New Zealand Statistical Association (Inc.)

newsletter

ISSN 0112-2649

February 1985

Number 5

DATA PROCESSING AND STATISTICS: SOME INTERACTIONS

by Murray Jorgensen

Modern data management technology impacts on the statistician in two ways: firstly it opens new possibilities and greater flexibility for a statistician to manage his or her data, making feasible the analysis of much larger data sets. Secondly this new capability gives the statistician an interest in the design of information systems that are not in the first instance statistical, in order that these systems may be constructed in such a way as to permit the extraction of data appropriate for statistical analysis.

A Database Management System (DBMS) is a computer software product under the control of which data is stored, organised into related files, and retrieved. To use a modern DBMS it is not necessary to know the physical details of how the data is actually stored, but only to have a conceptual picture of the logical relationships among the data items. Commonly databases are organised hierarchically. As hierarchies are common in Statistics as well it is useful to give some examples.

(i) A national crop variety evaluation might have data available at the following levels:
region, site, trial, block, plot, sub-plot.

At the region level weather information might be available while soil-type, farmer's name, etc. might be available at the site level.

(ii) A survey might collect data at the levels of:
city, mesh block, household, person

(iii) In fisheries data collection at least 3 levels can be identified. Firstly we have data about the sampling "station" such as latitude, longitude and depth. Secondly data pertaining to the occasion of sampling such as date, time, gear used, name of responsible person. At a third level we have information on the composition of the catch. If this is rather complex the catch level may itself be decomposed into more than one level.

In a modern database, at each level, a different type of record is used to store the information pertaining to that level. The fields that make up a record are divided into **key** fields and **data** fields. The key fields contain sufficient information to uniquely identify a record. For instance in example (ii) above the key of a record of the "household" record type will contain sufficient information to uniquely identify the household referred to.

A DBMS will provide a means of storing information about the fields of a record type. This facility may be called a "schema" or a "data dictionary". In it we store such information as the data type (real, integer, character) of each field or variable, variable names, value labels, variable ranges, and so on.

But statistical packages in general expect their data input files to be in the form of a simple rectangular matrix of observations by variables. When a database is interfaced with a statistical package a retrieval program is written in the language of the DBMS to build such a "flat" file. Information must be communicated to the statistical package at a particular "level" of the hierarchy. If this level is the lowest, such as the sub-plot level in the first example, information must be "brought down" from higher levels by chasing up the tree. This information will then be written out many times at the lowest individual level. If we wish to move information to a statistical package at a higher level then the lower-level information must be summarised in some way: by summing, taking means, medians, counts, variances, or percentage points of the lower-level data. The richness and flexibility of the retrieval language will determine exactly what can be done. It is at this point that many commercially-oriented DBMSs fall down: they are often limited to such simple summaries as means and totals. Facilities for handling missing values are also absent from many DBMSs.

It may seem almost frivolous to talk about the transfer of labels and variable names to statistical packages but experience has shown that this facility is very important. P-STAT solves the problem by being a statistical package as well as a DBMS. SIR (Scientific Information Retrieval) has an interface which permits the creation of SPSS, SAS and BMDP system files. If a DBMS language is sufficiently expressive it is possible to develop procedures which create files containing the data to be transferred and expressions in the language of the statistical package which control the reading in and labelling of the data. The Ministry of Agriculture and Fisheries uses SIR procedures to build input files for GENSTAT and MINITAB in this way.

The other alternative is to do things the hard way and write out a pure data file from the DBMS and write a statistical package program to input, label and analyse the data. Errors and inconsistencies can often creep in when this is done.

Whatever approach is adopted the use of statistical packages in partnership with DBMSs provides a very powerful tool for studying, analysing and reporting on large sets of data. In Hext (1983) seven data management systems are described and evaluated for their usefulness in statistical applications.

A knowledge of what can be done with statistical packages leads the statistician on to wonder what could or should be done statistically with the very large amounts of data held in commercial dp systems. Most

Published by the New Zealand Statistical Association (Inc.), P.O. Box 1731, Wellington, New Zealand. Printed by Bayleys Allprint Services Ltd., Wellington.

people have only a vague idea of the difference between scientific and commercial computing along the lines of "they use COBOL, we use FORTRAN". Increasingly language is coming to seem less important, each package having its own internal language owing something to a variety of predecessors. I believe that a better distinction between the two types of computing is to say that scientific computing involves intensive processing of small, static, datasets. (It may even involve no data at all and be totally theoretical in nature). In contrast commercial computing involves the processing of large and constantly changing datasets. The processing itself is usually of a fairly simple nature.

