COMPSTAT 1982

part 2:
short communications
summaries of posters

COMPSTAT 1982

5th Symposium held at Toulouse 1982

Part II (Supplement): Short Communications Summaries of Posters

Edited by
H. Caussinus
P. Ettinger
J.R. Mathieu

Physica-Verlag · Wien 1982

CIP-Kurztitelaufnahme der Deutschen Bibliothek

COMPSTAT:
COMPSTAT . proceedings in computational statistics ,
symposium – Wien Physica-Verlag
5. Held at Toulouse 1982.
Pt. 2 Short communications, summaries of posters. –
1982
ISBN 3-7051-0001-7

Printed in Germany by repro-druck „Journalfranz" Arnulf Liebing GmbH + Co., Würzburg

ISBN 3 7051 0001 7

PREFACE 原书模糊

Le présent volume contient les résumés des affiches et communications courtes présentées au congrès de COMPSTAT-82 (Toulouse, 30 Août-3 Septembre). On remarquera que les sujets abordés sont très divers.

Certains travaux concernent la définition et l'étude de modèles spécifiques de statistique appliquée tant exploratoire que décisionnelle.

D'autres articles présentent des logiciels permettant la mise en oeuvre de méthodes statistiques ; en ce domaine il est intéressant de noter qu'à côté de logiciels permettant le traitement de grands fichiers, de nombreux statisticiens, suite au développement récent des micro ordinateurs, présentent des logiciels adaptés à ce type de matériel : ils donnent ainsi une nouvelle impulsion à l'utilisation des méthodes statistiques et apportent une contribution importante à leur enseignement.

Enfin, dans d'autres communications, sont développées des applications de la statistique à divers domaines scientifiques (sciences sociales, archéologie, biométrie,...).

Ces divers thèmes témoignent des préoccupations actuelles des statisticiens, de la complémentarité et de l'enrichissement mutuel de leurs recherches. Aussi, sommes nous heureux qu'ils puissent s'exprimer ici.

This book contains the summaries of the short communications and posters presented to the COMPSTAT-82 Conference held in Toulouse (August 30 - September 3). You will notice that the matter dealt with here are of varied kinds.

Some of these works are concerned with the definition and study of specific models of applied statistics, either exploratory or decisional.

In other papers the authors present statistical softwares. In these latter works, it is interesting to note that besides packages to be used in analysing large data bases many statisticians, considering the new developments in micro computing, have worked out packages adapted to micro computers. This may give a new impulse to the use of statistics and brings new incentive to their teaching.

Finally, in other papers you will find applications of statistics to various other scientific fields (social sciences, archaeology, biometry, ...).

These different items, which are some of to-day's statisticians' concerns, are complementary aspects of statistics and therefore any development in one of them contributes to the others. This is why we feel particularly satisfied to have them all here presented.

Toulouse, Juin 1982

The editors.

Exploratory Analysis of Counted Data Using Log-Linear Models

M. A. Adena, Canberra

Counts are one of the most common forms of data, and their analysis has been revolutionised by the recent development of log-linear models (Bishop, Fienberg and Holland, 1975), especially when the counts are summarised in contingency tables.

In a log-linear model, the natural logarithm of the expected value for each count is modelled with a linear model in the explanatory variables, and the counts themselves are assumed to have a Poisson error, with mean given by the exponentiated linear model. In the simple case of two cross classificatory variables, the log-linear model with only additive main effects corresponds to multiplicative main effects in the scale of the counts, that is, to statistical independence of the two cross-classificatory variables. Likelihood ratio tests can be used to evaluate the goodness of fit of models and to distinguish between competing models in a hierarchy. Log-linear models with a Poisson error are analogous to linear models with a Normal error - the most important parametric models for continuous data.

However, many analyses of contingency tables make only limited use of this similarity, and restrict their attention to log-linear models corresponding to simple multi-factor analysis of variance, ignoring the full range of log-linear models corresponding to regression, analysis of covariance and the general linear model. Similarly, the concepts underlying random effects analysis of variance are often useful in practice, even though the distribution theory is still incomplete.

One common problem with log-linear models encountered in practice is that they often have a great many parameters. This means that such a model may be very good at conforming to the particular quirks of a data set, so masking possibly important features from the data analyst. In addition, though not a difficulty with the iterative proportional fitting algorithm, estimation of many parameters may cause computational problems of speed, storage and accuracy with the iterative reweighted least squares method (used, for example, in GLIM).

