagement via explorations of evidence*. [Online: www.procivicstat.org]

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*. [Online: onlinelibrary.wiley.com/doi/10.1111/insr.12110/full]

Ridgway, J., Arnold, P., Moy, W. & Ridgway, R. (2016). Deriving heuristics from political speeches for understanding statistics about society. In J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the IASE Roundtable Conference*. [Online: iase-web.org/Conference_Proceedings.php]

Tampa Bay Times (n.d.). *PolitiFact*. [Online: www.politifact.com]

United Nations Statistics Division (n.d.). Table with data on Refugees. *United Nations data base*. [Online: data.un.org/Data.aspx?d=UNHCR&f=indID%3AType-Ref]

JOACHIM ENGEL
Institut für Mathematik und Informatik
Ludwigsburg University of Education
Reuteallee 46, 71634 Ludwigsburg
Germany