STATISTICAL LITERACY AS THE EARTH MOVES

CHRIS J. WILD
University of Auckland
c.wild@auckland.ac.nz

ABSTRACT

"The Times They Are a-Changin'" says the old Bob Dylan song. But it is not just the times that are a-changin'. For statistical literacy, the very earth is moving under our feet (apologies to Carole King). The seismic forces are (i) new forms of communication and discourse and (ii) new forms of data, data display and human interaction with data. These upheavals in the worlds of communication and data are ongoing. If anything, the pace of change is accelerating. And with it, what it means to be statistically literate is also changing. So how can we tell what is important? We will air some enduring themes and guiding principles.

Keywords: *Statistics education research; Effects of technology; Guiding principles; Research agenda; Big data*

1. AS THE EARTH MOVES

We all know that the internet is continually spawning new modes of communication and dissemination, and disrupting the old ways and that all this is changing the way we live, receive ideas and information, and conduct business and governance. We know about the growing deluge of data, about big data, open data and the advent of data journalism (Gould, 2010; Ridgway, Nicholson, & McCusker, 2013; Ridgway, 2015). This is accompanied by worries – worries about personal privacy, about trampled ethics, about being overwhelmed, and about becoming disempowered spectators (even victims) of technocratic decision making (Philip, Schuler-Brown, & Way, 2013). And in statistics education there are, and should be, worries about the extent to which we are relevant any more. It is not just the times that are a-changin', the very earth is moving under our feet, with increasingly strong shock waves moving in surprising directions.

As forms of communication and ways of presenting and receiving data change so too does what it takes for someone to be statistically literate. There is an urgent need for a research agenda targeting, "What will it mean to be statistically literate as we start heading into the 2020s?" This matters a great deal because the core purpose of today's education is to prepare students for the world to come, and the core purpose of today's research is to inform tomorrow's education.

So how can we tell what is important? We will propose some guiding principles. The first is that "statistical literacy" should focus on the *human understanding* of data and what data can and cannot do. A second is that for something to be classified as important for statistical literacy, we need to think it is likely to be *empowering* and of *lasting value* in the face of technological advances and social changes. Third we need to be able to answer, "What *transferable lessons* can we extract here?" Unless the number of such lessons is small, no one will ever be able to remember and apply them. At the base of all

Statistics Education Research Journal, 16(1), 31-37, http://iase-web.org/Publications.php?p=SERJ

of this are existential questions about, "What can we come to know and how can we come to know it?" and "How then can we think?"

2. WHY STATISTICS SHOULD BE LIKE COOKING

Literacy, unadorned, is the ability to read and write – typically at some basic level. Literacy coupled with some subject name, e.g., computer literacy, often conjures up a minimal subset of basic skills in that subject desired, if not expected, for all citizens (Gal, 2004). While statistical literacy is sometimes used with connotations of basic skills, increasingly, following Wallman (1993) and Gal (2002), it has been used with reference to the ability to understand and critically evaluate data-based arguments and reports of statistical work because these have been the ways in which most people have been impacted by statistics.

A mainstay of the way we have thought about needs for statistical literacy is a dichotomy drawn between the producers and the consumers of statistics; literacy has focused on the needs of the latter. But the value of such a dichotomy is eroding fast. Increasingly people are no longer just passive recipients of data-based reports. Dashboards created for decision makers, and visualizations created by others such as data journalists, already let us vary settings and explore data to discover relationships and play what-if games. Easy-to-use data analysis tools are increasingly accessible in our plugged-in world (apps and web apps) and it will not be long before public agencies connect powerful analysis tools even to huge sets of unit-record data. "Self-service analytics/business intelligence" is an emerging field. As Roger Peng said on Simply Statistics (Peng, 2015), "In the future, everyone will need some data analysis skills." Large proportions of the population will not just be consumers, they will also be producers of statistics.

"Statistics should be like cooking" (*Xiao-Li Meng*).