Interactions on several degrees of freedom can often be simplified and understood better after decomposition into several specific contrasts, each representing some interpretable effect. Contrasts can be constructed using sub-factors (created by combining several levels of a factor) and linear, or other components of ordinal factors. It may even be worthwhile attempting to simplify the experimental constraints (margins of the contingency table usually assumed fixed because of the design of the data collection).

Sometimes seemingly complex interactions can be resolved by identifying and removing outliers, that is cells with unusually high or low counts. In some situations, outliers can be paired and high and low counts balanced against each other. For some analyses, the most important outcome may be the identification and explanation of outliers.

For flexible modelling of counted data, the full range of log-linear models should be available simply through suitable software. Such software should also allow easy analysis of sub-tables (including the exclusion of particular cells), estimation of the parameters of the fitted models, and effective display of residuals.

These considerations are illustrated in a reanalysis of a recent contingency table (D'Alessio *et al.*, 1981). In particular, the different conclusions that may be drawn depending on the treatment of certain cells emphasises the importance of screening counted data for unusual values, as is routinely already done for continuous data.

This reanalysis used the widely available computer package GLIM (Baker and Nelder, 1978). GLIM is a general purpose linear models package for interactive modelling of data where the error distribution is from the exponential family. Hence it exploits the conceptual links between linear models with a Normal error and log-linear models with a Poisson error, and indeed, these analyses differ by only a single directive. The full range of log-linear models is simply available, as is the selection of which counts to model and the display of parameter estimates. However, although residuals are easily printed sequentially or plotted, tabular output does require additional programming with this package.

Baker R.J. and Nelder J.A. (1978), The GLIM system Release 3. Generalised linear interactive modelling, NAG, Oxford.
Bishop Y.M.M., Fienberg S.E. and Holland P.W. (1975), Discrete multivariate analysis, MIT Press, Cambridge, MA.
D'Alessio D.J., Minor T.E., Allen C.I., Tsiatis A.A. and Nelson D.E. (1981), A study of the proportions of swimmers among well controls and children with enterovirus-like illness shedding or not shedding an enterovirus, Am.J.Epidemiol., 113, 531-541.

Prelude-composing in ALGOL 68: Applications in Datamanipulation and User Languages

H. J. Adèr and G. C. van der Veer, Amsterdam

Programming languages like Ada or ALGOL 68 offer the possibility to define one's own datatypes as well as the operators and routines that act upon them. This provides a way to create a notational system which can be used in a specific problem-area.
We will outline what concepts are offered in this respect in ALGOL 68 (A user-extension in this language is called a "Prelude"). Then two examples are presented in more detail. Finally, we will comment upon the construction of this kind of notational systems.

An Algol 68 Prelude usually includes definitions of useful datatypes. For instance, in a notational system for the manipulation of pieces of data a (well-defined) datatype "matrix" or "variable" could be convenient. When writing a program with the Prelude as a library, the user can declare and use datatypes of the prescribed structure. Often, in the Prelude some handy declarations are already made for the convenience of the user.
Obviously, the meaning of these new datatypes is highly determined by the associated operators and routines. These also are predeclared in the Prelude. In his program, the user handles his datatypes with the aid of these operators. In this way, construction of a Prelude is very similar to the construction of a programming language (or a user interface, for that matter).
Now, let us consider some examples of ALGOL 68 Preludes. From literature the matrix/vector-system TORRIX is well-known. It supplies a set of operations on vectors and matrices over a (more or less) arbitrary field.
Our own Preludes, PREDAT and LIBMULT, differ in objectives:

- PREDAT is meant to provide the programmer with a tool to facilitate information handling on file. He has to describe the structure of his data, like defining it to be a correlation matrix. Datastructures may be combined to form larger configurations. A rather elaborated indexing-system is present, to access elements or substructures of the given datastructure. In the poster an example of the construction and accessing of a hierarchical file is shown.
  The PREDAT-user should know some simple ALGOL 68. He can use PREDAT efficiently if he knows more about the language.

The second example might be more interesting to the computational statistician:

- LIBMULT offers a userinterface for Finn's program MULTIVARIANCE.
  The output is a setup that has to be input to this program to get the desired

Multivariate Analysis.
Here the predeclared operators are used as keywords. For instance, "between" is used to denote the names of between-subject-factors, "nested" to indicate the nesting structure. In this case the user isn't expected to have any special knowledge about ALGOL 68: he should know the keywords which are represented by ALGOL 68 operators.
The poster gives an example of a straightforward Multivariate Analysis of Variance-design with two Within-subject-factors.

