Data Visualisation and Statistics from the Future

Theodosia Prodromou [1]
[1] University of New England, Armidale, NSW, AUSTRALIA
theodosia.prodromou@une.edu.au

Abstract

Our world is increasingly data-rich and data-dependent. Every day, 2.5 quintillion bytes of new data are created — so much that 90% of the data in the world today has been created in the last two years alone. Data visualization allows us to explore and effectively communicate relevant information about this voluminous data through graphic representations. This includes visualisation of all kinds of information, not just data, and is closely associated with research by computer scientists. From the perspective of statistical pedagogy, data visualisation can be viewed as computer-assisted exploratory data analysis of voluminous complex data sets.

Data visualisation has blossomed into a multidisciplinary research area, and a wide range of visualisation tools have been developed at an accelerated pace. Admittedly, statistical data analysis necessitates data visualisation to form the basis for decision-making. There is a greater need for people to make good inferences from visualisations. The flexible nature of current computing tools can potentially have a major impact on the discipline of statistics and allow easier use of visualisations in the educational process. There is evidence that a wholly visual approach can help students learn statistics, especially the informal inferences, and the abilities used by students to make informal inferences based on a wholly visual approach are the same abilities needed to make inferences from and judgements about the statistical graphs that are so common. This paper proposes a framework for considering the way that youth might employ data visualisation when working with data in a pedagogical context.

Keywords: data visualisation, statistics, education, inference, students

1. Introduction

Developments in computing power have been of great benefit to graphical representations in recent years. Computing advances have enabled the drawing of precise, complex displays with great ease, so graphics drawn for the purpose of illustrating and explaining results become available. Moreover, computing advances have also benefitted exploratory graphical representations that in turn provided support to exploring data further. The quality and quantity of graphical representations has been improved. Many different displays of the same data can be drawn to shed some light into the information in data.

The affordances of different displays of data graphical representations have been gradually appreciated and capitalised on. These advances provide an opportunity for new avenues of education for young students of statistics.

The great importance of computer software availability and popularity in determining what analyses are carried out and how they are presented is a topic of major importance. In the world of business, itself, the spreadsheet Excel has long been in common use for creating graphical representation of data. In the world of statistics, there are several sophisticated software packages used by statisticians—SAS, SPSS, S and S-PLUS, and more recently R—however, "None of these packages provide effective interactive tools for exploratory graphics, though they are all moving slowly in that direction as well as extending the range and flexibility of the presentation graphics they offer" (Unwin et al., 2008, p. 6).

As already noted, a great mass of data is being continually generated. Citizens of present and future eras will be constantly and increasingly bombarded with tables of data and graphics in media, economic forecasting, and social activities. People's decisions about their everyday lives depend on data visualisation and numerous reported figures, and these are not always what they seem: "There may be less in there

than meets the eyes" (Huff, 1954, p. 8). Although Huff wrote this several decades ago about graphics used in the print media of his day, his ideas are still applicable.

It is essential to have citizens who have the skills and understandings required to become informed about our world and understand the data that politicians, advertisers, and other advocates are using to promote particular causes that will impact the future of our planet. This embraces the capacity of not only understanding the underlying messages that the data attempt to reveal but also critically examining the statements presented in news media in terms of data and data representations, so as to recognize misleading or distorted graphs that misrepresent data, and are too often used. These graphs may be created intentionally to hinder the proper interpretation of data and subsequently incorrect conclusions may be derived from it. Misrepresentations of data may be also created accidentally by users who are unfamiliar with using the graphical software, or because the data cannot be accurately conveyed. Whether the deception is intentional or not, being able to recognize such distortions is a skill to be valued in educated people.

As Huff pointed out:

> The secret language of statistics, so appealing in a fact-minded culture, is employed to sensationalize, inflate, confuse, and oversimplify. Statistical methods and statistical terms are necessary in reporting the mass data of social and economic trends, business conditions, "opinion" polls, the census. But without writers who use the words with honesty and understanding and readers who know what they mean, the result can only be semantic nonsense. (Huff, 1954, p. 8)

These concerns are made no less pressing during the current data revolution. The field of data visualisation describes ways to present information that avoid creating misleading graphs, but the problem of filtering out any massages from data still exist. This suggests that visualisation tools could be very important in statistics education.