When challenged in this regard about his picturesque wine-producers/wine-connoisseurs version of the producers/consumers dichotomy, Harvard's Xiao-Li Meng quickly came up with a new metaphor above, one that was much more aptly inclusive (personal conversation, June 2016). While not everyone is a professional chef, almost everybody is involved in food preparation to some degree at some stage of their lives. *Statistical* cooking both provides food for our minds and helps us digest it.

Meaning trumps mechanics.

The implications of the basic-skills versus consumer-needs ways of thinking about statistical literacy are converging. With the advances in ubiquitous technology, the basic skills of value to the broad populace are no longer mechanical (e.g., the ability to calculate an average), they are conceptual understandings (what an average tells you and how it can be used and misused). The same trends apply to data analysts. For most analysis, the mechanistic elements are handled by pre-packaged software leaving the analyst to concentrate on decision making and interpretation. Over time, any particular mechanical or procedural function can be expected to be taken over by software. So the biggest needs for every statistical cook, from the person who simply heats up TV dinners to the Michelin-star chef, are needs for conceptual understanding. We might encapsulate this as meaning trumps mechanics.

The greatest hope for extending the accessibility of the statistical messages in data to a much wider spectrum of the population is visualization. But no matter where you are on the continuum between the webpage viewer and the professional data analyst, you have to be able to make sense of what you are seeing. An essential difference for the people at different points along the continuum is how much choice they have in how they can look and how much they have to know to be able to make those choices.

The varieties of graphic/visualization experience are mushrooming (see Gelman & Unwin, 2013, its discussion and particularly the rejoinder). We teach a tiny, historical fraction of what is available. There is immense creativity in the producers of infographics and data visualizations, especially those produced by designers for whom visual creativity and novelty helps to attract attention. It is not enough to expose people to graphics – we have to find and expose the lessons that can be transferred to other data sets at other times – lessons about how to read particular types of graphics and what they are good for. The pursuit of novelty for its own sake is the enemy of transferable skills. Transferability is provided by general principles and scaffolds for thinking.

How do you decide what sorts of displays and summaries people should be able to read? Bearing in mind that what can be taught in realistic timeframes is severely limited, what instruction is likely to be most useful? For data exploration I believe prioritization should be based on maximizing what people can see and how fast they can get to see it in data types they are most likely to encounter. For using visualizations on webpages or in the traditional media, prioritization should be based on the data encodings people are most likely to be exposed to, together with ingraining a propensity to examine critically graphics of types previously unseen because people will be coming across forms of graphics they have not seen before constantly.

3. PUTTING DATA ON THE WITNESS STAND

3.1. EVEN THE GOOD CAN DO BAD THINGS

"Nobody wants 'data'", David Hand was widely quoted as saying in press reports from 2015. "What they want are the answers." (e.g., Harford, 2014 b). The answers arrived at from data always involve *extrapolation* from the patterns seen in the data we have on the entities we have observed to some other or wider context and this is where the wheels tend to fall off.

Good statistically-designed studies use data collection processes carefully chosen to minimize distortions in their data that could bias the answers they provide for their questions of interest. These types of data are the gold standard, the least likely to be misleading. The way that the data comes into being is designed and controlled to produce data which is of high quality for addressing pre-specified questions or types of questions. The design process itself justifies certain types of extrapolation from the data (e.g., sample to population or process, experiment to causation).

But even good, gold-standard data can tell us bad (misleading) things. The three major reasons that we worry about in which data can mislead us are: *bias, random error and confounding*. These, and what we try to do to overcome them, must therefore play a central role in statistical literacy. There are always small biases in even the best studies but we never know how big they are and we cannot do anything about them. We just hope that the care we have taken is enough to have kept them small. Random error is something we *can* make allowances for in the analysis of data from designed studies (e.g., significance tests, confidence intervals, etc.). Confounding is central for thinking about causal

reasoning and inherently involves multivariate contexts. So all of this happens even under ideal conditions.

