STATISTICS EDUCATION ON THE SLY: EXPLORING LARGE SCIENTIFIC DATA SETS AS AN ENTRÉE TO STATISTICAL IDEAS IN SECONDARY SCHOOLS[1]

by James K.L. Hammerman, Ed.D
TERC, Cambridge, MA, USA

ABSTRACT

Many large scientific and social scientific data sets—for example, those about climate and the environment, medicine, population or economic trends, the human genome, astronomy—are now widely available. As secondary students explore these data they investigate fascinating and important topics that can help them better participate as global citizens. However, understanding the meaning of these data requires statistical understandings—e.g., of variability amidst underlying aggregate trends, statistical control in complex relationships, the meaning of interaction effects, expectations about small probability events, statistical versus practical significance—that are difficult and rarely taught at the secondary level. This paper explores how interest in the science can motivate exploration of statistical ideas, at an informal if not rigorous technical level, which in turn can lead to a deeper understanding of scientific ideas. The role of data visualization and analysis tools to support this learning is also explored.

INTRODUCTION

The number and variety of publicly accessible large scientific and social scientific data sets has been growing rapidly in the last several years. Many of these address some of the most compelling scientific and social scientific questions—including issues of climate change, biocomplexity and species extinction, the search for extra-solar planets or the origins of the universe, and the shifting distribution of economic resources. Interest in tapping this increase in accessible data for scientific and educational purposes has also been growing. The National Science Board (U.S.) in a study entitled *Long-lived digital data collections: Enabling research and education in the 21st Century* (National Science Board, 2005), is almost effusive in describing the potential of these widely accessible data sets:

> *It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.*

Much attention has been paid to the technical and infrastructure issues associated with making these data available to scientists and the general public—an essential step to ensure wide access and use. Less attention has been paid to whether ordinary people and students will have the statistical understandings needed to make sense of these data and their implications. This type of *cognitive access* requires learning, at least in an informal way, about important statistical ideas, and can be facilitated by easy-to-use tools for representing and exploring relationships in data.

Unfortunately, study of statistics is still limited, despite calls for increased attention to statistics education by associations of math educators in several countries (Australian Education Council, 1994; National Council of Teachers of Mathematics (NCTM), 2000; New Zealand Ministry of Education, 1992; UK Department for Education and Employment, 1999). Although the AP Statistics Exam in the US is one of the fastest growing tests (averaging more than 9000

[1] National Science Foundation (U.S.) support for this work under Grant #0822178 is gratefully acknowledged. The views expressed herein are those of the author and may or may not reflect those of the Foundation.

new students a year and now the 10th largest exam), the 108,284 students who took the test in 2008 (http://www.collegeboard.com/student/testing/ap/statistics/dist.html?stats) still represent just 3% of the total number of seniors.[2]

Other secondary students' understanding of statistical ideas may be limited, in part because their exposure to statistical concepts is often restricted (at best) to techniques for calculating basic descriptive statistics (measures of central tendency and dispersion), techniques for creating specific representations of data (histograms, box plots, scatter plots), a few ideas about sampling, distributional shape, and study methods, and perhaps exposure to lines of best fit and simple hypothesis tests. Given this level of exposure, students have trouble understanding how aggregate measures such as the mean or median can really typify a complex and variable data set, don't really understand how variability or sample size contributes to determining the precision of statistical comparisons, and don't understand the role of randomness in sampling, among other essential ideas. And there is little space in the mathematics curriculum to further explore and integrate these interesting but difficult ideas.

With so few secondary students actually studying statistics, how can we help students come to understand the statistical ideas needed to explore these fascinating large data sets?

This paper presents a multi-part learning hypothesis to address this problem:

- That secondary students could and would be interested in statistical ideas in the context of exploring real and relevant large scientific and social scientific data sets;
- That in doing so, they would develop a deeper understanding of both data analysis/ statistics and of the scientific/ social scientific content; and
- That technological tools might be essential in facilitating such explorations.

