LETTER FROM THE PRESIDENT

Statistics the ‘Wiki’ way?

The president of ENBIS, Fabrizio Ruggeri, suggests that we could fulfil our mission by broadcasting reports of problems solved with statistical tools through a ‘Wiki’ world about statistical consultancy

In the last few years a colleague of mine (a computer scientist, and much more!) has been involved in the Wikipedia project (see www.wikipedia.org). I have followed its developments during lunches and corridor talks with him. The idea behind Wikipedia is simple: it is an international project where people are building the largest encyclopaedia ever, in many languages and covering a multitude of areas. Contributions are voluntary, although there are rules and an etiquette to be respected. If you type the word ‘Wikipedia’ in Google you can find (end of May 2006) 366,000,000 entries. It is a social phenomenon that has been the subject of discussions in both scientific and popular journals. I do not want to add my opinion on this aspect but I just want to express some ideas triggered by today’s talk with my colleague Alberto Marini (and the kind, but firm, pressure exercised on me by Tony Greenfield, the editor of the ENBIS pages within *Scientific Computing World*, to write these notes). Alberto talked about a ‘Wiki’ area in the web of our Institute and, to address my doubts, he explained that it could be useful to create a local Wikipedia collecting information useful for our research.

Could Wikipedia (or a similar system) be useful in our profession as statistical consultants? We are used to a system in which there is an available mix of scientific papers and books, tutorials (online and offline), courses, software and their manuals, and humans (from academic professors to statistical consultants, with both professions sometimes performed by the same person) which are set to solve problems customers might have. What would be the role of all of them in a world in which individuals and companies might have free access to very specialised, structured information on the web, as the one provided by Wikipedia? Would the role of statistical consultants be undermined?

I think that, as learned from the Wikipedia experience, it is possible (but not easy) to get skilled people to devote some of their time to prepare entries for a version about statistical consultancy. It would be possible, especially if large associations, like ENBIS, would promote it. If committed people are a relevant requisite, a plan and a philosophy behind it are even more important. It is not just a matter of having tutorials and books on line: this is already possible, although for a fee in general. Practitioners can have easy access to statistical tools on line but this could be efficiently complemented by taking a new, ‘Wiki’ approach. In a ‘Wiki’ world about statistical consultancy, people should write about real problems they faced and how they solved them (or not...) with statistical tools, on how to use the latter in practice, the myths and the misunderstandings about them. Once these accounts are available on the ‘Wiki’, I don’t think the role of the statistical consultants would be diminished, but it would be better understood. There could be practitioners who would find inspiration from the entries and address (or, at least, try) their problems by themselves. The toughest, and most stimulating, problems would be left to consultants and, last but not least, practitioners would communicate with better trained people, more keen to understand their language and tools. At the same time, the information on the web could attract people with problems, people who had no previous ideas of the stochastic nature of their problems and, even more important, of their possible solution with statistical methods. The ‘Wiki’ approach could be useful to tackle the problem for all actors in consultancy (industry, private consultants and academia) to try to understand each other, and overcoming language barriers, as I discussed in previous letters.

Of course, I do not believe in a ‘magnificent and progressive fate’ (translated quote from a famous poem by Giacomo Leopardi, one of Italy’s most important poets), but I am convinced that the actual implementation will be far from the idealistic representation I gave a few lines above. Nonetheless, I think we should reflect on this ‘ideal world’, consider seriously the opportunities offered by the ‘Wiki’ approach and think how we can exploit them, although not in their entirety. There are many issues to be addressed, such as the content and the organisation of the system, the control of the quality of the information, industrial secrets preventing the diffusion of most data and methods, copyright by publishers (and their opposition, in general). I think ENBIS is very suitable, because of its mission and composition, to promote such discussion and to check the possible implementation of such system.

