The impact of computer analysis on the design of dietary surveys

By ELAINE C. TAYLOR,* *Social Paediatric Research Group, Department of Child Health, University of Glasgow and Health and Welfare Department, Corporation of Glasgow*

Introduction

Data-processing by digital computer is now an accepted part of most nutrition and dietary surveys. When this new tool was introduced a decade ago it was commonly regarded by research workers as a more efficient and rapid method of performing tedious and time-consuming calculations than the desk calculator. A consequent increase in the rate of publication of survey findings was expected.

These high hopes diminished as it was realized that the programmes or software necessary for data analysis were not available, that developing them took an unbelievably long time and that computer processing demanded re-assessment of methods of collection and recording data. In addition, if the specification of the required computations was not very precise the resulting output was massive and difficult to interpret.

Programmes are now available for most routine tabulations and calculations but difficulties and delay will continue until it is generally recognized that preparation for computer analysis is a vital and integral part of the structure of surveys.

General survey planning

Few surveys on man are completed without unexpected difficulties arising. A firm definition of the aims of a study and thorough prospective planning are made even more imperative by the use of the computer.

Many surveys are beset with difficulties because the time required for planning is grossly underestimated. Decisions which have to be taken and implemented include those on a feasible method of sampling, the type of equipment necessary, the tabulations and calculations required, the most efficient method of analysis, and the content and design of the record sheet or source document. A pilot study to detect ambiguities, omissions and difficulties in organization and in collection, analysis and interpretation of data is particularly valuable when computer analysis is being used for the first time. Minor aspects of methodology, not directly related to the analysis, are often overlooked and later give rise to intractable problems.

Additional time given to planning in the early stages of a study will not only facilitate the analysis of data, it will also result in a more critical approach to the objectives and methodology and thereby contribute towards the ease with which the field work can be accomplished. Time taken to explain to staff how the new type of analysis will affect their own area of work is well spent.

The most carefully designed surveys can be ruined by careless, uncritical, uninformed or unco-operative personnel at any level from executive to clerical assistant.

*Present address: 103 Fotheringay Road, Glasgow, S1.

The thorough training of staff and good liaison between them ensures that, for example, no changes in methodology can take place in the field without effects on the validity, reliability and subsequent processing of the data being assessed.

Source document

The design of the source document, log-book or data-recording sheet is influenced by the purpose of the study, the personnel or subjects who will complete it and the chosen method of analysis.

All irrelevant data, however interesting, increase the amount of analysis possible and should be ruthlessly excluded.

If the record sheet is to be completed, checked and prepared for analysis by the designer, less rigorous attention to details of layout and clarity of instruction is required.

If, however, more than one person will contribute to it, it is vital to make the layout of the document as logical, self-explanatory, detailed and easy to complete as possible. Additional instruction sheets to interviewer or subject are usually necessary, for example, where subjects are making a record of food intake, but these should be cut to the minimum.

A pleasing and unambiguous layout aids in the collection, checking and punching of data. Answers to questions or other data are often recorded as either numerical or alphabetical codes (Table 1). They may be inserted in boxes numbered to correspond to eight columns of a punch card. If so, there should be a good visual link between the question and the coding area for ease of completion. Listing the possible answers to a question and their codes adjacent to each other so that the appropriate one can be indicated is a good method but, although it is easily completed, it results in a long document. Unfortunately, subjects and some interviewers are deterred by a thick log book and a compromise is made necessary.

If codes are to be inserted in boxes, these should be large enough to allow for legible entries. The codes from the documents will probably be transferred on to cards or paper tape by young girls who have no interest in the data except that they are easy to decipher. Checking that data sheets or coding sheets are complete and legible in every detail before sending them for punching is a common-sense practice but sometimes neglected. Although cards are verified in the key-punching department, it is not unusual for errors to be repeated or for duplicates to be produced and it is advisable to have a system for detecting such carelessness. The supervisor of the key-punching department may be a useful ally.

The commonest type of card used for data processing has eighty vertical columns in each of which one class of data can be recorded. There are twelve positions in which holes can be punched, thus providing a possibility of twelve responses to a question. Paper tape is similarly punched and is useful for large amounts of data since it is not restricted to eighty characters.

Whether cards or tape are used, codes for entries such as 'unknown', 'not relevant' and 'other' should be decided upon when devising the coding system in consultation with the key-punch department and depending on the computer being used. The use

Table 1. *Specimen questionnaire*

		Col. No.
Survey number	O B 2	1–3
Subject number (limit 99)	1 1	4–5
Sex (M or F)	F	6
Age in years	4 5	7–8
Date of interview	Year 6 9 / Day 3 0 / Month 0 9	9–14
Is sugar taken in tea? (Ring appropriate code with felt-tipped pen)	Not relevant "–" Never 0 Always 1 Sometimes 2	22
Alternative Layout		
Is sugar taken in tea?	2	22
"–"=Not relevant; 0=Never; 1=Always; 2=Sometimes (Enter appropriate code in box)		

of the same convention throughout a document helps the interviewer who has to work quickly and unobtrusively often in the presence of the subject, the checker and the person who will have to interpret the computer tabulations.

