STATISTICAL LITERACY IN DATA REVOLUTION ERA: BUILDING BLOCKS AND INSTRUCTIONAL DILEMMAS

THEODOSIA PRODROMOU
University of New England
theodosia.prodromou@une.edu.au

TIM DUNNE [†]
University of Cape Town

ABSTRACT

The data revolution has given citizens access to enormous large-scale open databases. In order to take into account the full complexity of data, we have to change the way we think in terms of the nature of data and its availability, the ways in which it is displayed and used, and the skills that are required for its interpretation. Substantial changes in the content and processes involved in statistics education are needed. This paper calls for the introduction of new pedagogical constructs and principles needed in the age of the data revolution. The paper deals with a new construct of statistical literacy. We describe principles and dispositions that will become the building blocks of our pedagogical model. Our model suggests that effective engagement with large-scale data, modelling and interpretation situations requires the presence of knowledge-bases as well as supporting dispositions.

Keywords: *Statistics education research; Visualisation; Dispositions*

OPEN DATA AND STATISTICAL LITERACY

Nicholson, Ridgway, and McCusker (2013) argue that our view of what constitutes statistical literacy needs to change rapidly. Their remarks were a source of motivation to formulate a literacy framework for the data era. It is essential "not [to] strip important information away but [to be] presenting the full complexity of the data for context and highlighting what is important" (Smith, 2013). In order to take into account the full complexity of data, we have to change the way we think about controlling and handling data. This view calls for the introduction of new constructs and principles needed for the age of open data, which we will build upon the values, knowledge elements, and dispositional elements used to describe the constructs of statistical literacy. The principles we propose are listed in Table 1; we discuss the dispositional elements in it in turn.

CONTEXT, CORRELATION AND CAUSALITY

Knowledge and dispositional elements Initially students tend to approach a data set as complete in itself. The practice of explicitly noting or articulating a context permits channelling of exploration and deeper critical attention. This practice has to be acquired, and its inherent value appreciated. Experience of having teachers or peers elucidate context can lead to appreciate preconditions latent in the data structure and composition.

Table 1. Literacy framework for the open data era

Knowledge elements	*Dispositional elements*
1) Big ideas: open data, big data, multivariate data sets, data visualisation, correlation, and causality. 2) Language. 3) Context. 4) Critical view of data.	• An ability to consider context, correlation, causality. • An ability to assess data: – Evaluating the quality of evidence; – Conceiving statistics as modelling; – Action-oriented statistics and data visualisation.

Interplay between context, association and causation The act of articulation grounds description of relationships and associations between variables in the richness of the context. A tendency in the search for meaning in data is assertion of cause and effect relationships from what are merely associations. In many contexts this strategy can be demonstrated to be incomplete, unwise, or false. Open-data settings with an inherent time ordering of multiple observations may lead to a focus on the derived variables of change in category, in count or in measure over specified sequence lengths or time intervals of interest. Contextual cues permit an appreciation of the complexity of establishing cause-and-effect relationships. The strongest levels of evidence involve association of at least one change in an explanatory variable with a change in a response variable at a suitable lag that permits the development of effects over time, and their eventual observation.

Correlations Associations between changes over time in two variables may have a regularity of pattern within and across cases. These conditions will more strongly invite causal inferences, though they are necessary but not sufficient for that purpose. The slippery notions for the student involve the data context with its latent set of assumptions, agency or imposed action, control of extraneous variables and the role of time. One has to verify from the metadata that a suitable time-lag was allowed for in the observations. The lag has to be adequate to allow an effect to develop, but not be so large that the intrusion of other factors might become plausible explanations.

Context Proper understanding of open-data messages by students depends on their ability to distinguish the different types of data, reduce multitudinous amounts of data to the meaningful parts, and use this information to make informed decisions. Also needed is knowledge about the contexts that generate different types of data, and consideration of their broad range of sources (e.g., web data, science data, graph data, user data, and transition data). Students must know that to understand all data, even open data, they must go through a process of data integration and cleaning, reduction, indexing, analysis, and mining. Context knowledge is the major determinant of people’s familiarity with reducing large amounts of data down to the meaningful parts to investigate associations between such meaningful parts. Open-data correlations could point towards promising areas to further explore causal relationships. If the student is not aware of the context knowledge within which the data is generated, it is very difficult to identify appropriate parts and seek relationships amongst those parts.

