Numerical indigestion: how much data is really good for us?

blogs.lse.ac.uk/impactofsocialsciences/2012/11/27/goldstein-how-much-data-is-good/

11/27/2012

We are swimming in 'big data' and despite their performances as advocates of data freedom, policymakers don't seem to bear any responsibility for educating the public on how to read it. ***Harvey Goldstein*** *believes that academics must make it their mission to explain that evaluating statistical information is far from trivial.*

Modern cultures are deeply imbued with notions of measurement, and it is often assumed that the introduction of new measurements or the expansion of existing ones needs little justification. The default assumption is almost always that more measurement is a good thing.

It is the rapid development in information technology over the last 30 years that has generated enormous possibilities for the collection, processing and publication of data of all kinds. In countries such as Denmark there exist linked files for all citizens containing personal data about education, health, employment etc. In the UK there is a national pupil database which tracks moves and achievements of every pupil in the maintained education system. Data on crime, health outcomes and educational test can be accessed in the UK via the portal www.data.gov.uk In the US the Open Data movement is well established and the Government website, www.data.gov has very large numbers of freely available datasets. In Canada there is www.data.gc.ca .

There is no doubt that the existence of such data, where it has been collected with care and reliably, provides a valuable resource for policymaking, for research and for informing citizens about society. Indeed, there is a continuing debate about how to ensure reliability and relevance and how to avoid misleading inferences when such data are published and the RSS GETSTATS initiative is a recognition of how seriously this is taken by the profession. Yet almost all of this debate happens among professionals themselves and there is little public discussion of these issues through the media. Policymakers in general tend not to see that they have any responsibility to 'educate' the public and highly visible public advocates of 'data freedom' such as Heather Brooke (2010) fail to touch on such issues.

What I would like to do in here is pose some questions about the data tsunami threatening to engulf us. The data really are big. For example, Twitter in 2010 is estimated to have generated 3 petabytes (of data and there are fascinating insights available to those who are interested in exploring such datasets. In particular, their size means that subtle interactions can be explored in ways that were not previously possible. How do we cope with the existence of such huge amounts of data, and how might we develop ever more sophisticated machinery to handle it, or do we simply let it wash over us and pick up the pieces? In this article I want to sound a warning by looking at one particular use of a relatively big data set.

Comparing schools

Every January the Department for Education, responsible for school education in England (other parts of the UK schooling system are administered differently) publishes increasingly large quantities of data about every maintained (public) secondary school in England. Data available for researchers and include items such as test scores and exam results, among many others.

Table 1 is an extract showing percentages achieving higher grades in the GCSE (16 year) examinations as well as percentages making 'expected progress' in English and Maths between year 6 (just prior to entry to Secondary school) and year 11 (end of compulsory schooling). Note that private fee paying (independent) schools have no data on this since their students are not subject to regular testing throughout their schooling. The measure of expected progress used is based upon dividing the test score distribution at year 6 into seven or more groups and seeing what percentage of students in each group achieve each of nine year 11 grades at GCSE. A (not clearly explained)

decision is then made for every year 6 test level, what counts as 'expected progress' in terms of the grade achieved at year 11. The overall percentages achieving expected progress is reported. As can be seen, some two thirds are deemed make 'expected progress' in English and mathematics.

TABLE 1 Summary of performance

School name	School type	% of pupils making expected progress		% achieving 5+ A*-C GCSEs (or equivalent) including English and maths GCSEs				% achieving all English Baccalaureate subjects	% achieving grades A*-C in English and maths GCSEs
		English	maths	2011	2010	2009	2008		
England - all schools		NA	NA	58.9%	53.5%	49.8%	47.6%	17.6%	59.5%
England - state funded schools only		71.8%	64.8%	58.2%	55.2%	50.7%	48.2%	15.4%	58.7%
Schools *(click box to add schools to your selection)*									
Fortismere School	Foundation School	87%	88%	79%	73%	73%	70%	48%	81%
Highgate School	Independent School	NP	NP	100%	0%	0%	0%	78%	100%
Channing School	Independent School	NP	NP	100%	98%	100%	100%	35%	100%
Bishop Douglass School Finchley	Voluntary Aided School	66%	68%	45%	39%	35%	41%	2%	45%
The Compton School	Community School	85%	70%	65%	58%	68%	54%	18%	65%
The Compton School	Academy	No KS4 data available for this school							
Highgate Wood Secondary School	Community School	86%	76%	68%	46%	51%	46%	24%	69%

What we do not see in this table is any expression of uncertainty, despite some of these percentages being based upon small numbers. This issue becomes starker when we look at some more detailed comparisons as shown in Table 2 for two of the schools in Table 1.