At ENBIS we have been discussing how to promote statistical awareness and improvement of statistical knowledge. We have some special interest groups, open to all members, devoted to this. A first answer comes from our European Union project PRO-ENBIS, funded within the Fifth Framework Programme. This project produced a series of tools to be freely available to members and a series of papers on different fields of industrial and business statistics that are being gathered into a single volume. More can be done and it would be great if, in future, an initiative about the spread of statistical knowledge through freely, certified, information could be achieved and ENBIS could definitely play a role in it.

In the meanwhile we are working with our specialised workshops where leading statistical experts are illustrating, for free or a low nominal fee, the tricks of the profession at our annual meeting, in Wroclaw (Poland) on 18-20 September. ENBIS will also start another activity about training and certification but you will learn more about it in the near future.

ENBIS
European Network for Business and Industrial Statistics
www.enbis.org

Statistics for industrial engineers:

a personal reflection of their education

Elisabeth Viles insists that engineers must learn statistics. Part of an engineer's job is to analyse varying data. But many engineers become aware of this only when they start to work professionally. So what statistics should engineers study?

The use of statistics in business and in industry continues to expand. This includes applications to new product design and development, and optimisation and control of manufacturing and services processes. But there seems to be a communication gap between engineers and statisticians. This gap may be because statisticians and engineers have different views of the world; they think differently. Engineers tend to keep things simple, to take action and focus on the bottom line, whereas statisticians depend on the scientific method and the desire to be thorough. Statistical education of these engineers should reduce this gap.

Since the Eighties, several authors have described the need to revise the statistics education in university, particularly in the United States. Some engineering schools have modified their statistics courses to make them more practical and useful. But little has changed over the past 25 years, at least in Spain.

Two questionnaires applied to industries compared the topics taught in Spanish engineering schools with those used in industry.

Information from industry

The two surveys focused on the industrial use of specific statistical techniques. One survey was in northern Spain's Basque country, whose industry is recognised throughout Europe for its quality and prestige (Figure 1). TECNUN is in the Basque country.

The other survey was included in a European Project (EURobust), which investigated the use of robust design methodology (RDM) in companies from five European countries: the Netherlands, Germany, Ireland, Spain and Sweden (EURobust 2003).

In our study, we focused solely on the use of statistical tools and industry techniques to get information about the real use of these techniques within the majority of companies. Answers came mainly from manufacturing companies, and the respondent was in most cases the quality manager. From the analysis of this research, we deduce:

- The surveys revealed that regular use of the tools investigated didn't reach 50 per cent in most cases. (Figures 2 and 3)
- Although different tools were analysed in both surveys due to varying sources (SPC, FMEA, Capacity studies, DoE), the results of their industrial use are similar.
- Both surveys have shown that the more sophisticated the statistical techniques are (Taguchi, regression, reliability, simulation techniques…), the less they are used.

Why is this? Could it be that engineers choose not to use these tools because they don't see the need for them or because they don't know how to use them? It would be reasonable to think that the use of tools is limited by an engineer's knowledge.

Information from universities

I collected data, in November 2005, about the content of statistics and quality courses at Spanish engineering schools from their web pages.

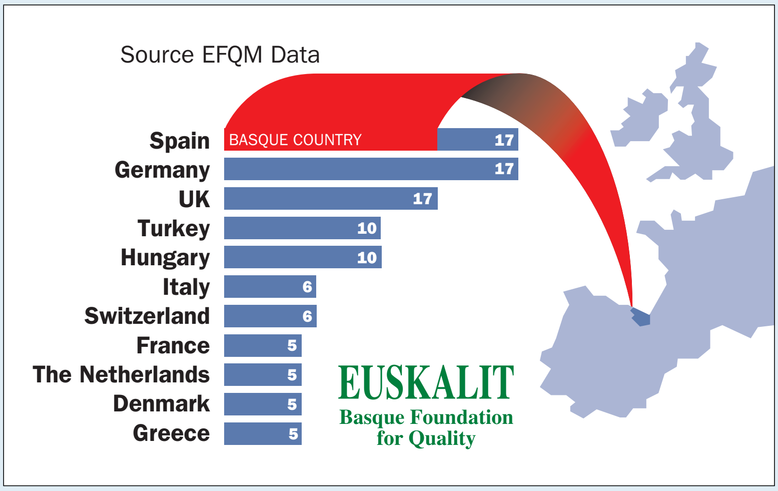

Figure 1. European Quality Awards 2000-2005 (Finalists, Prizes, Awards).

