DEVELOPMENTS IN OFFICIAL STATISTICS AND CHALLENGES FOR STATISTICAL EDUCATION

Dr Paul Cheung, Chief Statistician, Singapore

INTRODUCTION

The development of statistics can be traced over two parallel paths. The first path, which can be traced from the work of Pascal and Bernoulli in the 17th century, is regarded widely as the academic side of statistics. However, the second path, which is the development of official statistics, can be traced back to the census in biblical times.

The importance and the need for comprehensive and timely information about the state provided the subject with its name. The word "statistics" or "state arithmetic" was coined by Achenwall, a German professor in 1749. "State arithmetic" refers to the counting and calculating activities necessary to the operation of a modern nation-state. The term was adopted by John Sinclair in his preface to a comprehensive survey of Scotland in 1791. He wrote that he was interested in:

> "... the quantum of happiness enjoyed by the inhabitants and the means of its improvement: yet as I thought a new word (*statistics)* might attract more public attention, I resolved to adopt it."

The concerns of Achenwall and Sinclair are the concerns of official statistics, as we understand it today. Official statistics provide the basis for the quantitative assessment of the "state of the society and the economy". The stated objectives of national statistics offices, including the Singapore Department of Statistics (DOS) are to provide accurate, relevant and timely statistics to facilitate informed debate and discussion, policy monitoring, evaluation and formulation.

ACADEMIC AND OFFICIAL STATISTICS

Up to the final decades of the 19th century, the two paths of academic statistics and official statistics developed in parallel with little contact between them. Academic statistics, which was almost entirely mathematics, was rigorous and demanding. Official statisticians who laboured to collect data were generally untrained in the mathematical aspects of statistics. Mathematical tools were often rejected as happened for surveys.

The major economic crisis experienced in Europe between 1875 and 1895 brought about the need for a wide range of new statistics: number of paid workers, consumer price indices of workers, family budgets, unemployment and strike days. Driven by economic and social forces, official statisticians in England, France, Germany and the United States set out to organize and record new statistical data.

The establishment in 1885 of the International Statistical Institute (ISI) provided a forum for official statisticians to discuss their problems among themselves and with academic statisticians. By this time, official statisticians were no longer defined only by administrative skills but also, and increasingly, by their use of technical tools, the three most important of which were "representative enumeration" (the forerunners to sampling), "statistical machines" (the ancestor of computers) and regression and correlation (the ancestors of mathematical statistics).

The first two of these innovations are directly related to the work of official statisticians, but the third has its origins from a different world, that of biometrics and heredity. Official statisticians were innovate in their use of survey techniques and the mechanical processing of information, but were not particularly concerned with the development of analytical methods. These were initiated by Galton and Pearson in their study of heredity in the world of biometrics, and further developed by Fisher, Gosset (known more by his pseudonym, Student) and Neyman. These ideas were introduced to official statisticians by social scientists like Bowley, Yule and March through their presentations to the ISI, which became a forum for scientific debate. The debates on data recording and classification techniques took place instead at the International Labour Office and the League of Nations (subsequently the United Nations).

The result of these developments meant that the two paths remained separate. Mathematical statistics became more and more formalized, and taught in specialized university departments. The construction of statistical "data", which posed tremendous problems of definition, nomenclature, organization of data, and harmonization of varied sources, remains the major domain of official statisticians. They demand intense work and know-how that is more difficult to formalize and teach than mathematical statistics.

The divergence of these two paths is unfortunate, and has given rise to some commentators describing the two groups of statisticians as "blue" and "white" collar statisticians. The white collar statisticians are the academic or mathematical statisticians working in a university or research institute, whereas the blue collar statisticians are those

labouring to compile official statistics to inform society and policy makers. Their activities are very different; with blue collar statisticians collecting data and producing information in response to societal needs, and white collar statisticians engaged in academic and theoretical research with little concern for messy, real world problems.