If a second or third data card is required the survey and subject should be re-identified in a pre-determined way on each. The transition from one card to another should be indicated by these codes on the record sheet so that they are automatically punched at the beginning of the new card. This precaution helps to avoid loss of cards when they are dropped, mislaid or mixed up with cards from other studies either in the department or when they are sent for processing.

Whether cards or tape are used it is often useful to leave one or two columns free at the end of each discrete section on the record sheet to allow unanticipated material to be inserted.

The foregoing remarks have applied principally to the transfer of data recorded

https://doi.org/10.1079/PNS19700031 Published online by Cambridge University Press

on a questionnaire to the eighty column punch card where answers to each question will occupy one or more columns. In nutrition and dietary surveys socio-economic, environmental, clinical, biochemical, anthropometric and dietary-history data will usually be recorded in this form and, when the record sheet has been completed and checked, punching can be done directly from it.

Records of daily or weekly intake of foods require more preparation for input to the computer. Each food is assigned a code number corresponding to the numbering in the food table programme and these codes together with the quantities of food in oz or g (a small programme takes care of any conversion) are then punched on to cards or paper tape or typed directly on to a teleprinter and thence to the computer. Clarity of layout on a coding sheet is as important as it is in the source document.

If there is any possibility of hand or sorting machine analysis being required some aids can be incorporated at this stage. For example, if the age in weeks of a child is required, date of birth and date of interview will be sufficient for computer analysis. In the absence of a computer the insertion of the calculated age in weeks will aid sorting and tabulation against other data on the same card.

The more people there are who contribute to a source document, the more necessary it is to make it unambiguous. Similarly the more people there are who are going to be involved in checking, punching and analysing the collected data, the more important it is to see that they are fully informed and that instructions for handling the data have been worked out beforehand in consultation with the persons from different disciplines who may be contributing.

A simple outline of the project and methodology given to all staff involved would often save confusion and error.

All personnel who are involved in surveys should be familiar with the basic data-processing equipment which is available, whether hand or desk calculating machines, desk computers, card punching machines or sorting machines. It is often taken for granted that otherwise experienced staff who are new to survey work know about methods of data processing whereas they may never have used even a hand calculator or set eyes on a punch card. Familiarity with the possibilities of these basic methods makes it easier to judge whether it is worth insisting on a computer programme being obtained or written for a particular calculation or whether time would be saved by using the more conventional method.

If possible the research worker should gain some experience in writing simple programmes. He will be able to save time by producing these for himself and will be better able to understand the problems of the programmer when he requires more sophisticated software.

Choosing a method of analysis

Data from most studies are analysed by a variety of methods. The method chosen will depend on the type and volume of data collected, the complexity of calculations and the time available for completion of the project.

If, having chosen computer analysis, you find the results are slow in appearing it is often convenient to do some preliminary analysis in the laboratory and even when

the results become available certain aspects can be more easily and quickly followed up personally than by trying to find another programme.

These considerations should be taken into account when deciding whether to use cards or paper tape for storage of data. Cards are suitable for storage and the data on them can be conveniently checked if they are printed along the top of the card. The cards can be sorted by simple equipment, errors are reasonably easy to find and error-free cards can be quickly punched. Tape does not lend itself so well to checking or sorting although it is cheaper than cards.

There are two main operations underlying analysis; these are sorting and calculation. Sorting of data may be done by hand, by a sorting machine or by a computer. Calculation may be done by a desk calculator, a desk computer or a large computer.

Many surveys can be analysed satisfactorily by straight sorting and counting from which tabulations can be produced and incidences, percentages, means and simple statistical calculations easily obtained.

A desk-top computer is invaluable for statistical calculations. Formulas are fed in on programme cards and any number of calculations can be performed without repeating the formulas. Only the data have to be typed in. Some of the most powerful machines can have programmes stored in them permanently. If a desk computer is available, giving the benefit of enormous saving of time in calculations, hand sorting of the record sheets should not be scorned if the volume and type of data is suitable. This may at first seem to be a laborious method of analysis but it eliminates the errors and delay associated with key punching.

In all non-computer methods the rate at which computations are performed depends almost entirely on the person conducting the analysis and further calculations can be selected in the light of intermediate results.

Use of a computer with standard programmes

Most university computing departments, regional computing centres and manufacturers provide an array of standard programmes covering the more usual computations required. Superficially these are very attractive but it cannot be overemphasized that before the final structure of the source document is decided the exact specifications of these programmes should be studied if their use is contemplated. Any deviation from the fixed standard programme introduces sources of error and may result in weeks or months of delay in tracing and checking these.