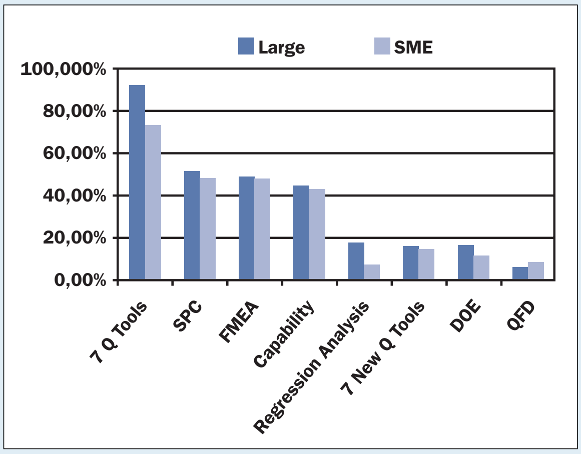

Figure 2. Regular use of Quality Tools in industry in The Basque Country, Spain (Source: Ricondo, 2005).

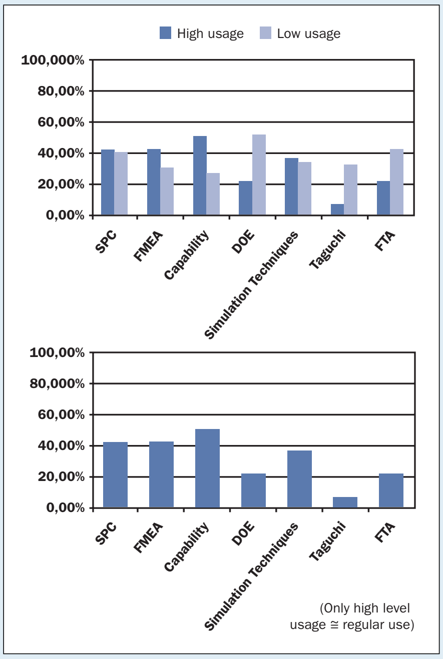

Figure 3. **Use of statistical tools in European industry (Source: EURobust, 2003).**

In Spain, educational law calls for a minimum of 60 hours of statistics at industrial engineering (IE) schools. There are no minimum requirements for industrial quality-related subjects. From this point onwards, each school decides its own strategy to best develop and train competent engineers.

What do these schools include in their statistics courses? Figure 4 shows the percentage of Spanish industrial engineering schools, which include each specific topic in their statistics courses. In this case, this graph reflects the subject matter of 70 per cent of engineering schools.

Figure 5 shows the percentage of IE schools that train their students in statistical tools for process and product improvement.

In summary, the data showed that:

- About 45 per cent of engineering students have only between 60 and 75 compulsory hours of statistics and quality theory.
- For these students, subject matter is traditional, more common in mathematics courses than engineering courses (such as probability, random variables, parametric estimation).
- Overall, there is no mention of statistical thinking nor the use of statistics for problem solving.
- In general, engineering schools do not promote the use of statistical tools for the design of QFD (quality function deployment), DFMEA (design failure modes and effects analysis), nor for the design of experiments. Generally, these topics are included only at schools with more developed statistics programs, so they reach fewer than 20 per cent of engineering students.

It is impossible to establish a cause and effect relationship between the data registered from the industry and the information collected from the curriculum of IE schools. Oddly enough, however, the measurement of knowledge of a tool goes hand-in-hand with its industrial usage.

Interesting statistical skills for engineers

From a general educational viewpoint, it is clear what maths, physics and chemistry skills engineers need whereas the opposite seems true when in comes to concepts of statistics. I firmly believe that engineers not only need to learn concepts of statistics and techniques, but also need to be skilled in implementing them. They should know the theory as well the practical applications of statistical techniques. But engineers must be effective problem solvers, so they must understand the realities of statistical practice and be able to extend and develop statistical methodology.

