



Dealing with the data deluge

2013 is International Year of Statistics. **Chris Skinner** explains why this multi-faceted subject merits global celebration.

Over the last decade or so the explosion in availability of digital data in field after field, sometimes called the data deluge, has greatly expanded the need for people who can make effective use of these new kinds of data. Indeed, the report on “big data” from the McKinsey Global Institute last year anticipated a major expansion in jobs worldwide for people with the expertise to use modern statistical methods to gain insights from large data sources. This can only be heartening news to today’s students of statistics. Google chief economist Hal Varian’s widely cited 2009 prediction that statistician will be the “sexy job in the next 10 years” doesn’t look hollow. Varian contrasted the huge explosion of availability of essentially free digital data with the scarcity of the ability to understand data and to extract value from it.

A key type of statistical technique which has found particular success with new large data sources is predictive modelling, often called predictive analytics in its applications to business. Here, the aim is to predict unknown outcomes given observable features: for example, translate a piece of text into English from another language. Models are fitted to “training data”, where a rich source of humanly translated pieces of text is available, and then applied to new pieces of text where translation is needed. A remarkable feature of tools like Google Translate is their reliance purely on statistical methods and rich data sources instead of knowledge of grammatical rules or linguistics. The same applies to Google’s matching of advertisements to content or to Amazon’s personal recommendations, which operate without reference to behavioural or social science knowledge about how people form preferences. To be sure, such developments cannot be attributed to the discipline of statistics alone. Computer science has played a key role and the skill set in demand

in this new data-centric world is often referred to as data science, drawing on skills from both statistics and computer science.

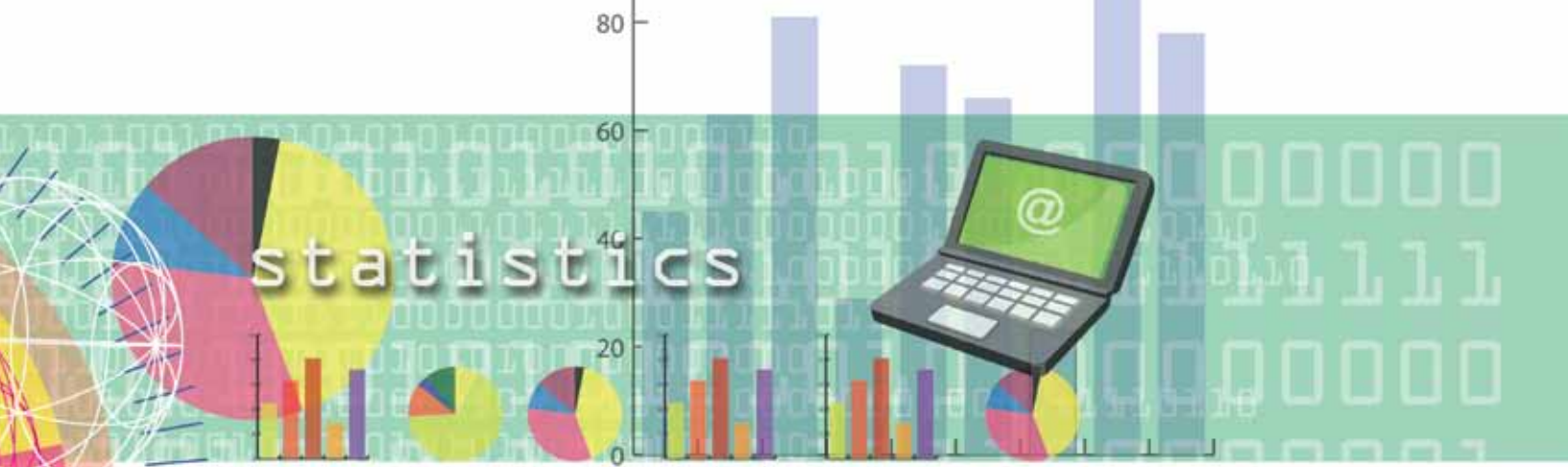
Successful new applications of sophisticated statistical techniques have certainly not been confined to the business world. Very many academic disciplines have seen similar massive expansion of digital data, for whose exploitation new statistical techniques have proved vital. Genetics and astronomy are just two. Nearer to home, on my return to the Department of Statistics at LSE after 30 years, I’ve been impressed by how developments in the field of finance have stimulated particular expansion of the department’s research.