A survey programme prepared for Atlas (Rees, 1965) illustrates the type of simple specifications which should be considered at an early stage in planning, e.g. a limit of ninety-nine is set to the characteristics recorded for each subject, only a numerical input will be accepted and the input has to be on paper tape. Minor modifications to the recording of any survey data will enable them to be used with this very useful programme. If, however, the data have been collected in an unsuitable form, perhaps with many alphabetic entries, alteration and checking of a large number of record sheets is a daunting task. The situation is even worse if the data have been punched on to cards before advice on analysis is sought.

This programme is of general use and supplies the following facilities, although all do not require to be used: (1) One, two and three dimensional contingency tables; (2) chi-squared tests and Kendall's tau tests; (3) means and standard deviations; (4) variance-covariance and correlation matrices; (5) principle components factor analysis.

If such a standard programme covers most of the analysis required it is often better to use it and do any additional computations with a calculator using data supplied by the programme rather than to have modifications made. Programmers prefer writing their own programmes to adapting standard ones and developing programmes is a lengthy process.

One example of a useful set of standard programmes is the Biomedical Computer Programs (1965) (B.M.D.), developed by the Health Sciences Computing Facility, Department of Preventive Medicine and Public Health, School of Medicine, University of California, at Los Angeles. These are available at National Engineering Laboratory (N.E.L.), East Kilbride, and the Regional Computing Centre, Edinburgh, offers use of the N.E.L. facilities as a service. Notice of 1 month is requested.

Another useful programme which basically counts, tabulates and produces simple statistics is available at Atlas Computer Laboratory, Science Research Council, Chilton, Didcot, Berkshire. All British Universities have access to this computer. This multi-variate counter programme (M.V.C.) can be of value particularly where data collection has not been planned with a programme in mind. It accepts data in a sophisticated form and its specifications are flexible.

Standard programmes are now becoming available for specialist areas including the calculation of total intake of food and nutrients from records of food consumption. These provide in effect automated calculation from food tables. The first programme of this type to become widely available was developed at Queen Elizabeth College, London (Miller, 1965). Further information about it can be obtained from the Scottish Medical Automation Centre, Edinburgh, or from Elliott Medical Automation Ltd. The Ministry of Health, the Ministry of Agriculture, Fisheries and Food and some university nutrition units have also developed their own programmes which may be available to investigators on request.

Use of a computer with special programmes written for the project

The project may be of such a type that no standard programme will fit and special programmes have to be written. This process should be reckoned in terms of years.

Programming has been defined by Nicholls (1967) as 'the totality of the attitudes involved in the solution of problems by machine'. This is a much broader definition than is normally used and includes 'defining the problem to be solved, planning the broad strategy of the solution; preparing a flow chart or outline specification of the programme; coding the programme; testing and validating the programme; documenting the programme and preparing instructions for its use; and using the programme'. It is a useful definition because it emphasizes the complexity of the task and thus the painstaking work which has to go with it. Checking a programme for errors when it has been written may take many weeks. Few people who have not

attempted to write fairly large programmes can appreciate the difficulty in locating and removing errors without disturbing other functioning parts of the programme.

Programmers are also hindered by machine breakdowns and, especially in university computing departments, overstrained resources lead to too many demands on them.

Conclusion

'Too often it is presumed that the amassing of large volumes of unselected data will allow its "digestion" by the computer, with the emergence of correlations, tabulations or hypothesis more or less at the touch of a button. Nothing could be further from the truth' (Taylor, 1967).

The use of a large computer in survey work does not at first appear to save a significant amount of time. Only when programmes are established and error-free and the collection and preparation of data are directed towards the method of analysis does the real time-saving become apparent.

Much time and thought is given to the feasibility and accuracy of sophisticated survey techniques. Unfortunately data which have been painstakingly collected may wait many months for analysis unless basic requirements of data-processing are anticipated and careful consideration given to relatively mundane aspects of survey planning.

The clearer the aims of a project and the closer the attention to the design of the data-recording sheet, the quicker will be the analysis and the smaller the output.

The password to successful computer-processing is, 'First find your programme'.

REFERENCES

Miller, D. S. (1965). *Progress in Medical Computing* p. 33. London: Elliott Medical Automation Ltd.
Nicholls, J. (1967). *Computer Weekly, London*, 5 October.
Rees, D. J. (1965). *Computer Unit Report* no. 2 (revised ed.). University of Edinburgh.
Taylor, T. R. (1967). *The Principles of Medical Computing*. Oxford and Edinburgh: Blackwell Scientific Publications.

Statistics and computers—a worked example

By R. THOMPSON, *Agricultural Research Council Unit of Statistics, Edinburgh*

Computation plays a large part in applied statistical work and it is no wonder that the advent of high-speed computers has had an impact on its development. Yates (1966), indeed, believes that computers have started a second revolution in statistics. Computers have undoubtedly provided the impetus for research into new or previously unsolved problems in methodology but they can also provide speedier and more complete analysis using techniques known, if not used, in the desk calculator era.

I will illustrate various aspects of the latter use with data from two experiments (1 and 2) on the effect of stocking density on live-weight gain and feed conversion