To achieve this, as statistics teachers, we should work harder at developing the concepts of uncertainty and physical variation and the use of scientific method within the statistical education of engineers. In short, we need to introduce the statistical thinking approach as a problem-solving methodology. This will allow engineers to identify and understand the variability of the process, to learn how to collect appropriate data for a specific purpose, to recognise limitations in existing data, to use basic and more complex tools in order to analyse data, to understand the limitations of statistical analysis and to infer conclusions and improvement actions from data analysis.

As Hoerl and Snee say in their book, the main goal of statistical education for engineers should be that 'the students develop the attitude that statistical thinking and methods can help them do a better job…' (Statistical Thinking: Improving business performance, Hoerl and Snee. Duxbury, 2002).

Elisabeth Viles is associate professor
of TECNUN Engineering School,
University of Navarra

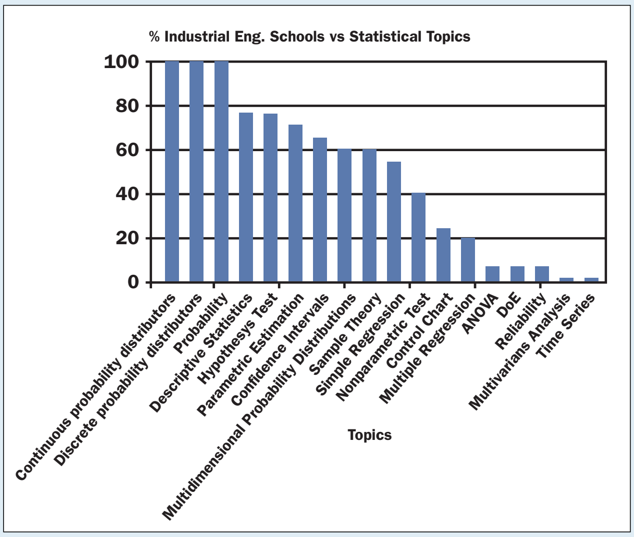

Figure 4. **List of subject matter of industrial engineering statistics courses.**

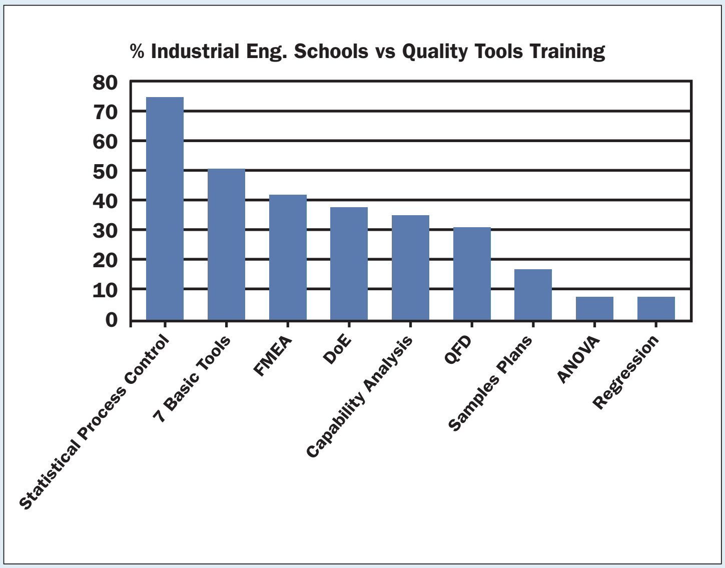

Figure 5. **Quality Tools Training in Industrial Engineering Schools.**

STATISTICS UNRAVELLED

Include the science if you can!