2. Exploratory Data Analysis

In 1960s and '70s, John Tukey issued a call for the recognition of data analysis as an authentic branch of statistics distinct from mathematical statistics. He introduced Exploratory Data Analysis (EDA) and invented a wide variety of new, simple, and effective graphic displays such as stem-leaf plots, boxplots, hanging rootograms, two-way table displays, and so forth. Tukey's Exploratory data analysis, aided by the use of graphs, appreciates the central role of data in statistical analysis. EDA looks for patterns in the data and allows the data to reveal underlying structure and suggest admissible models that best fit. As Tukey (1977) argued, graphs let us see what may be happening over and above what we have already described.

Tukey suggested that statistics ought to be concerned with data analysis, thus the EDA approach attributes special attention to dynamic visualisation of concepts as a more general and useful idea to display characteristic features of a concept and not only to depict data in an easily accessible way in order to facilitate interpretation.

EDA is discussed as having the following objectives: "(1) maximize insight into a data set, (2) uncover underlying structure, (3) extract important variables, (4) detect outliers and anomalies, (5) test underlying assumptions, (6) develop parsimonious models, and (7) determine optimal factor settings" (NIST, 2003, p. 2).

EDA encourages users to look for the underlying structure of the data and make visual inferences from data, emphasising (large central modes or distribution of data), while deemphasising those minor ones in the tails of the distribution, such as outliers and anomalies. Then EDA places emphasis on looking at data sets in order to form hypotheses worth testing, instead of putting forward hypotheses concerning the possible kind of probability model the data follow.

With the advent of EDA, a new and innovative way of interactive data analysis was born that was not inextricably bound to mathematical probability. Tukey (1972, p. 51) claimed that Exploratory Data Analysis (EDA) enables people to avoid the notion of

probability and use it only where it is needed. Exploratory data analysis "[helps] people to overcome the problems they had with the discovery of the collective distribution, which is, in fact, psychologically the cornerstone for probability, it places emphasis upon the analysis of data" (Prodromou, 2008, p. 274). In fact, new ideas introduced by Tukey prompted many statisticians to give a more prominent role to data visualisation and more generally to data.

New data visualisation tools create the possibility of tools for visualising collections of actual data usable by students at the middle and high school level (Finzer, 2001; Konold & Miller, 2005), providing opportunities to include key statistical concepts not previously taught to this age, such as inferential reasoning (Ben-Zvi, 2006) and the process of statistical investigations (Wild & Pfannkuch, 1999). Concepts important for statistics can be mastered with a "wholly visual approach" (Wild et al., 2011, p. 252) that will, "attempt to minimize the conceptual distances between the concrete realities, including precursor practical experiences, and the dynamic imagery envisaged." This novel approach encourages an inferential step to be performed without having "students taking their eyes off their graphs" (p. 252) so that students can easily draw connections between question and data and answer as quickly as possible. This new approach places working with data and making visual inferences at the core of instruction, a shift towards more holistic, process/problem-oriented approaches to learning statistics.

Admittedly, the shift in research from intensively focusing on mathematical tools of statistics (e.g., averages, distribution, graphs, samples, modal clumps) as embedded in the processes and contexts under investigations to statistical processes, has raised new issues. Hancock, Kaput, and Goldsmith (1992) highlighted the challenges that students encountered when attempting to connect their statistical questions to the data required as evidence; and then aimed at associating their conclusions with the questions under investigation. They argued that school statistics curricula ignored two dimensions of a statistical process: (1) connecting statistical questions to data and (2) linking conclusions from data to the questions under investigation. Focusing on the statistical *investigative cycle* (Wild & Pfannkuch, 1999) as a process of investing the context domain of a real problem and seeking explanations in that problem context, "entails understanding the statistical investigation cycle as a process of making inferences. That is, it is not the data in front of us that is of great interest, but the more general characteristics and processes that created the data. This process is indeed inferential" (Makar & Rubin, 2009, p. 84).

This process of making inferences has attracted a great deal of interest in learners' informal ideas about statistical inference and people's intuitive ways of reasoning about statistical inference in diverse contexts. Researchers of statistics education over the last several years have focused on informal aspects of inferential reasoning. Makar and Rubin (2009) grappled with developing a framework for understanding the building blocks of informal statistical inference and inferential reasoning within a context of statistical investigations. Three main principles of informal inference to making informal inferences from data have been identified by Makar and Rubin (2009): (1) generalising, predicting, estimating parameters, and drawing conclusions that extend beyond describing the given data; (2) using data as evidence for those generalisations, and (3) employing probabilistic language when describing the generalisation and making informal reference that include levels of confidence or uncertainty in the conclusions drawn. A fourth principle involves (4) comparison of datasets with a model (Bakker et al., 2008). The first principle is particularly related to the process of inference, whereas the second and third are related to statistics. When generalising beyond the data one makes generalisations (inferences) from a sample about a specified population from which the sample is drawn. It is often difficult to distinguish descriptions of a data sample and inferences about the population from which that data was drawn (Pfannkuch, 2006), but visualisation techniques can help students when shifting their attention from describing the sample to making inferences