From the point of view of language-design ALGOL 68 Preludes offer the possibility to do it the easy way (f.i. LIBMULT took a month or so to program). The structure of the underlying language is very well-defined. So the designer is forced to formulate a very precise definition of his new dialect (a practice which, especially in userlanguage-design is not too often encountered). In this way the technique might be used to check the merits of a new language.
Of course, there are disadvantages too, like the impossibility to have the user enter his program interactively.

As far as practical programming is concerned, "composing" a Prelude is a special art: it has nothing to do with writing a "normal" program: it much more ressembles writing that part of a compiler that embodies the translation. The parsing part has been handled implicitly by the ALGOL 68 compiler.

References:

-C.H. Lindsey and S.G. van der Meulen. Informal introduction to ALGOL 68; Revised edition, 1977 North-Holland Publishing Company.
-J.D. Ichbiah et al., Rationale for the Design of the Ada programming Language, SIGPLAN Notices, ACM, 1979, Vol 14, Numb. 6, June 1979, Part B.
-J.D. Finn, MULTIVARIANCE, Univariate and Multivariate Analysis of Variance, Covariance, Regression and Repeated Measures, User's Guide, Version VI, Release 2, October 1978, National Educational Resources, Inc.
-C.H.A. Koster, User Languages and Application Languages, March 1979, Informatica/ Computer Grafics, Faculty of Science, Nijmegen University, The Netherlands.
-S.G. van der Meulen, M. Veldhorst, TORRIX, A programming system for operations on vectors and matrices over arbitary fields and of variable size, 1979, Mathematical Centre Tracts 86, Mathematical Centre, Amsterdam.

Logit Models for the Analysis of a Very Large Survey of Unemployment in France

M. Aitkin, Lancaster

This paper reports the analysis of data on unemployment in France from the EEC Labour Force Survey of 1979. The data were supplied by Eurostat, and were analysed at the University of Lancaster, in conjunction with the Centre d' Analyse Statistique des Structures et des Flux at the University of Paris X.

The object of the analysis was to examine the relationship of unemployment rate to age, sex, industrial sector and region. Data were available from a 1 in 300 sample survey in the form of a four-way contingency table defined by the cross-classification by age (10 5-year classes), sex (2 classes), industrial sector (11 classes) and region (8 or 22 classes).

Within each cell of the table, the number of unemployed persons r and the total working population n were recorded. Unemployed people who had not previously worked in an industry were treated separately, and were not included in this analysis.

We index the table by i(1-10), j(1,2), k(1-11) and l(1-8 or 1-22), and let p_{ijkl} be the population proportion of unemployed people in cell ijkl. We examine first a probability model of the form

$$\theta_{ijkl} = \log p_{ijkl}/(1-p_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l$$

This is a binomial logit model, which includes main effects of age, sex, industrial sector and region. This model can be fitted to the table by maximum likelihood (ML), which is equivalent to minimum cross-entropy, using the statistical package GLIM (Baker and Nelder 1978). The ML estimates of the parameters are found by an iteratively reweighted least squares algorithm, which also provides large-sample standard errors of the estimates, and an overall measure of goodness-of-fit of the model to the data, the deviance. If the fitted model adequately represents the data to within binomial sampling variation, the deviance has in large samples a χ^2 distribution, with degrees of freedom given by the difference between the number of non-empty cells in the classification and the number of parameters in the model.

For the 8-region classification, the four-way table has 1760 cells, but 64 of these are empty. For the 22-region classification, the four-way table has 4840 cells, but 406 of these are empty. The main effects model, when fitted to the 8-region table, gave a deviance of 1823 with 1668 degrees of freedom (28 parameters in the model). This model gives an inadequate fit to the data, and large interactions of sex with the other variables were found, so separate main effects models were fitted for each sex, equivalent to adding to the main effects model all the interactions of the other variables with sex. The resulting deviance of 1700 on 1642 degrees of freedom now shows a good fit of the interaction model.

