From the President

I work with many statistical practitioners from business or industry, and with people from other disciplines who use statistical methods in their work. Whenever I meet them, which is often, I am encouraged by the pleasure we share in arriving at results that make positive differences to their companies. There is a pride and even a joy in working together as a team, applying our professional skills in new areas, identifying sources of variation, solving problems and communicating our findings.

But there is another emotion we share: a feeling of isolation. An industrial statistician provides a service to the mainstream of company activity: R&D; production; marketing; and management. There is usually great pressure to deliver tangible results quickly but without being at the centre, or even a part, of the department that demands the service. We must be experts in our field and keep up with the continuing growth in computer power and data-capture in industry. It is not easy to sustain this momentum, as well as keeping abreast of the subjects of our clients. Amid such galloping changes, we are strongly challenged to choose the best tool for tackling the problem presented to us.

The ENBIS network helps statisticians to meet this challenge. Our conferences and interest groups provide platforms for exchanging information between peers with similar work experience and professional needs. The internet has enabled us to create our own global marketplace in which we can seek information, discuss solutions to technical problems, and establish good practice.

But the amount and value of the information shared in this marketplace depends on its visitors. The more we use it, the better the value. By being active in an interest group, you can benefit yourself and add value to our network.

The interest groups will meet during the ENBIS conference in Copenhagen and plan future activities. We shall discuss ideas for creating an interest group for statistical software. I invite you to attend, and urge you to participate.

Poul Thyregod
President of ENBIS

Why is statistics ignored?

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

John Logsdon wonders why there are many companies that do not use statistical methods and suggests some answers

Statisticians are convinced of the benefits of statistical methods in industry. Is industry as convinced? Some companies use statistics, but many are sceptical, deny the need, or don't even consider the question. Faced with quality problems, some have turned to policies such as Taguchi, TQM, and Six Sigma, for help.

There is much to recommend these policies; they give systematic frameworks for quality improvement and persuade management that something is being done. At least they promote the idea that errors and uncertainty are present in industrial processes. Many contain good basic statistical modelling techniques, but they are more management processes than statistical processes and can lead management to believe, as Tony Greenfield described (*SCW* June 2004, p46) that uncertainty is not allowed. Clearly, statisticians are failing to convince.

Here I address the problem from the other point of view – why industry does not use statistics rather than why statistics is not used in industry. It is a subtle, yet important difference. I focus on manufacturing, which, in all countries, has its back to the wall.

Apart from ignorance, the key is in the essential accounting view of the marginal cost of labour. Table 1 shows the distribution of sizes of manufacturing companies across Europe and the United States – including mining, quarrying, and the utilities. High-tech or advanced manufacturing units with more than 500

John Logsdon is Executive Director of Quantex Research Ltd in the UK and Visiting Research Fellow at Lancaster University. He is also on ENBIS council.
j.logsdon@quantex-research.com

Country	Germany	Spain	France	Italy	UK	EU	US 1998 companies	US 1998 units
Manufacturing companies	139785	170549	261298	566319	168889	1614322	318537	366443
Units with > 500 staff	1970	358	959	560	1232	6206	4838	38572
Percentage of total per country	1.41	0.21	0.37	0.10	0.73	0.38	1.52	10.53

Table 1 Employees in manufacturing economies – EU and US EU data are drawn from an average 1995-2000 data from Eurostat; US data are from 1998 on the census.gov web site. The two US classifications represent total companies and separate establishments, where a single company may have a number of manufacturing establishments or units.

employees employ typically ten to 20 per cent of the workforce in research and development. Table 1 shows that the US has about six times as many such manufacturing units as the EU – proportionately, about 25 times as many.

Larger units may therefore have 50 or more staff in R&D; an additional person therefore represents a tiny marginal cost. Small companies with 10 to 20 employees are subject to much more variation – perhaps up to half the staff are in R&D – but even one additional person, such as a statistician, would represent a very large increase in relative marginal cost that needs to be justified.

Few statistical problems can be solved instantly. Statistical modelling can take many days, even months, of intensive work. It is not a matter of offering an opinion and walking away with a large fee, but of collecting, analysing, and interpreting data in an iterative way, working all the time with the subject experts, to produce a verifiable solution to a real problem. The statistician must be more than an occasional visitor.

The larger the company, not only is it easier to employ a specialist, but also more knowledge is contained within the company. This endogenous knowledge is part of the intellectual capital and it is much more than the sum of the parts. Over the past 20 years or so, fashion has suggested that downsizing, outsourcing, and the client-consultant model is a cost-effective way to access specialist knowledge. The much higher proportion of small manufacturing units in the EU suggests that this will be much more difficult to manage than in the US.