Many statisticians see themselves as Data Analysts and the modern statistician has a range of powerful computer tools at his or her disposal to undertake this role. The question I wish to raise is: can we bring our analytical tools to bear on data within commercial information systems? Is there any need to do this? I would say certainly! Many businesses have captured within their computer systems a reflected picture of the firm and its changing performance in the different sectors of its market. Statistical analysis of data extracted from corporate information systems could provide valuable information to managers about the performance of their institution. The interfacing of databases with statistical packages discussed above provides one means of doing such extraction of data.

There are, however, problems connected with the dynamic nature of commercial information systems. There are limitations on the amount of historical information that can be held in disk storage and extraction retrievals will usually produce a "snapshot" of the data when what is required is a time series. Yet a statistician is normally only interested in a small fraction of the information in such a system. Furthermore advances in the theory and practice of the analysis of categorical data means that the data storage requirements can be compressed substantially. What a statistician might best be able to work with is an automatically generated "state description" which tracks the evolution of the system through time. This would usually be available only if a statistician were involved in the design of the information system.

I have indicated in the first part of my article how modern data management technology can assist the statistician in taking on large sets of data for analysis. I also feel that the statistician has a role in the field of Management Information. However I feel that the statistician will never be fully effective in this area unless he or she has an input into the design of management information systems. Both the statistician and the computer scientist recognise an area called "Data Analysis". At the moment each means something different by the term. But it is possible that both areas are part of a larger whole and could be profitably knit together. I hope that this happens, not only in an academic sense, but in the working together of all who are responsible for the collecting, processing and interpreting of data.

REFERENCE

Hext, G.R. (1983) *Database systems for statistical applications.* (Civil Service College Handbook 24). HMSO, London.

THE SEVENTH AUSTRALIAN STATISTICAL CONFERENCE

by Ken Russell

The 7th (biennial) Australian Statistical Conference was held at the University of Queensland from 27th to 31st August 1984. The list of participants records full-time attendance by 250 people, of whom ten were from New Zealand. Another 50 people attended for one day only.

The featured sessions at the Conference dealt with Medical Statistics, Analysis of Sample Survey Data, Geophysical Signal Processing, Estimation using Capture-Recapture and Related Methods, Modelling of Biological Systems, Inference, Multivariate Analysis, Discrete Multivariate Analysis, Design and Analysis of Experiments, and The Role of Statistics in Public Policy. The invited speakers were George Seber (Auckland; "Recent Developments in Population Estimation"), Richard Smith (Imperial College; "Statistical Problems of Extreme Values"), T.W. Anderson (Stanford; "Components of Variance in MANOVA"), Robin Sibson (Bath; "Smooth Variation Across the Plane : Analysis and Presentation"), Shelby Haberman (Hebrew University of Jerusalem; "Canonical Analysis and Maximum Likelihood"), and Richard Cormack (St Andrews; "Loglinear Models for Capture-Recapture").

The Conference began with the Presidential Address of John Darroch (Flinders). This was a most entertaining and illuminative lecture entitled "Probability and Criminal Trials—Some Comments on Bayes' Theorem Prompted by the Splatt Trial and Royal Commission". The accused in this celebrated South Australian case had been convicted of murder on the basis of circumstantial evidence and some "probabilistic" arguments by the prosecution. This had been followed by petitions for his release, a Royal Commission, and suggestions that a policeman may have planted some evidence. (I am sure that the New Zealanders in the audience shared my feeling of *deja vu* as parts of this case were detailed.)

The Presidential Address set the stage for what was a most stimulating conference. Scheduled talks were held from 9 a.m. to 5 p.m. each day except for Wednesday and Friday, when they finished at noon. In general, there were three or four parallel sessions at any one time, except when the Invited Addresses were being presented. I felt that the general presentation of papers was better than at earlier Conferences. There were fewer speakers who couldn't be heard, and distinctly fewer transparencies which couldn't be read—but the exceptions still remain. When will some speakers learn that photocopies of typed material are almost never legible, even from the front row?

Pleasant memories remain of the Brisbane Conference. The University campus was attractive, especially with its riverside location, and the Lunch Cruise on the Wednesday afternoon was a definite highlight. The Conference organisation was good, the morning and afternoon teas lavish, the Welcoming Party and Conference Dinner very pleasant—and the weather was superb! As well, some of us were fortunate enough to attend a barbecue held at the home of Tony Swain, one of the Conference organisers.

Less memorable were some of the bathrooms at Union College and the disappointing organisation of the poster display: there were only three posters, and these