And yet, it is important for these two groups to meet again. A natural bridge exists between them: both groups had to come to terms with uncertainty in their work. The blue collar statistician knows that there is uncertainty, i.e. errors in his estimates and labours to reduce these errors. On the other hand, white collar statisticians provide the theory and the methodological basis for estimating and reducing these errors. The innovative analytical tools developed by academic statisticians are necessary to improve the quality of official statistics, particularly with the growing demand for more and better statistics. In turn, the problems encountered by official statisticians would provide fertile areas of research for academic statisticians.

RECENT TRENDS

As we look forward to the next millenium, it is both timely and necessary to review some recent developments in official statistics. As suggested, these developments provide new and fertile areas of research for academic statisticians. In addition, they pose substantial challenges for statistical education.

Just as the development of new statistical indicators in the 19th century were driven by economic and societal changes, recent developments in official statistics are driven by major economic and societal forces, in particularly, rapid technological developments and the trend towards globalization. These developments have brought about an increasingly complex and more knowledge-intensive economy and society, increasing both the need for more information as well as the difficulties of developing appropriate statistical measures.

The shifting of the economy from tangible activities (within manufacturing and agriculture) to intangible service-based activities has, for example, made it extremely difficult for official statisticians to measure economic output arising from these activities. The proliferation of new activities, more as well as better quality products has implications not only for the development of statistical classifications, but also on the accuracy (or relevance) of price indicators, and consequently measures of real output and productivity.

Besides the increasing complexity of the “new” economy, the official statistician has to develop a vast range of new statistical indicators to contend with globalization, which has led to the rapid growth of international trade and sharp volatility in capital flows. Financial innovations and financial derivatives have provided tremendous challenges in the development of appropriate concepts and methodology, which are as demanding as those posed by problems in mathematical statistics.

MEETING THE CHALLENGES

These challenges are extremely difficult to meet, and it is here that official statisticians will require the assistance of academic statisticians. With limited resources at their disposal, official statisticians could not continue hope to meet the demand for more information simply by collecting more data, i.e. to launch more and more surveys to collect data from businessess and households. Indeed, even if sufficient resources could be found to conduct the additional surveys, official statisticians would have to contend with the need to avoid too large an increase in the burden imposed by the additional surveys on respondents or information providers.

This means that surveys would have to be efficiently designed and conducted. More efficient sample designs will ensure not only better use of scarce resources, but also that fewer establishments and households need to be surveyed. The smaller sample size would also enable better quality checks. The result is more timely and more accurate data with lower respondent burden. Sample designs should be reviewed periodically to ensure that sample sizes do not grow unnecessarily. With this in mind, DOS has recently completed a review of its sampling design for economic surveys. The review has resulted in a halving of the effective sample size without loss of precision.

Simply reducing sample sizes is not sufficient. The administrative procedures and survey forms would have to be similarly reviewed to ensure that those respondents who remained in the sample are not placed under unduly heavy burden. The further reduction of burden can be achieved through a re-design of the administrative procedures for the conduct of the survey. DOS has also undertaken such a re-design in its economic surveys, reducing the previous eight-page survey form to a four-page form with information normally available in financial accounts excluded from the survey form. The extraction of these information directly from the financial accounts has enabled a significant reduction in the burden imposed on our survey respondents.

But as indicated above, with the need for more and better statistics to be made available on a timely basis, there are challenges not only in the collection of data, but also in obtaining reliable estimates. This is the forte of academic statisticians. In addition to the traditional approach of obtaining estimates using simple empirical-based approaches based on sample statistics (which official statisticians traditionally excel), there is the additional need to consider more innovative complex model-based approaches. These model-based approaches are especially needed when the information available to official statisticians are incomplete or insufficient. The appropriate use of such models for imputation and the derivation of reasonable estimates will enable the provision of good statistics on a timely basis.

Because economic statistics are compiled within an integrated framework with the input-output tables providing the basis for the framework, the application of more sophisticated techniques to identify and reconcile conflicting estimates has tremendous potential. These reconciled estimates could then be verified against the original data. Such data confrontation (or reconciliation) exercises are increasingly being undertaken within DOS, as they are in the national statistics offices of the more statistically developed countries such as Australia, Canada and the United States.