COMPELLING DATA

As mentioned above, the kinds of data that are now available can be quite compelling. As one example, a number of government agencies and universities make available a wide variety of environmental data. Those focusing just on aspects of climate change include 18 different data sets from the National Snow and Ice Data Center (NSIDC) (http://www-nsidc.colorado.edu/), data about a variety of greenhouse gases and ozone from the Earth System Research Laboratory (ESRL) (http://www.esrl.noaa.gov/), and data about climate change related gases from the US Department of Energy's, Carbon Dioxide Information Analysis Center (CDIAC) (http://cdiac.ornl.gov/home.html). There are data about other environmental issues, as well—environmental toxins and radiation levels, data about tides and ocean currents, data about abiotic conditions, biocomplexity and species distribution in different environments. Other biologically related data include a wide variety about diseases and health, about reproductive choices and outcomes, about nutrition and water quality, about the human genome, about population trends, and so forth.

There are also a variety of data about physical characteristics of the world—the sun and stars and planets and a wide variety of exotic and fascinating astronomical objects such as pulsars and black holes, data about the structure of crystals and subatomic particles and the origins of the universe. And there are data about human activities—who we are, how long we live, our sizes and shapes and family structures, our beliefs and values and political life, our economic activities, the ways we communicate or trade, the books we read, the websites we visit, and on and on.

A number of US government agencies have supported curriculum projects and educationally linked data sources to tap the potential of these fascinating topics. For example, the National Science Foundation (NSF) has recently funded projects such as the "Pulsar Search Collaboratory" (http://www.pulsarsearchcollaboratory.com/) which asks students and teachers to use radio astronomy data to look for new pulsars; "OssaBEST," (http://ossabest.armstrong.edu/)

[2] Based on 1/4 of the 16.2M 9-12 grade students in 2008 (http://nces.ed.gov/fastfacts/display.asp?id=65).

which explores environmental issues on a Georgia barrier island, and "Sharing the Message of Global Change" which explores climate data. Other agencies, too, have developed curricula and educationally linked data sets: e.g., the National Institutes of Health (NIH)-funded curricula about the human genome (BSCS & Videodiscovery, 1999; Horn, 2002); the National Aeronautic and Space Administration (NASA)'s "My NASA Data" project (http://mynasadata.larc.nasa.gov/data.html) and "Advanced Composition Explorer" (http://www.srl.caltech.edu/ACE/) providing data on the solar wind; the National Oceanic and Atmospheric Administration (NOAA)'s "National Climatic Data Center," (http://www.ncdc.noaa.gov/oa/ncdc.html) and the National Estuarine Research Reserve System (NERRS)'s System-Wide Monitoring Project (SWMP or "swamp" data: http://nerrs.noaa.gov/Monitoring/) providing weather, water quality and biotic data, among many others.

These and other science and social science topics can be of interest to secondary students who are interested in similarities and differences among people, in themselves, in the nature of the world around them and in their place in it, and in issues affecting their future and the future of the planet. Issues of environmental and social justice may be particularly compelling for some students (Gutstein, 2003), and data is often a useful tool in coming to understand these issues.

Science and social science teachers may be able to build on students' interests in these topics, but may not have the statistical tools and resources to help them explore these issues deeply. In fact, there are a number of cognitive issues about understanding large data sets which can be addressed in these contexts.

TECHNOLOGY TOOLS

Gaining access to interesting data is a critical first step. Unfortunately, too often when representational tools are available as part of these data sets (beyond just data tables), they are often designed by scientists to reflect research uses, rather than with an eye on learning or understanding by the general public. This needn't be the case, though. Some sources of data are also built around new ways to display and represent data, such as Google's online software "Gapminder" (http://www.gapminder.org/world/) which shows relationships among a number of international demographic variables over time, providing a more user-friendly representation of data.

Tools for multiple linked representations of data built into commercially available educational data analysis software such as *Fathom*, *TinkerPlots* and *InspireData* (formerly TableTop) (e.g., Fathom: Finzer, 2005; InspireData (formerly TableTop): Hancock, 2006; and TinkerPlots: Konold & Miller, 2004) can also provide powerful tools for exploring data, though interfaces to access available data may need to be created. One powerful example of this is Fathom's capacity for downloading and exploring US Census micro-data.