Models of unemployment of this kind have two advantages. First, they identify immediately major subgroups of the population with high unemployment rates - for example, men aged 20-24, women aged 14-24, men in region 6, and men and women in

industries 5 and 6. Second, they allow the estimation of accurate unemployment rates for individual cells of the table where the sample size is very small. For example, the "fitted" logit for women aged 20-24 in region 6 and industry 6 is -1.141, with a corresponding unemployment rate of 24.2%. In this cell the observed unemployment rate is 41.2%, based on a sample of only 17.

REFERENCE

Baker, R.J. and Nelder, J.A. (1978). General Linear Interactive Modelling, Release 3. Numerical Algorithms Group, Oxford.

er Specifications for the Data Management in a Distributed Data ıalysis System

Alfano, L. Colazzo, A. Fambri and A. Silvestri, Trento
Zito, C. Dell'Aquila, E. Lefons, M. T. Pazienza and F. Tangorra, Bari

his paper describes the user facilities for the data management in a distributed da a analysis system: POLD. POLD is a part of the DATANET Task,in the applied research roject "Informatica" of the C.N.R.(National Research Council). POLD is a distributed lata analysis system that will run on a heterogeneous micro and minicomputer network. The system has been designed in order to provide the local user with all the facilities of a scientific data oriented DBMS.

1. Data organization and indexing facilities
2. Data selection and sub-sample creation
3. Linking facilities to user routines and catalogued query procedures
4. Data presentation in various graphic and alphanumeric mode.

The most peculiar features of the system can be recognized in the remote user operation mode.

Most of the problems in the non-homogeneous computer network arises when different da ta base organization are present in the co-operating nodes. We have solved this problem by using a very simple data structure.

The data record is an ordered couple <informative part,indexing part>. Both are in relational form (CODD): the indexing part is a bit-string containing values of simple conditions(tests). No other data structures are present in the system: so,all the nodes have information stored in the same form.

The global user has access to a catalogue in which all the relevant information on global data are stored. The remote query is quite similar to local query: the only difference is in the file qualifications needed: the name of the node and the mode of the data base is requested.

As the transmission of data is to be realized by Hermes(DATANET) system(a distributed filing system in project by the DATANET/HERMES Group) the POLD system has been designed in order to reduce all remote operation to standard file operation. The problem of indexing data ranging in continuous intervals and many other related problems (definition of join and selection conditions, mono and bidimensional plotting) have been solved by the session concept. The used data base schema can be modified by any user at any time. We call this a session; the user is requested to define his own subschema: to define all the tests that he will use in the session execution. For example, before any execution of a join operation, conditions on the data attributes keeping continuous values must be precisely defined, giving a set of boolean variables.(Equality of continuous values must be defined in terms of discrete intervals).

As a result of the definition session,for each entity, a bit string is created containing all the relevant information for subsequent research. (For example, in geographical data base, that belonging to a definite region is defined at session time). At execution time, if possible, only the indexing (bit string) part of data is processed. A relevant concept is that of "multiple test". A multiple test is a boolean vector. A generalized histogram is defined as the procedure of adding to a vector accumulator some calculated quantity in a predefined form from the data stored in the informative part of the record. In the vector accumulator the channel is selected corresponding to the first (left-to-right) one bit, in the multiple test representation.

COMPSTAT 1982 © Physica-Verlag, Vienna for IASC (International Association for Statistical Computing), 1982

Multidimensional plotting is obtained in the same way, from the Cartesian product of multiple tests. The join is defined in a very general form: the user defines, in a general way, a boolean function B of the chained string $S = S_1.S_2$ (S_1 and S_2 being the bit strings corresponding to the index parts of records of the files 1 and 2, respectively). The join is executed on the pairs of elements corresponding to "true" values of B(S).

At present, the local system is running on several different computer systems:PDP 11/34, IBM Series 1, Selenia GP 160, HP 21 and VAX 11/750-780. The memory requirement is less than 64 kb without any segmentation. Due to the structure of the program, the segmentation can reduce drastically the memory occupation. Despite the physical sequential organization the response time is very small, due to a fast scanning algorithm(ARMENISE), based on a tree valued logic(CAPASSO), in comparison with the performance of other methods for range search, based on B-trees and similar indexing techniques (KNUTH).