The innovations in statistical science associated with such developments should not be taken to suggest that statistical methodology is, in any predominant way, a matter of fashion. The core principles and ideas of statistical science remain crucial and are well established. I was struck to hear how a statistician at Google found some of the principles of experimental design established in the first half of the 20th century to be crucial for the many experiments that Google runs with its search engine. My own special interests are in the statistical methodology of surveys, where many of the core methods were established decades ago. At the time of writing, I am about to participate in a technical advisory group for a new international survey of higher education, the feasibility of which OECD is exploring, with the aim of testing student and university performance globally. The relevant statistical skills I expect to bring are the core long-established ones of survey statistics.

Much of this core methodology of statistical science does not require high technical sophistication, and much of the richness of the modern subject comes from it not just being practised by professionally qualified

statisticians. Much more widely, it is used by those whose primary expertise is in substantive fields and who bring subject-matter enhancements to the application of statistics. This more widespread use of statistics does, however, raise questions about how the appropriate skills are acquired. The teaching of statistics at LSE goes right back to its origins, when Sir Arthur Bowley was appointed part-time lecturer. There is a long tradition at the School of statistics providing key underpinning for teaching and research in the social sciences. But, while some of the greatest enthusiasm for statistical data analysis can come from non-statisticians, it is clearly not the case that statistics is uniformly loved! In recent years there has been much comment on the shortfall in quantitative skills among researchers in the social sciences in the UK, and the ESRC, the Nuffield Foundation and the British Academy have all invested in initiatives designed to combat this shortfall. This is certainly not just a UK matter. I heard clear echoes of such concerns when I directed a four-year European Science Foundation programme on Quantitative Methods in the Social Sciences, which sought to build capacity among young researchers in 19 European countries.

The democratisation of statistics across academic and professional fields is nothing new. Of more recent note has been its much wider promotion with the increased availability of statistically orientated software on laptops, tablets or even smart phones, combined with the increased availability of a vast range of data sources. The UK government recently issued an Open Data White Paper, subtitled “unleashing the potential”,



“Statistics can be used to help make well-informed choices, take sound decisions and understand the society and the world we live in”

which seeks to promote further ways of expanding the free availability of data, “the 21st Century’s new raw material”, as it is termed.

Statistics here can represent a valuable life skill, enabling each of us to convert data into useful information. Statistics is a tool for problem-solving and decision-making, usable in many areas of life. It can be used to help make well-informed choices, take sound decisions and understand the society and the world we live in. The Royal Statistical Society has initiated an ambitious ten-year “getstats” campaign (www.getstats.org.uk) aiming to reach

out to new and wider audiences, to raise awareness of the benefits of statistics, and to create paths to the know-how and skills needed. The wider use of data is also being promoted by a new generation of “data journalists”.

Statistics in the public arena are also, of course, at risk of misuse in support of different interests and political ends. Public distrust of official statistics has been a particular issue in the UK and common reasons cited in surveys of trust include concerns that figures are misrepresented or spun by politicians and the media. The getstats campaign is also seeking to address statistical literacy in the media, parliament and politics.

These are just some of the many facets of this discipline and its relevance to so many kinds of constituency. There are certainly reasons for celebration. Look out to learn more in the International Year of Statistics (statistics2013.org). ■



Chris Skinner is Professor of Statistics at LSE.



UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

LSE-UCT July School 2013

Cape Town, South Africa, 1-13 July 2013



An exciting new opportunity to study important social science issues relevant to Africa today, taught by world-leading faculty from LSE and UCT in the beautiful, cosmopolitan city of Cape Town.

Courses offered include: Economics, Geography, History, Management, Economic History, Government, International Development.



twitter.com/LSEUCTJulySch



facebook.com/LSEUCTJulySchool

LSE-UCT.July.School@lse.ac.uk

For more information please see lse.ac.uk/LSE-UCTJulySchool