The style of natural or naïve sense-making seeks evidence for dependable (replicable) interventions within a context, to achieve specific desired effects. It is the purposefulness of sense-making that has to be both invoked and harnessed in open-data contexts.

Causal effects One of the prospects of data visualisation for open data is valid real-world inferences, especially when evidence for causal effects for one or other set of changes is desirable or invited. The student has to be introduced at least to some appreciation of the stringency required for formal causal inferences.

If we are to encourage explorations that tap into cause and effect, we are obliged to introduce students to the conditions or assumptions necessary to warrant any causal inference. Usually these background conditions for validity of particular causal inferences will have to be assumed to hold. Thus it is important that students exposed to open-data visualisations for the elicitation of possible causal effects have sufficient awareness and reverence for the background assumptions that are required, before available graphical evidence is deemed adequate for the purpose. The argument here is the all-other-things-being-equal (AOTBE) assumption, naturally made at first pass in many common-sense approaches; this has to be confirmed in data analysis with the contrasting assumption almost-all-other-things-being equally unknown (AAOTBEU). This second step re-alerts the evidential enquiry to the latent contextual assumptions driving causal inferences. Essentially the confirmation exercise provides stimulus for the elimination of redundant variables and for increasing specification of variables that may generate causal effects.

Of the two acronyms, AAOTBEU is more frequently the correct description of the context for which we have data. Thus any inference of causality must be carefully phrased to ensure its validity.

ASSESSING DATA

Evaluating the quality of evidence This element involves: a) expecting, seeking and obtaining information about data sources, and inserting specifications to access the metadata into displays that present multivariate data, and b) discussions that draw attention to the source of data generation and associated quality.

Conceiving statistics as modelling Statistical modelling features the use of standard models to ‘fit’ the data. Applications of standard models need to be exercised with particular care (Box and Draper, 1987). Students have traditionally used standard statistical models and their use of these models focuses mainly on describing phenomena to make connections between data and chance (Bakker, Kent, Derry, Noss, & Hoyles, 2008). The revolution of data provokes the need to reflect on the nature and purpose of modelling, and the use of different models of phenomena. The data deluge gives rise to new situations that necessitate the invention of new numerical methods. Engagement with a new set of problems has created the need for custom-designed methods, the use of visual methods as analytic tools, and the use of grounded theory to explore open data (Ridgway, 2015). Open data is likely to increase students’ opportunities to deal with social science, since open data offers a “royal road into social science” (p. 5).

Action-oriented statistics and data visualisation In the case of open data, new methods need to be developed that are founded on additional basic skills, which include an understanding of techniques suited to analysing high-volume data for a variety of purposes. Key data providers, such as the OECD (Organisation for Economic Co-operation and Development) or the ONS (UK Office for National Statistics) provide powerful visualisations with the aim to make their data more accessible to the public. The purposive nature of data investigation facilitated by powerful visualisations such as data displays (e.g., dynamic population pyramids, dynamic maps of commuter flow, and choropleth maps) and consequent actions will engage and enrich student learning. These circumstances permit approaches to deeper issues: “Statistics is seen in the context of an investigative cycle, where the end point is a theoretical account and some action designed to change the current situation” (Ridgway, Nicholson, & McCusker, 2013, p. 9). Data visualisation is the key element of pedagogy centred on helping students develop data analysis skills and engage with the investigative cycle in the process of exploring open data. In the case of open data, data analysis is seen in the context of a “cycle of visual analysis”, that, according to Morton, Bunker, Mackinlay, Morton, & Stolte, 2012),

> starts with some task or question about which a knowledge worker seeks to gain understanding. In the first stage, the user forages for data that may contain relevant information for their analysis task. Next, they search for a visual structure that is appropriate for the data and instantiate that structure. At this point, the user interacts with the resulting visualisation (e.g., drill down to details or roll up to summarise) to develop further insight. Once the necessary insight is obtained, the user can then make an informed decision and take action (p. 807).