Representations are important because the inter-relationship between representations of data and how people are able to think about data—the affordances these representations provide—are critical for educational research and design. Bakker & Gravemeijer say: "We have come to see this back-and-forth movement between graphs and informal statistical notions as an important heuristic for instructional design in data analysis" (2004, p. 9).

Even with smaller data sets, people use the available tools to find ways to reduce the cognitive complexity of what they're seeing—e.g., using proportional reasoning around cut points to create measures for grouping and summarizing data (J. K. Hammerman & Rubin, 2004). Other tools will be necessary to help people deal with additional types of complexity associated with large data sets—relationships of multiple variables, cyclical patterns in time series data, relationships among variables in space and time.

CONCEPTUAL CHALLENGES AND OPPORTUNITIES

What would people need to understand to make use of these interesting data sets, given appropriate tools for exploring them? (And what additional tools might help them explore

specific different types of data?) The literature suggests a variety of challenges in what students (and people more generally) understand or don't understand about statistics needed to make sense of large scientific and social scientific data sets. At the same time, there may be some non-intuitive conceptual advantages to starting with large data sets rather than smaller ones.

Aggregate characteristics. Students often have a difficult time coming to attend to emergent aggregate features of data sets as a whole (e.g., measures of center, spread, shape) rather than attending to individual or groups of cases (Bakker, 2004, p.13-14; Hancock, Kaput, & Goldsmith, 1992; Konold & Higgins, 2003; Konold, Higgins, Russell, & Khalil, 2003; Konold, Pollatsek, Well, & Gagnon, 1997; Lehrer & Schauble, 2004). Some suggest this move to attending to aggregates is developmentally key to making progress in statistics and may go through several intermediate stages (J. K. L. Hammerman & Rubin, 2006; Konold, et al., 2003). However, it may be that in large data sets, distributional shape (center, spread, etc.) increases in salience as individual points seem to matter less when there are so many of them. In such circumstances, are students more likely to talk about and compare aggregate features?

Sample size. People often have a hard time understanding the ideas behind the Law of Large Numbers—that aggregate characteristics of data become more stable as sample sizes increase (Rubin & Bruce, 1991; Saldanha & Thompson, 2002; Sedlmeier, 1998; Sedlmeier & Gigerenzer, 1997). In part, they think of samples as requiring a proportion of the entire population. The persistence of such proportional ideas may explain why people are reluctant to believe that political polls tapping the opinions of 500 to 1000 people can possibly be representative of a country of several million citizens. Yet, starting with a large amount of data may provide some non-intuitive advantages. When data sets are large enough, then any differences or relationships that *can* be observed are statistically significant, essentially based on the logic of the Law of Large Numbers—when N is big enough, expected variability in aggregate measures is small enough that observable differences are unlikely to occur by chance. Once students have seen a relationship in data, then asking them whether *all* these data would be needed to find such a relationship can address the Law of Large Numbers from the opposite side. That is, now that we know a relationship exists, how *few* data points would we need to notice it, and how sure could we be about it in such cases?

Practical significance. The fact that small differences observed in large data sets are *statistically significant* raises further questions about the *practical significance* of any such differences. For example, the graphs below show a small statistically significant difference in the means of two groups. Yet by most intuitive measures—range, shape, location of the mean (blue