about the population (Wild, et al., 2011). The EDA approach that focuses explicitly on "looking at data to see what it seems to say" (Tukey, 1977, p. v), does not necessarily promote inferential reasoning if the attention focuses on the data from a sample. Other tools such as box plots can be used as a bridge "between reasoning entirely from graphics to reasoning from summaries in ways that converge, qualitatively" (Wild et al., 2011, 254).

3. Data visualisation and learning statistics

The application of visualisation methods to an ever-expanding array of substantive data structures and the new developments of dynamic graphic methods that allow instantaneous and direct manipulation of graphical objects and related statistical properties bring novel applications from new Technologies and new approaches towards teaching (e.g., modelling aspects of applications, observe valid data to adequately answer the questions) suggesting new ways to support students' learning.

Incorporating data into curriculum and learning statistics in schools has been of increased emphasis. Despite calls for more emphasis on collecting and then making sense of data, the focus in school statistics continues to be on calculations, procedures, and graphs (Sorto, 2006). However, in their national curriculum, several countries encourage students to make sense of data and data representations and unravel the story that data tell (implicitly or explicitly), and they promote visualisation of data as a way to do this (e.g., see National Council of Teachers of Mathematics [NCTM], 2000; Australian Curriculum, Assessment and Reporting Authority [ACARA], 2010). For example, the Australian Curriculum, Assessment and reporting Authority notes for students at the middle school level, interest in, "data representation and interpretation," especially with identifying and investigating "issues involving continuous or large count data collected from primary and secondary sources" (p. 42) and describing and interpreting data sets in terms of location (centre) and spread. It is uncertain, however the extent to which schools in these countries have successfully implemented data visualisation when students are engaged in interpreting data.

The introduction of digital technology into schools has prompted increased interest in Exploratory Data Analysis (EDA) as a means of engaging students in statistical analysis, arguably reducing the need for a sophisticated understanding of theoretical statistical principles, demanding an appreciation of probability theory, prior to meaningful engagement. The technology is ideally suited for supporting students as they manipulate data and portray it in a range of different representations in order to infer underlying trends. New data visualisation tools promote a perspective on visualising collections of actual data at middle and high school level (Finzer, 2001; Konold & Miller, 2004), providing opportunities to include key statistical concepts not previously taught to this age, such as inferential reasoning (Ben-Zvi, 2006) and the process of statistical investigations (Wild & Pfannkuch, 1999). Rubin, Hammerman and Konold (2006) argued that the use of technological tools such as TinkerPlots can rapidly enculturate their users into unblocking stories in data. Conceptions of statistics, and in particular statistical inference, can be mastered with a "wholly visual approach" (Wild et al., 2011, p. 252) that will, "attempt to minimize the conceptual distances between the concrete realities, including precursor practical experiences, and the dynamic imagery envisaged." This new approach portrays the foundational difference in newer approaches to working with data as the core of instruction, towards more holistic, process/problem-oriented approaches to learning statistic.

The emphasis should be on making inferences based on a wholly visual approach and justifying those inferences based on characteristics of the visual graphical representations. The same basic skills used by students to perform informal inferences based on a wholly visual approach are the same strategies needed to perform informal inferences and judgments about the statistical graphs that are so common.

It seems there might be some essential competencies required from youth's engagement in data visualisation and future statistics. These competencies, which

include the essential skills for graph comprehension (Friel et al. 2001) and elements in acquisition of statistical literacy (Gal, 2002), are: 1) data grappling skills: the ability to move data around and manipulate data with some programming language or languages; 2) the ability to create informative pictures of data when they when they encounter new data; 3) finding relationships between the data, formulate and communicate a reasoned opinion about statistics (average, variation around average), confidence intervals, error bars; 4) reading beyond the data, including identification of relationships, in order to perform inferences or generalisations based on a wholly visual approach; 5) forecasting and prediction, both general and specific; 6) communication of a reasoned opinion of the information stemming from data; and 7) finding a model that best fits a dataset. Some software representations help students in making inferences by exhibiting regions of the data that strongly suggest the presence of underlying models of the data.