Cycle of visual approach This cycle of visual analysis is centred around and driven by students’ use of visualisation tools, effectively, as a habit of mind. The cycle requires that the visualisation system be flexible enough to support students’ feedback and allow students to investigate a variety of exploratory tasks in order to develop stronger understandings of the possible relationships between multiple variables, and reason about covariation between multiple variables using the power of digital tools to represent quantities and measures in new ways. This approach was further developed and described as a *cycle of inquiry and visual analysis* (Prodromou, 2014) for promoting students’ inferences from data visualisations. This cycle is more than a pedagogical strategy; it is intended to become a personal disposition because of its empowering effects in dealing with the dynamics of exploratory tasks when using visualisation tools. All of these processes necessarily unfold in specific contexts as the settings for inference. We need to understand how students conceptualise and experience the *cycle of inquiry and visual analysis* in order to make the cycle accessible, natural and habitual for every student.

An example for the cyclic approach Modes that lead students into an appreciation of multivariate structure will be necessary and useful. An example of an innovative approach for multivariate categorical data is offered in Wild (2013). He represented all combinations of (eight) binary responses in a multiple response item by a string of 0s and 1s; the visualisation in Figure 1 of the tabular presentation of the frequencies for the binary string patterns is, organised from more frequent to least frequent. This visualisation updates in two-dimensional space both the observations and the frequencies of correlated binary variables in a higher-dimensional space.

There is a need to see visualisation as product and process. As a process it has passive and active connotations. While it is important for students to develop a smooth reading of graphical images, this objective is insufficient. They also need to actively attempt to pre-visualise in order to engage and internalise deeper modes of interaction with visual and graphical forms. Students have to rely on their common sense in addressing complexity. Common sense seeks threads of insights as first steps into comprehending aspects of complexity. Students start to locate elements that matter in order to begin with tentative partitions. The teacher needs to emphasize to them the iterative process of seeking relevant partitions rather than any particular product fit for purpose. Initially the product is unknown, and that insight into its initial state may be the first element of their learning. What is to be communicated to the student is not just the device of partitioning as a building block of process, but also the value of the final partitioning once identified and explicitly labelled. Partitioning leads to a focus on a part or segment of the data. When this segment is rendered relatively homogeneous with respect to some features of the complete data, its internal complexity is reduced. Thus the selected data segment's own particular internal patterns are more likely to emerge in any data summarising activity.

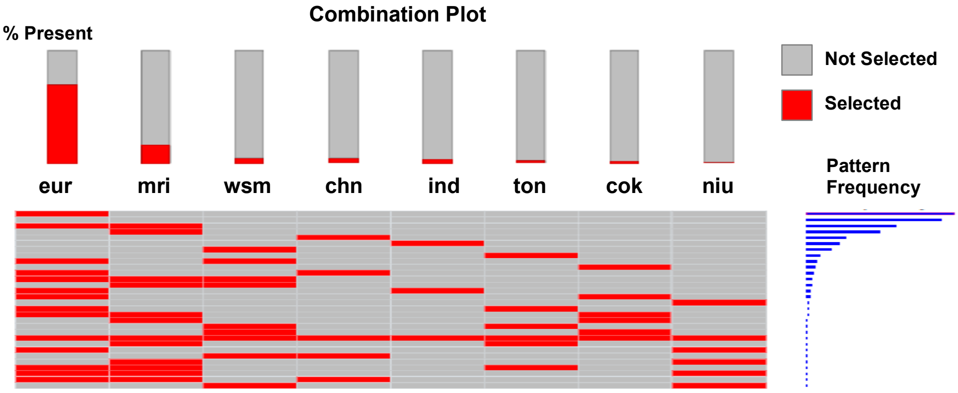

Figure 1. Possible combinations of responses in a multiple response item

Those emerging conditional patterns are the stuff of inference, whether in visual or statistical forms, and the inferences across these two forms will generally admit parallel conclusions. This fact keeps open the prospect of a formation in inference that is initially wholly visual being later complemented with a formalised statistical approach.