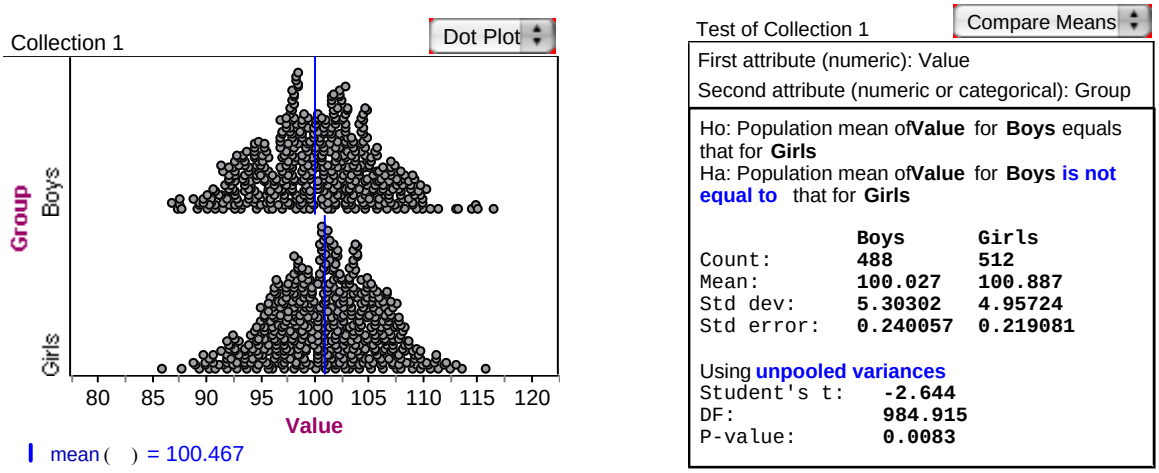

Figure 1: Statistically significant difference may not have practical significance

line)—most students would likely see these distributions as the same. And, for most practical purposes, they *are* the same. If these graphs represented gender differences in scores on some

test, a statistician might wonder why girls scored slightly higher than boys on average. But teachers and parents would see a lot of variability in scores for both genders with distributions that are almost the same. Focusing on the meaning of statistical findings can help students learn to use, interpret, and question statistics they encounter in the world.

Low probability events. At the same time, large data sets provide an opportunity to look for low probability events that may be scientifically interesting. Even as students come to think of data as a proportionally representative subset of a larger population (Saldanha & Thompson, 2002), they may have a hard time understanding that this proportional view extends to rare events as well—that rare events *will* (likely) occur if there are enough cases. This idea is tricky because it requires students to extend their hard-won multiplicative view of data even to events whose most salient feature is that they are highly unlikely, to predict that they *will* be found.

Sample v. census. In some cases involving large data sets, the data are *not* a sample of the population but are the entire population—a census. Statistical methods don't make sense in such cases—we already *know* that a counter-factual hypothesis *isn't* true, so statistical methods exploring its likelihood under the null hypothesis are nonsensical. Some researchers argue, however, that all data are a sample of something—even the complete record of some set of events is seen as a sample of the process which produced the record, with the idea that findings from one situation could be used to make predictions for future behavior (Frick, 1998). This distinction may be essential in such cases, may be an interesting one to explore with students, and builds on the exploration of how *few* data are needed to observe the same relationships noted above.

Multiple attributes. Sometimes large data sets are large not (just) because of their number of cases but (also) because of their number of variables or attributes. Data sets from the social sciences, ecology, health and medicine, and about internet behavior, among others, often contain a large number of variables. In some sense, it is these sorts of data sets that the National Science Board is so excited about because further analysis could find previously undiscovered relationships. Yet there are important conceptual difficulties in coming to understand relationships among more than two variables at a time. The idea of statistical control—that we can look at *all* the data and still be able to say something about relationships between two variables holding a third constant—is difficult. (Why don't we take a sequence of subsets?) Representing these relationships can also be a challenge. Three dimensional graphing tools can certainly address this problem but require some effort to be able to understand and interpret the resulting "planes of best fit." Sometimes color can be used in a 2-D scatter plot to show levels of a third variable. which can help explain predicted relationships after statistical control (see Figure 2). Animations have also been used to show a third variable, especially (but not exclusively) when that variable is time. The complexity of depicting these multi-variate relationships increases further when there are interaction effects—that is, when the relationship between two variables varies by the level of a third. It is an open question whether and how these and other representational tools can make such complex data sets accessible to secondary students.