3.2. WHEN OPPORTUNITY STRIKES

Opportunistic (happenstance or found) data is data that was not specifically collected to address our problems. It came into being for other reasons but looks as though it might be relevant to our problems. Most big data is of this form. In a very real sense, we have walked into the theatre half way through the movie and have then to pick up the story. For opportunistic data there is no extrapolation which is justified by a data-collection process specifically designed to facilitate that extrapolation. The best we can do is to try to forensically reconstruct what this data is and how it came to be (its *provenance*). What entities were measures taken on? What measures have been employed and how? By what processes did some things get to be recorded and others not? What distortions might this cause? It is all about trying to gauge the extent to which we can generalize from patterns in the data to the way we think it will be in populations or processes that we care about.

It is helpful to think of any data as being like a witness in a criminal trial. We need to put the witness on the stand and cross examine both it and its credentials. Our aim, to help us decide, "*What, of the evidence that this witness provides, can I trust?*" We have to tread carefully because:

Data can tell lies.

It is not so much that the witness (the data) is itself telling us lies, but that our witness can only speak to us through interpreters (analysts and communicators) who may either misread or misreport what the witness can truthfully say. We think that the variables are measuring something when they are really measuring something rather different. The make-up of the set of entities we have data on can be misread in ways that lead to false generalizations. "Human beings," said David Hand, "are great at finding narratives that aren't really there" (see Adler, 2015).

Big Data can tell bigger lies.

This time it would probably be more correct to say "more lies" rather than "bigger lies". It can tell more lies and we are more likely to believe them. Why? With big data, the "answers" come from huge amounts of data and "cutting edge technology" – making it hard for a still, small voice of skepticism to surface, "this is probably rubbish." In business, the company that acts on new knowledge first can make a lot of money, thus building in an incentive for quick decisions. Additionally, anywhere where there are important problems and we are desperate for answers there is an unreasoned willingness to believe data that on the face of it looks relevant, particularly if the answers it gives seem to conform to our preconceptions.

It is hard to accept that, on the information it is (currently) possible for us to get, it may simply be impossible to determine the answers. That sounds like such a defeatist attitude. Positivity keeps on urging, "It has to be possible." All of this can sweep us up into wishful thinking. And capping this off are the people who make a great deal of money by exploiting our wishful thing, the present-day peddlers of cargo cults, silver bullets and snake oil.

The big thing for small data is random error. The big thing for big data is bias.

Why can big data tell more lies than smaller data? With small data many possible stories are ignored or never read, being screened out as falling beneath the fog of random error (noise). With big data the random noise is much smaller, so there are many patterns that are obviously "signals". Although these effects are not random, they may still just be *artefacts* of the way the data was assembled (i.e., be lies) *rather than facts* about some aspect of the real world that we are interested in.

3.3. GOOD LUCK VERSUS GOOD MANAGEMENT

Any substantial learning from data involves extrapolating from what you can see in the data you have to how it might be in some wider universe. The only way to do this is to assume that, at some level, *what you have not seen is very much like what you have already seen.*

Random sampling guarantees unbiased estimation of characteristics of the process or population sampled – good results (within limits) produced by good management. This is what happens with data purposefully collected using well implemented, statistically-designed studies. As previously noted, for opportunistic data there is no extrapolation which is justified by the data-collection process itself. Biased data *can only be relied on* to provide unbiased analyses if we have other sources of information that tell us exactly how and to what extent the data has become biased in enough detail to enable us to correct for those biases. Otherwise extrapolation is either largely a shot in the dark or is justified by other things you know.

When rich data on human behavior and dispositions (e.g., how they feel about issues) is wanted, opportunistic data generated by social media is something that is relatively easy to access.

As goes the Twittersphere, so goes the world. (Not!)

The mere suggestion is ridiculous. We should all know to expect that self-selected groups, e.g., regular tweeters (or any other social media users), are likely to differ from the general population in many ways. For some things tweeters will tend to be different from the public at large. For others they may well be much the same (e.g., the way their body chemistry works). The features we are interested in may be those that tend to be different or may be those that are largely the same. In the latter case we get good results by good luck. And there have been some great advances made by generalizing from opportunistic sources of data such as automatic translation of text, and voice and image recognition. But, although the occasional free lunch may fall into our laps we should not rely on a regular supply.