The above competencies are presented in the form of seven essential skills that might be seen as both a hierarchy of competence, and a framework for considering the way that students and students might employ data visualisation when working with data in a pedagogical context.

4. Discussion

These simultaneous and independent stimuli for statistical analysis based on visualisation occurred as a result of significant developments in information technology. 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 visualisation that contributed to giving a prominent role to data visualisation and more generally to data. The increase of the volume of data necessitated the need for exploratory analyses that coupled together with the graphical methods very quickly demonstrate their potential in this kind of analysis. In fact, the key element in the success of data analysis is the strong contribution of visualisation because it exploits the human capability to perceive the three dimensional space.

Data analysis and modelling have become integral components of high school courses. New technologies provide students with dynamic, visually compelling environments that help them obtain detailed graphical visualization quickly to help explore, analyse, and model data. Software packages such as Fathom and TinkerPlots can help students to explore and learn within an *enquiry cycle* (Wild & Pfannkuch, 1999) using data representations that would help students to build concepts underpinning statistical thinking. Moreover, learning environments (e.g., simulations, animations) might support students with their dynamic features (e.g., animated histograms, animated box plots) to examine and scrutinise both samples and parent populations of data.

While the amount of data continues to grow and new technologies have become indispensable tools, there is also a need to teach youth the related computer science topics such as combinatorial optimization data, data structures, database management, etc. Should this point of view ever be taken into consideration, a big change would be required in teaching practice, academic programs, and school curricula. The curricula can also be expanded to include current computer-oriented data analysis methodologies, many of which have been developed outside the field of statistics.

Advances in computer software and hardware have greatly simplified and eased the production of graphical representations. These advances have also contributed to raising the expected standards. Future developments of graphic methods will indisputably contain more flexible and powerful software packages that better integrate modelling and graphical representations. There will probably be developed novel and innovative graphics and some of the general design of displays will be improved.

This attention increases the need to understand the psychological aspects of data visualisation that will provide us with feedback about how youth reason with the help of improved graphical displays, what aspects of formal inference are needed given

current visualisation tools, and which methods foster students' ability to understand conventional formal conceptions and characteristics of big data sets. Ideally, there should be further progress in the formal theory of data visualisation. Nevertheless, current growth of the field already gives rise to the challenge of integrating data visualisation in statistics education for youth.

References

Australian Curriculum, Assessment and Reporting Authority (2010). *Australian Curriculum: Mathematics*. Online: Version 1.2. http://www.acara.edu.au

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), 131-146. Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ7%282%29_Bakker.pdf

Ben-Zvi, D. (2006). Using Tinkerplots to scaffold students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.

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

Finzer, W. (2001). *Fathom: Dynamic Data* .Emeryville, CA: Key Curriculum Online: www.keypress.com/x5656.xml

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124-158.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337-364

Huff, D. (1954). *How to Lie with Statistics*. New York, NY: W.W.Norton.

Konold, C., & Miller, C. (2005). *TinkerPlots: Dynamic data exploration*. Emeryville, CA: Key Curriculum. Online: www.keypress.com/x5715.xml

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105. Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of teachers of Mathematics.

NIST (2003). *NIST/SEMATECH e-Handbook of Statistical Methods*. Online: http://www.itl.nist.gov/div898/handbook

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Research Conference on Teaching Statistics, Salvador, Brazil.* [CDROM]. Voorburg, The Netherlands: International Statistical Institute.

Prodromou, T. (2008). Connecting thinking about distribution. *Unpublished PhD Thesis*, University of Warwick, United Kingdom.

Rubin, A., Hammerman, J. K., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Research Conference on Teaching Statistics, Salvador, Brazil.* [CDROM]. Voorburg, The Netherlands: International Statistical Institute.

Sorto, M. A. (2006). Identifying content knowledge fir teaching statistics. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.

Tukey, J. W. (1972). Data analysis, computation and mathematics. *Quarterly for Applied Mathematics, 30*, 51-65.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison−Wesley Publishing Company.

Unwin, A., Chen, C., & Hardle, W. (2008).Introduction. In C. Chen, W. Hardle, & A. Unwin. *Handbook of Data Visualisation*. New York: Springer-Verlag.

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

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Conceptions of Statistical Inference. *In J. R. Statist. Soc. A*, 174, *Part 2*, 247-295.