So here is the challenge. We statisticians must find an effective way to communicate and work with companies that are tiny in size and financial resources, yet we must still make a living. In the longer term, the industrial community must re-engineer the knowledge and innovation base so that small businesses can benefit from the methodologies capable of quantifying quality with confidence.

This is the issue that statisticians and statistical practitioners in ENBIS must confront if we are to succeed in promoting the widespread use of sound, science-driven, applied statistical methods in European business and industry. Restricting our attentions to the few large companies will not be enough.

WHAT IS STATISTICS?

The historic and popular understanding of 'statistics' is the collection, tabulation and presentation of data. Important though these functions are, statistics has expanded into a much more intellectual discipline with many theories and methods. It has become central to all sciences and an essential aid to all technologies. We need new interpretations. I offer these and invite debate:

Statistics is the science of uncertainty; a coherent framework that enables us to characterise measurements and events that we cannot otherwise understand.

Statistical analysis is the use of tools that enable us to estimate probabilities of events of interest, to make predictions. This is sometimes known as…

Statistical modelling, even when the model is very simple. It is needed to draw unbiased inferences from history or to predict the future. Statistical modelling is a fundamental part of statistics. Practical methodologies such as design of experiments and process modelling can all be derived from statistical modelling approaches leading to deeper understanding and the development of bespoke or new techniques.

How can statistics provide value to its customers?

Ron Kenett *says a workshop in Copenhagen will lead to better services to statisticians' clients*

The ENBIS Statistical Consulting Interest Group (SCIG) is unique in that it is facing, head on, the ultimate question posed in the title of this article: *How can statistics provide value to its customers?*

Our customers include CEOs, engineers, scientists, marketing managers, production managers, R&D managers, purchasing managers, HR managers, scientists, the media, the public, the justice system, the public at large, and so on.

John Shade, who founded and led the group until this year, has drafted statutes setting out the objectives of the group as 'all aspects of statistical consulting and the enhancement of statistical consultancy in business and industry'. Furthermore, 'The section may engage in specific activities appropriate to the pursuit of the objectives set forth above, including but not limited to the following:

1. Promote the consultancy role of the statistical practitioner in business and industry;
2. Provide for the regular interchange of information on statistical consulting through the internet;
3. Organise or otherwise assist with contributions to the ENBIS annual conference;
4. Provide for liaison with other Sections of ENBIS and with other societies and organisations; and
5. Promote the development and use of ethical standards in statistical consulting.'

Our slot in ENBIS2004 is Monday, 20 September from 14:00 to 16:00 in 'Room 1'. Following talks by Jonathan Smyth-Renshaw on sustainable improvement; a long awaited workshop on Tom Lang's Puzzle (audience participation in the communication of experimental results) by Antje Christensen; and Minimax Estimation (minimisation of maximum risk) by Nahid Sanjari Farsipour; we shall have an opportunity to discuss future directions for the group.

Ron Kenett is leader of the Statistical Consulting Interest Group and is CEO and senior partner of KPA Ltd.

Predicting profitability for a book club

***Ilkka Karanta** and **Jussi Ahola** show that classifying customers can save on marketing effort for a book club*

Uudet Kirjat is a major Finnish book club. It is a part of SanomaWSOY, the leading media group in the Nordic region.

A book club is a marketing and logistics organisation. A core question, in terms of printing and mailing expenses, is targeting the marketing efforts. When a new customer is acquired, not much is known about them. Finnish law prohibits combining of databases, such as taxation information and census registry data. The situation changes when a customer has been a member for some time. The information given by the customer on joining the club, and purchase data up to the moment of analysis, are available. This enables the club to target customers most likely to purchase.

Although profitability is a continuous variable, the book club wanted to classify customers into five profitability groups to plan its marketing. This changes a prediction problem to a classification problem. The club also wanted the implementation to run automatically in its information system.

This case study was part of a larger project aimed at using data mining and statistical methods in marketing support, and the resources allocated to the study were relatively scarce. Thus, there was no time for large comparison studies or experimentation.

An attack plan

Methods for classification include neural networks formalisms, discriminant analysis, logistic regression, and decision trees.

Statisticians may apply their personal experience and judgment to choose a method. A rigorous approach would be to compare the results from different methods when each is used against the same data set, and then select the method that best meets some predefined criteria. A semiformal compromise would be to use the methods of decision analysis, systematically applying judgment and facts to arrive at a choice.

We adopted a very simple strategy of multi-objective optimisation, familiar from product comparisons in consumer and popular technology magazines. Criteria are formed, weights are assigned to each criterion, each product is evaluated on each criterion, the results are scaled to a single scale (in our case, from one to nine), and finally the overall score of a statistical method is the weighted sum of its rates on each criterion.