4. UNBLINDING THE BLINDINGLY OBVIOUS

Bad data leads to bad decisions.

"*Garbage in, Garbage out*" (GIGO) has been a standard catch-phrase in information-technology since the 1960s. It definitely applies to statistical data analysis. Relying on bad information can lead us down a garden path that leads through false conclusions to bad decisions. That is blindingly obvious! So we can translate GIGO to "bad data leads to bad decisions".

There are a number of things about extracting messages from data the truth of which is blindingly obvious – obvious that is, until the blackout-curtains of technological magic, wishful thinking and sales hype have been drawn over them so that the clear light of common sense is blocked from shining into our mental chambers. We have seen examples above and will now add more.

Forecasting projects patterns from the past into the future.

All forecasting is based on projecting patterns from the past into the future. If the mechanisms that generate the data-patterns change in important ways (e.g., major structural changes to the economic environment), then the future patterns will be different from the past patterns and the forecasts will start to fail.

Purely predictive patterns only work until they fail.

If you only have patterns in data (associations, correlations) and do not understand the causal mechanisms (what is producing those patterns), then there is no way that you can spot when they change, so you cannot know when your forecasts will start to fail. This is obvious when you think about it but easy to overlook. A great example is the Google Flu Trends story (told well in Harford, 2014 a) in which flu epidemics were predicted early and well by the extent of relevant internet-searches – until the time came when they were not. In general, prediction assumes that patterns in your data still apply in the situation where you want to make the prediction. Finally, we conclude with this:

Computer algorithms feed on information (data) and re-present it.
They cannot produce information that was never there in the first place.

The knowledge bases that need to be activated for someone to be statistically literate are changing fast as a result of ongoing technologically-enabled revolutions in how we communicate, and how we can obtain and reap benefits from data. Consequently, the statistics education community must investigate emerging needs and reconceive statistical literacy in ways that ensure it can still deliver lasting benefits in a rapidly changing world. This involves prioritizing: conceptual understandings that enhance *human understanding* of data and what data can and cannot do, elements that are *empowering* and of *lasting value* to citizens (see also Section 6 of Philip, Schuler-Brown, & Way, 2013), and result in small numbers of widely *transferable lessons*. In Sections 3 and 4 we have discussed a set of fundamental messages about learning from data (whether big or small) that are entirely obvious when not obscured by complexity, technological glitz, wishful thinking and sales hype. Dragging those simple messages out into the open and exposing them in memorable ways may be one useful, if small, contribution. For statistical literacy, we can, at the very least, endeavor to prevent the blinds being drawn over the obvious.

ACKNOWLEDGEMENTS

"The Times They Are a-Changin' " is the title track of Bob Dylan's 1964 album. "I Feel the Earth Move" is the opening track of Carole King's, 1971 album Tapestry.

REFERENCES

Adler, T. (2015). How predictive analytics can reveal more than you'd like. *Business Reporter*. [Online: business-reporter.co.uk/2015/03/23/how-predictive-analytics-can-reveal-more-than-youd-like/]

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

Gal, I. (2004). Statistical Literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47–78). Dordrecht, The Netherlands: Kluwer.

Gelman, A., & Unwin, A. (2013). Infovis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics*, *22*(1), 2–28.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, *78*(2), 297–315.

Harford, T. (2014 a). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Harford, T. (2014 b). Big data: are we making a big mistake? *Financial Times*. [Online: www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0]

Peng, R. (2015). The massive future of statistics education. *Simply Statistics. A statistics blog*. [Online: simplystatistics.org/2015/07/03/the-massive-future-of-statistics-education/]

Philip, T. M., Schuler-Brown, S. & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning*, *18*(3), 103–120.

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]

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]

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), l–8.

CHRIS J. WILD
Department of Statistics
The University of Auckland
Private Bag 92019, Auckland 1142
New Zealand