CONCLUSION

We have looked at skills that will be central to statistical literacy in the era of big data and have argued for an initial framework for teaching statistical literacy for an open data era. Our suggested framework incorporates new elements and dispositions to address the open data era. In particular, technological innovations in computer architecture not only allow the storage of large volumes of data but also enable users to obtain higher-quality graphical representations that have contributed to giving a prominent role to data visualisation and more generally to data. The increase of the volume of data precipitated the need for exploratory analyses, coupled with graphical methods that very quickly demonstrate ways of productively engaging with any large data set. The teachers of statistical literacy have to marshal many facets of visualisation as a result of significant developments in information technology. All of these facets seek elicitation of patterns and salient pictorial representations of a particular specified context. The production of contextual meaning and interpretation involves familiar cognitive strategies. These strategies encompass descriptions, profiles, partitions, contrasts, comparisons and associations that may focus on values for variables, contrasts of values, and changes in values. These features are discovered from and embedded in visualisations.

In fact, a key element in the success of pedagogical attempts to help students develop data analysis skills is the strong contribution of visualisation that exploits the human capability to perceive three-dimensional space and time and invoke those elements as windows into the data. Nevertheless, in our view, a solely visual approach has to be complemented with explicit language that controls unsubstantiated attribution of cause-and-effect. On this basis we may expect that a big change is required in teaching practice, academic programs, and curricula. In consequence we have argued for an initial framework for teaching statistical literacy for an open data era. This framework seeks to orchestrate context, language, plausibility of inferences within data, attention to causality and assessing data through a critical view. We have suggested that dispositional elements require explicit attention if statistical literacy is intended to enrich participation in the evolving world of the citizen. While the volume of data continues to grow and new technologies have become indispensable tools, there may also be a need to teach students some related computer science topics such as combinatorial optimisation, data structures, database management, etc. Should this point of view be taken into consideration, a substantial change would be required in teaching practice, academic programs, and school curricula. Curricula may need to include current computer-oriented data analysis methodologies, many of which have been developed outside the field of statistics.

ACKNOWLEDGEMENT

This paper was written with Professor Tim Dunne before he passed away on 18 April 2016.

REFERENCES

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal, 7*(2), 130–145. [Online: iase-web.org/Publications.php?p=SERJ_issues]

Box, G. E. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Morton, K., Bunker, R., Mackinlay, J., Morton, R., & Stolte, C. (2012). Dynamic workload driven data integration in Tableau. *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 807–816).

Nicholson, J., Ridgway, J., & McCusker, S. (2013). *Statistical literacy and multivariate thinking*. *Proceedings of the 59th World Statistics Congress*. The Hague, The Netherlands: ISI. [Online: 2013.isiproceedings.org/]

Prodromou, T. (2014). Drawing inference from data visualisations. *International Journal of Secondary Education, 2*(4), 66–72. [Online: www.sciencepublishinggroup.com/journal/archive?journalid=193&issueid=1930204]

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review, 84(3)*, 528–549. [Online: onlinelibrary.wiley.com/doi/10.1111/insr.12110/full]

Ridgway, J., Nicholson, J., & McCusker, S. (2013). 'Open data' and the semantic web require a rethink on statistics teaching. *Technology Innovations in Statistics Education, 7*(2). [Online: escholarship.org/uc/uclastat_cts_tise]

Smith, A. (2013). Emerging trends in data visualisation: Implications for producers of official statistics. *Proceedings of the 59th World Statistics Congress* (pp. 187–192). The Hague, The Netherlands: ISI. [Online: 2013.isiproceedings.org/]

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*(3), 223–265.

Wild, C. J. (2013). iNZight into time series and multiple-response data. In S. Forbes & B. Phillips (Eds), *Proceedings of the Joint IASE/IAOS Satellite Conference on Statistics Education for Progress*. The Hague, The Netherlands: ISI [Online: iase-web.org/Conference_Proceedings.php?p=Stats_Education_for_Progress_2013]

THEODOSIA PRODROMOU
University of New England,
Armidale, NSW 2351, Australia