John Logsdon issues some health warnings for your analyses

Creep, the slow continuous deformation of metals under steady load, has long been an interest of mine. Some years ago, I was investigating high temperature, high pressure creep failure times in steels. These are important materials used in the pipes in power stations that transport steam from the boilers to the turbines. Failure would be catastrophic to anyone in the vicinity – and very costly. I devised a simple statistical model that took into account the tests that were stopped before failure.[1]

I was astonished that one of the objections raised was that a zero stress led to a predicted finite failure time when an infinite answer was expected. This was extrapolating far outside the data, but it was impossible to persuade the sceptic of the validity of the estimates within the application range and that I was not suggesting a mechanistic relationship. I learnt two things:

1. Plaster health warnings about extrapolation everywhere.

2. Try to represent the science if at all possible.

Ernest Rutherford said: 'If you need statistics, you have done the wrong experiment.' He was well known for his aphorisms so perhaps, given his great contribution to science and the state of statistics in his day, we can excuse him. Unfortunately scientists and engineers less gifted have taken his joke at face value and totally out of context.

100 years later we can tackle complex statistical issues using computational facilities unimaginable at the time the structure of the atom was discovered. Yet statistical concepts are still applied only grudgingly across a wide area of science and industry. The lack of use of statistics in manufacturing for example is a scandal, particularly in the UK. Perhaps that is why we have so little manufacturing left.

In science and engineering, we are often faced with data that are not clear. Only if we are lucky do they follow the trend they should. Sometimes we are fortunate as the data are from an experiment designed to elicit the information needed. Other times the data are side-effects or measurements made 'just in case'. We may have to use prior judgements about some parameters, assigning distributions and using Bayesian techniques.

In earlier articles, I have made much of understanding the structure of data – for example the hierarchical or cross-classified organisation or repeated measurements aspects. Experiments have been written to illustrate these points. But particularly in science, there is something else – the science itself. In an ideal world, to draw reliable conclusions from data, we need two conditions to be satisfied:

'Statistical concepts are applied only grudgingly across science and industry'

1. The fundamental statistical model must be right.

2. The fundamental science must be right.

If either of these is wrong, the conclusions are wrong, even if by accident they give the 'right' answer.

It sounds simple but there are at least three major difficulties:

1. Much of science depends on covert statistics. The embedded equations may have been derived with statistical tools wrongly used or inappropriate data. Models almost always ignore the prediction errors. The next time you code that heat transfer coefficient, don't imagine that it does anything more than interpolate some laboratory measurements. It is probably wrong. Why else does the exponent change with Reynolds number?

2. Some science is extremely complex and difficult to represent as a structure, let alone solve. The information may be qualitative, little may be known of some aspects or there may be arguments. Again, to quote Rutherford: 'All of physics is either impossible or trivial. It is impossible until you understand it, and then it becomes trivial.' He could have been talking about all of science.

3. You need to find a common point where the accepted wisdom of the scientists is a reasonable starting point. If you can work from that point on using statistical approaches, you have a chance of taking your scientific audience with you.

In some sciences, we do not have too much fundamental knowledge; social sciences and some parts of life science are relatively poorly understood by physical science standards. In which case, the correct statistical model becomes even more important if reliable conclusions are to be drawn.

In these articles, I have tried to give a flavour of statistics and to show, in some cases numerically, how different assumptions can lead to different results. As it is generally unscientific to draw different conclusions from the same data, which is right? The answer is always the one where you have constructed a model that reflects the data structure fully. But remember that in general you have only a sample of data available, not all possible values. Those data are your only concrete information and you calibrate the model to the data. You certainly don't fit the data to the model. It is the data that are right, not the model.

[1] *The best representation for ½%Cr½ %Mo¼%V steels was:*

$$\text{Mean time to failure (hours)} = \exp\{40.52\text{-}4.812\text{x}10^{-2}T\text{-}5.547\text{x}10^{-5}T\sigma\}$$

where T is temperature in °C, σ is stress in Nmm^{-2} and the errors were distributed as a gamma distribution with shape and scale parameters of 1.587.

Errata: Due to a problem during production, John Logsdon's previous article (April/May 2006) contained an error and omitted the illustrative figures. The corrected version is available online at http://www.enbis.org/newsletter/SCWapr06Enbis.pdf