The rapid advances and developments of information technology (IT) may have made the economy more difficult to measure. But IT has also made possible new approaches for data collection, compilation, analysis and dissemination. The exploitation of IT for these activities is particularly important in the constraint of limited manpower. For this reason, it is crucial to explore and try out new arrangements for the collection of data and the transmission of information through electronic means. DOS has recently pioneered the Electronic Transmission of Returns (ETR), which has made it possible for companies to submit their survey returns through the information highway of the Internet.

As data are the basic raw material for national statistics offices, the development and maintenance of large databases is critical. Recent advances in data warehousing have opened up new possibilities in the integration and exploitation of these databases. More powerful computers and data mining programs have also make it easy for the application of modern approaches to data exploration and analysis through data visualisation and data mining. These are possibilities, when exploited, will enhance the ability of national statistics offices to provide additional insights into the state of the society and the

economy. DOS is actively looking into data warehousing and data mining to derive as much information and insights as possible from our vast data resources.

With regard to the dissemination of statistics, the new IT tools have increased the speed and extent at which statistical products can be disseminated. For example, most national statistics offices have established Internet web sites on which are posted recent economic and social statistics. DOS is no exception, and I am happy to note the very high hit rates achieved by our web site. In addition, DOS has recently launched a Data Shop through which value-added statistical products can be disseminated to the world at large.

CHALLENGES FOR STATISTICAL EDUCATION

The recent developments in official statistics posed several challenges for statistical education. What are some of these challenges? Will the white collar statisticians who provide statistical education to future blue collar statisticians rise to the challenge?

First and foremost, it is evident that our students must be equipped with a sound knowledge of sampling and basic statistical theory. But this is not sufficient. An understanding of basic statistical theory, the usual diet for most university statistics students, is useful but would not be adequate for our students.

Second, it is crucial for us to stress the importance of analytical skills by which I do not mean the ability to do simple analysis, or the ability to perform the standard "cookbook" statistical procedures. What our students need to do is to develop "data sense", i.e. the ability to make sense of data and to recognize "patterns" in the data. Since the recognition of patterns is central to the development of models, the development of data sense will enable our students to exploit the more advanced statistical concepts and models which are increasingly being implemented in data mining programs and applications. Students with such skills will be well-equipped to get the most out of incomplete data through the application of appropriate estimation and imputation techniques. These techniques will have to be exploited to get the most out of a mixture of administrative and survey data.

Third, it is important to realize that much of future statistical practice will be driven by developments in IT. Students will need to develop a whole range of IT-related skills which are not restricted to computational skills or the simple "crunching" of numbers. Students will need to appreciate the notions of a database, and develop the

ability to manage and integrate databases. In addition, they would have to develop the ability to exploit IT for the presentation and dissemination of their analysis and findings. IT-related skills will become increasingly a critical core competency for the statistical profession.

Fourth, it is worth asking if the difficult conceptual problems and issues, for example, those relating to index numbers theory and methodology and statistical classification, should not be included within the curriculum of university statistics courses. Inclusion of these into the curriculum will expose students to the practical problems and concerns that are likely to be encountered in practical statistical work, and narrow the gap between academic statistics and official statistics.

CONCLUSION

Would statistical education rise to these challenges? Are they met by our schools and universities? Would our students and graduates be adequately prepared? Is there a need for curricula to be reviewed to ensure that students graduate with sufficient statistical competence to produce good statistics, or the necessary competence to properly interpret and use statistics?

The cliche "*There are lies, lies and statistics!"* is not true. Statistics does NOT lie, but those who produce and use statistics can and do mislead if they are not adequately prepared. Their adequate preparation is the job and the challenge of statistical education! Statistical education, and its priests, the white collar statisticians, must rise to the challenge. If they succeed, we can be assured of more and better official statistics produced by competent and well-trained blue collar statisticians to inform society and guide public policies.