The performances are strongly dependent on the characteristics of the mass storage devices . With fast direct access units, a performance up to 10^3-10^4 scanned elements/second can be obtained. Finally, we give a brief list of the principals user operations:

- User features
- Session: definition; frame definition(node, data base, file,session)
- For each session : Creation of the schema
  - " " test (simple condition)
  - " " multiple test
  - " " index file (string)
  - " " selection or join condition(boolean function)
  - " " arithmetic function
  - " " user defined procedures

  Report and graphics definition
  Data file creation.
- Session: execution; frame definition(node, data base, file, session)
- For each session : String creation
  Relational operation : Selection, join, Projection
  Mono and bidimensional plotting
  Report generation
  Graphic output generation
  Graphic software execution(if the node has graphic facilities)
  Execution user defined procedures.

References

|1| Armenise N.,Silvestri A., at Al.(1979), POL:an interactive system to analyze large data sets, Computer Physics Commun. 16, North-Holland Pu.Co., 147-157.

|2| Capasso V.,Circella A, and Silvestri A.(1971), Una logica a tre valori per il calcolo del valore di verità di funzioni booleane complesse, Rapp.C.S.A.T.A.,Bari.

|3| Codd E.F.(1970), A relational model of data for large shared data banks, CACM 13, 377-387.

|4| DATANET(1981), Rapporto del gennaio '81, Sistema di gestione di DDB. Specifiche funzionali; versione 1.

|5| Knuth D.E.(1981), The art of computer programming...., "Sorting and searching" (Vol. III), Reading, Mass., Addison Wesley.

Flexibility in Statistical Software

A. J. B. Anderson, Aberdeen

Introduction.

In recent years, statistical software has become increasingly used by those whose contact with computers is only sporadic and who may, in addition, be statistically naive. In the social sciences especially, much current research activity lacks integrity because merit is gained from any successful computer interaction irrespective of the validity of the analysis. Software designers must therefore accept some responsibility to ensure that their products are sufficiently flexible to encourage greater appreciation of the fundamentals of statistical inference.

Preprocessor-driven software.

Anderson (1980) has discussed in general terms the benefits of hosting statistical software in some high-level language such as Fortran. Some more detailed considerations are here presented on specific aspects of this approach in relation to "user-friendliness".

(a) It is relatively easy to provide alternative syntactical forms of specification to satisfy differing user aptitudes and perceptual processes. Such flexibility could be attempted via full compilation into machine code but the availability of a well-established high-level target makes development much more secure. The naive user can, of course, remain entirely innocent of the embeddedness of the system until he is able to cope with more procedural aspects of analysis. Thus, the ability to name or number variables will allow him to develop from

```
SCORE1=0
SCORE2=0        to      DO 10 I=1,8     or even   SCORE(I)=0 ,I=1,8
   :                10 SCORE(I)=0
   :
SCORE8=0
```

(b) Because of the high-level target, we can minimise irritating restrictions due to compiler/generator limitations such as "Not more than 50 constants". Such restrictions are usually unimportant in practice but are psychologically unsettling. More serious are run-time constraints resulting from the rigidity of the implementation configuration, these

are also absent in the preprocessing approach. The more such restrictions can be removed, the more the user can concentrate on statistical essentials.

(c) It is clear that, for many analyses, some parts are more appropriate to batch working and some are better accomplished conversationally. In particular, the summarisation of large volumes of data in tables or sums of squares and products matrices is a batch exercise prior to conversational "modelling". Even for the batch run, the user's control statements should be interactively syntax-checked (but seldom are); the preprocessing phase is an ideal point at which to do this. Furthermore, a short simulated pseudo-analysis may follow to confirm that the output is roughly what was anticipated; this may also help to establish time requirements for execution. By these means, the user can have some confidence that the batch run will be successfully completed.

(d) Ready availability of computer packages has encouraged the presentation of "amateur" statistical analysis in research publications with consequent degradation of quality. Even when there is contact with a statistician, the collaboration often starts part-way through the analysis phase when the favourite package has revealed its limitations. The statistician may then be forced to switch to a more sophisticated but inconsistent system to achieve an adequate analysis. What is required, therefore, is a software environment appropriate to the skills of both the statistical adviser and his client i.e. a framework where analysis can be flexibly shared using a single system equally "friendly" to each.

Fortran-hosted software seems an ideal mechanism for satisfying such needs. For a simple analysis such as might be attempted by SPSS, the user need know nothing of Fortran; all can be achieved by the use of concise extended-language statements. At the next level of complexity, only a little advice from a colleague familiar with Fortran will permit the introduction of a wide range of "tricks". And for the practising statistician, there remains the advantage of being able to incorporate almost any sophisticated novelty either by ad hoc programming or by subprogram calls to other libraries.