Table 2. KS4 outcome by KS2 groups: Those below level 4 (Low attainers); those at level 4 (middle attainers); those above level 4 (high attainers).

Percentage achieving 5+ A*-C GCSEs (or equivalent) including English and maths GCSEs	All pupils	Low attainers KS2(Number)	Middle attainers KS2	High attainers KS2
Bishop Douglas	45%	17% (24)	48%	94%
Highgate Wood	68%	9% (32)	67%	96%

The aim of Table 2 is to provide an estimate of 'differential progress' for students starting with 'low', medium' and 'high' achievements at year 6. Thus, we see that despite students at Highgate Wood (HW) making more overall progress in English and Maths than those in Bishop Douglas (BD) (Table 1) and also obtaining more overall 'good' results at GCSE, among the low achievers twice as many get good GCSE results at Bishop Douglas than at Highgate Wood. If we look at the numbers involved, however, we can work out that these percentages equate respectively to just 4 and 3 students! So again, we see a disregard for any serious attempt at statistical accuracy amid a plethora of numbers.

It is tempting to view publications of official statistics such as these as an aberration that will be corrected when the issues are pointed out. With luck, this may happen, although the author's personal experience with this particular Government department suggests that, although it sometimes happens, on the whole this constitutes wishful thinking. More importantly, given the volume of data being produced and its take-up by the media, bloggers and others, much of what is dubious, unclear or simply misleading, is likely to thrive. It is also worth pointing out that misuse, deliberate or otherwise, appears to cross conventional political boundaries. There seems to be no obvious

mechanism in place designed to ensure quality control.

One tempting possibility is to consider setting up internet user forums, each of which specialises in a particular source of data, for example crime statistics, and making them open to anyone interested. To work effectively these would need some kind of moderation, as with many Wiki sites, to avoid simply becoming a collection of questions and statements. Evaluating statistical information is a more complex activity than rating a hotel room and it is, after all, what professional groups are trained to do through meetings, discussion forums and journals. In some cases there are existing auditing organisations that can carry out some of these functions. For example the UK has its Statistics Authority that monitors and reports on official statistics, including their quality. The media possibly has the key role since it is, currently, the source that most people have access to. On the whole, its record is not very commendable. Too often journalists are simply not qualified to understand the statistical issues, as well as succumbing to the temptation to report what may suit their own views or make good headlines. There are notable exceptions, which hopefully will grow, but these tend not to appear in the mass circulation media.

What about professional organisations such as the national statistical societies or the International Statistical Institute? The RSS, for example, has recognised the general problem of statistical literacy with its GETSTATS campaign, but this is not specifically focussed on the evaluation of databases. Nevertheless, professional bodies do get involved in commenting on particular issues, although again their resources are limited.

So where does this leave us? My own view is that the profession has to recognise these challenges and try to find ways of meeting them. It needs to recognise that it has a special responsibility to provide leadership and to explain that evaluating statistical information is far from trivial, requires effort and is not uncontentious. It will often mean confronting data providers with questions and criticisms and may involve condemning what is trivial, biased or misleading. It may also invove pointing out that simply publishing more statistical information may not lead to greater enlightenment. If professionals don't see this as their role then it is difficult to see who else is going to fulfil it.

Note: This article gives the views of the author(s), and not the position of the Impact of Social Sciences blog, nor of the London School of Economics.

About the author: Professor Goldstein is a chartered statistician, has been editor of the Royal Statistical Society's Journal, Series A, a member of the Society's Council and was awarded the Society's Guy medal on silver in 1998. He was elected a member of the International Statistical Institute in 1987, and a fellow of the British Academy in 1996. He was awarded an honorary doctorate by the Open University in 2001.