Lectio magistralis: Science, Data and Statistics

A. C. Davison[*]

December 18 2009

I greatly regret my inability to speak your own expressive language and to have to give this lectio in English.

May I begin by expressing my deep appreciation of the honour that you do me by the conferral of this *laurea honoris causa*. Over the years of my visits to Padova I have developed a lively respect for the high quality of your faculty and students, and it is particularly gratifying to have my work recognised so generously by one of Europe's oldest and most distinguished universities—especially in this anniversary year of your Faculty of Statistical Sciences.

Galileo and Gaia

This is a year of important anniversaries in the history of science. Four hundred years ago this month, here in Padova, Galileo Galilei built a telescope with a magnifying power of twenty, with which he quickly discovered the four brightest moons of Jupiter, found that the Milky Way is made of myriads of individual stars, and observed the craters and mountains of the moon. Modern academics are used both to international collaboration and to keen competition, but this is nothing new: Galileo had to work very quickly because a Dutchman had travelled to Venice to try and sell primitive telescopes—immensely valuable to an important maritime power—to the Venetian Senate. This then asked a friend of Galileo's to inspect the Dutch telescope, and he did so very slowly, while Galileo himself produced a better instrument—a good example of academic and commercial skullduggery. Once Galileo had made his first astronomical discoveries, he published them in March 1610—just a few weeks after they were made—no difficulties with important results being held up by slow refereeing and incompetent editors. But I digress . . .

Galileo's astonishing observations helped to overturn the immovable Artistotelian universe, but his impact on science was more profound even than this. He was not the first to use what we now call the scientific method—observation, experiment and the careful evaluation of hypotheses in the light of evidence—rather than investigation by pure thought as propounded since Aristotle by the 'Peripatetic' philosophers, but he was the first to apply it systematically to a wide range of phenomena—navigation, pendulums, the action of gravity, mechanics, hydrostatics, magnetism, the strength of materials, ... Now known as the father of modern science, he undertook his research as Professor of Mathematics; a precursor of the impact that mathematical scientists can make across the range of knowledge. He published numerous important books, among which *Il Saggiatore* (*The Assayer*), published in 1623, summed up his understanding of the scientific method. In it he wrote:

[*] Institute of Mathematics, IMA-FSB-EPFL, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, `Anthony.Davison@epfl.ch`

> *La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.*
>
> The book of the Universe cannot be understood unless one first learns to comprehend the language and to understand the alphabet in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures, without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth.

Today we would add to that list other characters, such as derivatives, matrices, and random variables.

The gathering of high-quality data is crucial, but without some organising principle, data are just arbitrary collections of numbers and facts—what Rutherford rather unkindly called stamp collecting. Mathematical reasoning provides coherent structures within which to classify data, to make connections between them, to extrapolate beyond them, and, with the aide of statistical ideas, to assess whether divergences between the data and some hypothesis indicate a deficiency in current theory.

Astronomy has come a long way since 1609. The Gaia Mission is a project of the European Space Agency, which intends to launch a satellite in 2012 that will collect data on one billion astronomical objects. One task that will result will be the classification of around 100 million variable stars into different groups, based on pictures such as those shown here. The human eye can do this for perhaps a few hundred pictures, but to achieve this objectively, quickly and automatically, classifying at least one star each second, will require sophisticated—and fast—classification tools. Here is a huge playground for statisticians familiar both with astrophysics and with modern classification.

Darwin and evolution

The year 2009 also marks 150 years since the publication of Charles Darwin's 'The Origin of Species'. Just as Galileo's contributions to the Copernican revolution helped to show us that our planet is not the centre of the universe, Darwin's great insight showed us that mankind is not unique in creation, that every living being is linked to every other, and perhaps to every being that has ever lived, by an evolutionary tree. Like Galileo, Darwin was an assiduous experimentalist and an acute observer—and like Galileo he generated data, some of which were later used in the statistician and geneticist R. A. Fisher's great book *Design of Experiments* to explain the importance of randomisation. This statistical principle has perhaps contributed more to human health, through the creation of the controlled clinical trial and all its variants, and to the gradual emergence of evidence-based medicine, than any single medical discovery. Fisher famously said:

> To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.

The design of experiments is so under-appreciated that many of the statisticians present will have had the unpleasant experience of having to explain to someone that their hard-won data cannot possibly give them the conclusion that they hope for, because of a failure to include a baseline by using controls, to avoid bias by randomisation, or to control variation through the use of blocking.