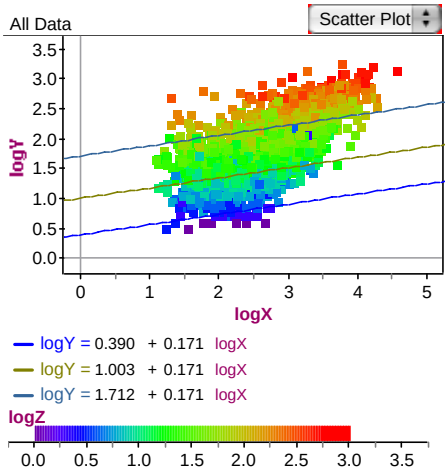

Figure 2: Color and regression lines tracking three levels of a third variable

Data involving time series or spatial proximity. Several of the interesting and available large data sets depict relationships where time features prominently. Accounting for cyclical features of such data to look for other relationships may require tools to factor out the auto-correlative behavior. While students should be able to understand the existence of cyclical (e.g., daily, weekly, seasonal) behavior, they may not need to understand technical aspects of how to

deal with these. Similarly, if students are exploring geographically/ location-linked data, they may not need to understand how to create spatially linked statistics, even if they can informally describe clustering/ patterns in the data. Creating tools that may facilitate these explorations will help in deciding whether students will be able to succeed in engaging with such data.

CONCLUSION

Large scientific and social scientific data sets hold much promise for engaging students in statistical explorations. However, much work remains to be done to understand how students think about such data, and to create software exploration tools and appropriate curriculum materials that will support deeper learning about the statistical and content-related ideas.

REFERENCES

Australian Education Council (1994). *Mathematics—A curriculum profile for Australian schools.* Carlton, Victoria: Australian Education Council.

Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools.* Utrecht, The Netherlands: CD Beta Press.

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In J. B. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, Netherlands: Kluwer Academic Publishers.

BSCS, & Videodiscovery (1999). Human Genetic Variation Retrieved 1/20/08, from http://science.education.nih.gov/supplements/nih1/Genetic/guide/pdfs/nih_genetics.pdf

Finzer, W. (2005). Fathom™ Dynamic Data™ Software (Version 2.0). Emeryville, CA: Key Curriculum Press.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers, 30*(3), 527-535.

Gutstein, E. (2003). Teaching and learning mathematics for social justice in an urban, Latino school. *Journal for Research in Mathematics Education*, 37-73.

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41.

Hammerman, J. K. L., & Rubin, A. (2006). *Rule-Driven and Value-Driven Measures: A Schema for Approaches to Using Distributional Data to Compare Groups.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hancock, C. (2006). InspireData (Version 1.0): Inspiration Software, Inc.

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

Horn, T. M. (2002). Background Paper on Human Genome Education at the Pre-College Level Retrieved 20 January, 2008, from http://www.genome.gov/10005289

Konold, C., & Higgins, T. L. (2003). Reasoning About Data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A Research Companion to "Principles and Standards for School Mathematics"* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics (NCTM).

Konold, C., Higgins, T. L., Russell, S. J., & Khalil, K. (2003). *Data seen through different lenses*.Unpublished manuscript, Amherst, MA.

Konold, C., & Miller, C. (2004). TinkerPlots™ Dynamic Data Exploration (Version 1.0). Emeryville, CA: Key Curriculum Press.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the 1996 IASE Round Table Conference*. Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635-679.

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

National Science Board (2005). *Long-lived digital data collections: Enabling research and education in the 21st Century*. Washington, DC: National Science Foundation.

New Zealand Ministry of Education (1992). *Mathematics in the New Zealand curriculum*. Wellington, NZ: New Zealand Ministry of Education.

Rubin, A., & Bruce, B. (1991). Using computers to support students' understanding of statistical inference. *ATMNE Journal*.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*, 257-270.

Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making, 11*, 281-301.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33-51.

UK Department for Education and Employment (1999). *Mathematics: The national curriculum for England*. London: UK Department for Education and Employment & Qualifications and Curriculum Authority.