Conclusion.

Experience with preprocessor-driven software in Aberdeen has confirmed the usefulness of the approach in satisfying the flexibility requirements listed above and development of the system will continue.

Reference.

Anderson, A.J.B. (1980), The use of preprocessors in statistical software, COMPSTAT 1980, Physika-Verlag, Vienna.

The Use of Payroll Files for Manpower Studies

B. I. Anderson, Aberdeen

Summary. The advantages and disadvantages of using payroll files for manpower studies are discussed, emphasising that such a source can provide data for a wide range of manpower analyses.

KEY-WORDS: Payroll file, Manpower, Cohort, Wastage, Prediction.

Introduction.

In recent years many large firms and government agencies have become increasingly aware of the necessity for manpower analysis to monitor staff changes over time, to plan for expansion or contraction of their workforce, and to anticipate likely future staffing problems. Often the need for manpower studies is urgent, precipitated by either a worsening of inherent staffing problems or the sudden necessity to increase/decrease the workforce over a short period. Time as well as other resources of money or personnel may be at a premium.

Where computerised systems designed specifically for manpower analyses have not been instituted, problems may arise in assembling appropriate data. Various options can be considered - using existing manual records would not be feasible if the workforce is large or the records incomplete; conducting a special survey would be time consuming and usually of limited value; setting up a computerised personnel information system, though perhaps the ultimate aim, requires very careful initial planning and data preparation as well as inter-disciplinary cooperation resulting in considerable delay before data are available. An obvious, though frequently overlooked source of data worthy of consideration is the payroll system.

Data from a payroll file.

The payroll file, almost certainly computerised, offers considerable advantages as a source of manpower data:-

(i) Traditionally the payroll file is related to the current financial year and includes a record of every employee currently in post as well as records for those who left during the financial year.

(ii) Each record contains basic variables essential in manpower studies e.g. date of birth, date of commencement, sex, grade, location.

(iii) From these data appropriate derived variables can be produced such as age, length of service, whole-time-equivalent, staff group.

(iv) Even if the payroll system spans several separate files, to accommodate say distinct staff locations or alternative pay frequencies, a single composite file, containing only manpower variables, can be abstracted fairly easily.

(v) The payroll system ensures regular updating of records, at least once per

pay period.

(vi) Using the payroll file relieves the investigator of the need to collect, prepare and maintain data.

The possible disadvantages of using data from the payroll file should not be overlooked, as such a file has not necessarily been created nor maintained for manpower analysis. Consequently some difficulties may be encountered - the data may be limited and probably not include sufficient personal information to permit analysis of careers, absenteeism, etc.; over time, the file may have developed historical or geographical inconsistencies in definition of some items but these can usually be satisfactorily resolved by a 'once-off' computer validation process; a few missing values and errors may exist amongst some data not directly related to pay but by reference to personnel records such can usually be traced and corrected; a compound file including records of different pay frequencies (e.g.weekly,monthly) may not be updated uniformly but this can be accommodated by a suitable choice of retrospective reference point for analysis.

Statistical analyses.

Data obtained directly from the payroll file are adequate for production of cross-sectional analyses, on a regular basis if required, to provide descriptive tables and graphs, of particular use to managers e.g. tabulation of sex by grade by location, or plot of age distribution of staff. By the end of a financial year the payroll file can provide sufficient data on staff stocks and flows in a year to permit further census analysis of wastage, prediction of future staff composition etc.

If a sequence of end-of-year payroll files can be accessed retrospectively, together with an item of record linkage such as National Insurance number, then data files of selected cohorts of staff can be compiled. This would allow more in-depth study of length of service, by considering, for example, models of wastage (e.g. two-term mixed exponential, lognormal) and investigation of likely future consequences of alternative planning strategies, as described by Bartholomew and Forbes (1979).

All such statistical analyses should of course be tempered by the managers' own experience and knowledge of the workforce.

Conclusions.

With relatively little extra outlay in resources, but essential inter - disciplinary cooperation, the payroll file can provide both census and cohort data permitting a wide range of manpower analyses of particular value to managers and planners.

Reference.

Bartholomew, D.J. and Forbes, A.F. (1979), Statistical Techniques for Manpower Planning, Wiley, Chichester.