In our view, there is a general set of criteria suitable for most statistical method selection problems; from these, the criteria to be used in a given problem can be derived. The following criteria were used:

- Time factors: the time to develop an adequate model within the model class, the time to compute a solution (if the solution goes to interactive use), and time taken by the validation of results. Development time consists of the time needed to study the method in order to use it properly, the time possibly taken by program development, and the time needed to formulate a satisfactory model;
- Monetary cost: the cost of purchasing programs, and of using external expertise;
- Quality of results: classification accuracy as measured from the development dataset, and generalisability of results, as measured by prediction power in the test dataset. Robustness is also an issue; and
- Acceptability: statistical and theoretical soundness, applicability of the presumptions of the model to the given situation, and intelligibility of the method from a decision-maker's point of view.

Selecting the method

Several considerations reduced the criteria. First, computation time needed in modelling was not an issue, because all the methods produce results within a reasonable time frame. Second, monetary costs were negligible, because it was decided that the solution would be based either on software we already had, on free software, or on programs we would construct. The final set of objectives is shown in figure 1.

We used the analytic hierarchy process to arrive at a set of weights for the criteria. It is a reliable workhorse of decision analysis; we had experience with it; and a program to implement it was readily available.

Figures 2 and 3 show the weights for the different criteria and the relative ratings of the methods.

Logistic regression won the contest, being the best in time factors: it is easy to learn; available programs support it well; and it is relatively easy to implement.

Forming the solution

Inputs include personal information, such as gender and post code, all of which are categorical data. Variables derived from transaction data should also be included in the inputs. Due to their dynamic nature, transactions aren't usable for logistic regression as such. They have to be transformed to static variables. Thus, for each customer, the number of different purchases (or returns) is counted from the transactions of her first half-year as a member. This is done

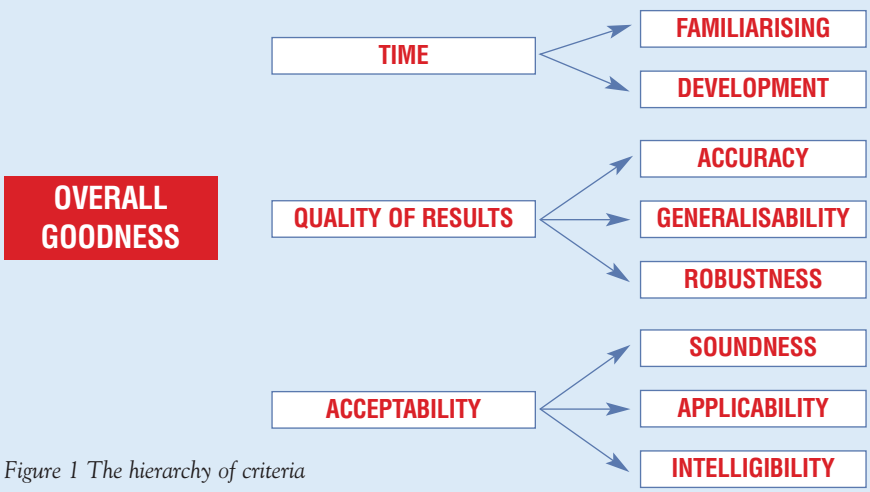

Figure 1 The hierarchy of criteria

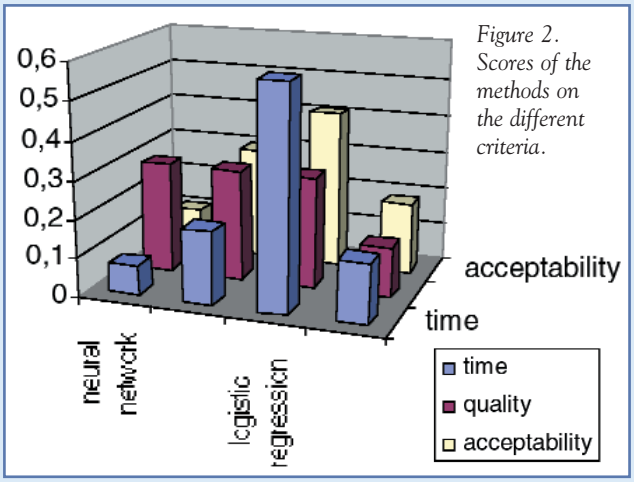

Figure 2. Scores of the methods on the different criteria.