Like astronomy, biology has come a long way. Here are some data from a series of experiments being performed by Amélie Dreiss and Charlène Ruppli at the University of Lausanne on conversations between young owls; the goal being to see whether the speech patterns of each owl can be characterized and whether their state of hunger affects the conversations, in order to understand the negotiations between the nestlings before a parent arrives back at the nest with food. Evolutionary game theory suggests that the birds should react to each other in order to reinforce sibling honesty and to ensure an optimal allocation of resources. Once the data have been recorded and treated acoustically, the statistical challenge begins: is it possible to characterise the calls of each owlet? What is the dynamic of a dialogue? Are there rules that determine who speaks when, and who receives the next meal?

Standard time series models are unlikely to be directly helpful here, and some new ideas may be required to give an incisive answer to these questions. Fisher's comment is partly true here: since the baby owls are never chosen twice, but always appear with one sibling, it is impossible to answer questions about individual characteristics: all that can be described is their behaviour in the presence of one other owl.

Datasets large and small

The two problems I have described exemplify how modern measurement technologies and data processing capacities provide scientists with unprecedented amounts of complex data, for which statistical tools and ideas are more than ever essential. This is true not only in science, but also in commerce: randomized experiments are performed when we click on links to web pages, to compare our reactions to different adverts. Not for nothing did the Chief Economist at Google, Hal Varian, say in an interview at the start of this year:

> I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but ... the ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.

One of the keys roles of a Faculty of Statistical Sciences such as yours is to train young scientists who can rise to this challenge, and who will reap the benefits that will come with having skills that are in high demand.

Massive datasets pose big challenges both in statistical and in computational terms, but such datasets may boil down very small once a very specific question is asked. For example, the Large Hadron Collider at CERN near Geneva, at last coming up to full capacity after a series of delays, is expected to should shed light on the origins of the Universe and on the correctness of the 'Standard Model' of particle physics. One key task is the hoped-for detection of the Higgs boson, the so-called 'God particle', which is thought to play a key role in giving mass to the other particles, but which is the only fundamental particle whose existence has not yet been confirmed. Yet although the LHC will produce more data than the entire European telecommunications network, once the noise is removed, the discovery or not of the Higgs boson will depend on relatively few observations, and statistical methods adapted for small samples will be needed. Many of the people present have contributed to the development of such procedures, which stem from fundamental notions of inference—likelihood, sufficiency, ancillarity, marginalisation, and so forth—originated by Fisher around 90 years ago, but which are still highly relevant today. It is the interplay between such general ideas, which stem from philosophical and mathematical considerations, with the huge variety of possible applications, that gives statistics such interest and charm—added, of course, to the thrill of scientific detective work in a haystack of data.

Rare events

To close, I'd like to turn to my main current preoccupation, the study of rare events through the statistics of extremes. Two recent happenings prompt this. The first is the current crisis in global finance, which is staggering from a sequence of major shocks, with more perhaps to come. One popular view of this, promoted by many newspaper articles and television programmes, is that 'a mathematical formula blew up Wall Street', by providing a simple way to measure the risk of two simultaneous defaults—simple but fundamentally flawed. According to this view the Gaussian copula was adopted wholesale by banks and other financial institutions, who then proceeded to badly under-estimate risks with it, because as the picture shows, under this model extreme events—such as the simultaneous collapse of two major financial institutions—cannot occur. In fact, as we have seen, they can occur with frighteningly high probability, because, as many respected academics said long before the crisis struck, the Gaussian copula is much too simple to provide a realistic mathematical analysis of such risks. Even after the shocks we have seen recently, these risks remain, and an important statistical challenge is to find better methods to estimate them, and then, much more difficult, to try and see that these methods are applied correctly.

An even more important challenge is the likely consequences of climate change. Even under optimistic scenarios, a 2–4°C rise in global mean temperatures over this century now seems inevitable, and this will have profound effects on our environment—on the crops and water on which we depend, for instance—and thus, perhaps, on our survival as a species. Events such as the heat-wave of summer 2003, which is thought to have led to the deaths of around 37,000 Europeans, are likely to be commonplace by 2050, yet at present we have no well-accepted and mathematically sound method for modelling the probabilities and likely consequences of these disasters. Windstorms, heavy rainfall and flooding, droughts and other hazards are also predicted to become more frequent, and will affect the built and natural environments that we now take for granted. Constructions now under way in certain places may have to survive events never before seen there, and engineers will have to rely on forecasts from mathematical and statistical models that are as yet poorly understood, and in which proper assessment of uncertainty will play a crucial role. Providing such assessments is what statistics is about, but if the results are to be useful then close collaboration between scientists from many different disciplines will be essential. It will be the task of all of us to meet such challenges to the best of our abilities, so that rational decisions can be taken informed by the highest values of our common humanity.

Closing

Please allow me to repeat my deep appreciation of this *laurea honoris causa.*