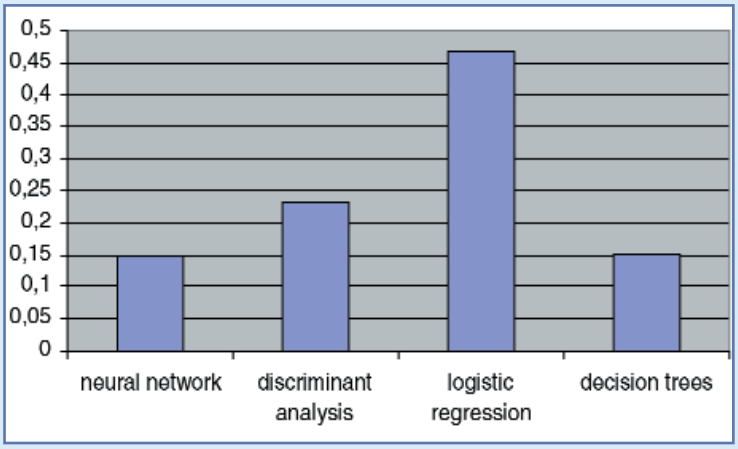

Figure 3. Final weights for the alternative statistical methods.

for each product type, products in different price groups, different sales channels and transaction types involved etc. These count variables can naturally be treated as continuous.

The output of the model is each customer's profitability group, computed from the transactions data. In multinomial logistic regression, five profitability groups translate to four models; the fifth group, 'new customers', is used as the base group.

Initially, more than 700 input variables were considered. To reduce the number, a straightforward but effective strategy was used: compute the significance levels of each variable in the full model; drop out the least significant variable and re-compute the significance levels. This is continued until all the variables are significant. This well-known method is easy to automate, and rather rapidly leads to a usable model. The multinomial regression tool of SPSS was used.

Results

Figure 4 shows how many customers were (correctly or incorrectly) classified to each class.

It seems likely that the model could be improved. However, the book club is satisfied with the results, which are a clear improvement over their previous practice. The book club is especially gratified by the fact that even the misclassified customers mostly fall in the adjacent categories. An implementation of the model has been installed at the book club, where it automatically predicts the profitability group of a customer.

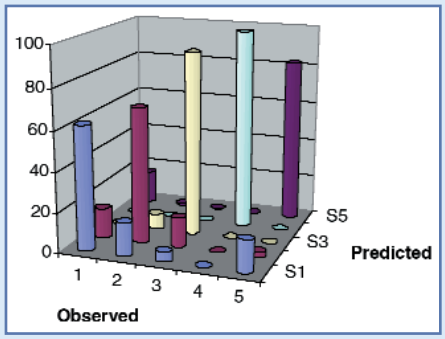

Figure 4. Observed vs. predicted profitability classes for the training set using the final model.

The input variables in the final model are of some interest. It is no surprise that purchases of 'books of the month' or regular-priced books are good predictors of a customer's profitability. Some of the less evident findings are that men are better customers than women; that people who purchase bargain-priced books are better customers than those who don't; and that purchasers of art, calendars and children's and teens' books are less profitable than average. No geographic area (as judged from the post code) turned out to harbour significantly good customers on the average; on the other hand, several post codes, most of them in cities with more than 100,000 inhabitants, turned out to have a population that make significantly less than medium-profit customers. We leave considerations of the meaning of these findings to the reader, and possible measures to the marketing experts of the book club.

Ilkka Karanta and Jussi Ahola are at VTT Information Technology, Finland
ilkka.karanta@vtt.fi, and jussi.ahola@vtt.fi

Six Sigma in Copenhagen

A course on Six Sigma will follow the ENBIS conference in Copenhagen. The course will last for two and a half days from 22 to 24 September. It will be levelled at managers, engineers and senior staff of companies.

Six Sigma is a management policy, a systematic framework for quality improvement characterised by its emphasis on data and by its focus on financial results. It was originally developed by Motorola in 1987, and adopted by multinationals such as American Express, Boeing, Citibank, Dow, Ford, and General Electric. They all claimed to have made billion dollar savings from the programme. Six Sigma has been adopted by many companies in Europe including DAF Trucks, Nokia, and Philips.

Six Sigma projects are led by 'Black Belts' who are recruited from middle management and are trained in statistical techniques for problem solving. Typically a Black Belt training has four modules, of four days each, spread over four months. The main topic of this training is an efficient problem-solving methodology called DMAIC: Define, Measure, Analyse, Improve, Control.

The teachers are experienced consultants in international industry and are well grounded in statistical science. They have implemented Six Sigma at, among others, Achmea Pensions, General Electric Plastics, Getronics, Perlos, Red Cross Hospital, and Samsung.

You can register for the course through the ENBIS website (www.enbis.org) where you can also find information about the conference.