

Output

The frequency tables formed in the 803 can be coded on tape for subsequent output, or directly printed. The output program provides many calculation facilities, including grouping and weighting of rows and columns; it is, therefore, often convenient to have more analysis columns in a table than it is expected will be required, and to group them at the output stage. The program includes facilities for page layout and descriptive row and column titles of tables.

The principal output facilities provided are as follows:

- (a) Percentage or actual frequencies may be printed.
- (b) Rows and columns may be printed in any order, grouped, weighted, or omitted altogether.
- (c) Rows and columns may be interchanged.

Conclusion

The programs described in this paper have been developed during the last six months; some of the complex editing facilities of the accumulation program are not yet complete. The programs have successfully catered for the analysis of several surveys, although in one case a modification was made to the output program to print special results. The programs have been written to provide general-purpose facilities, with the result that they are somewhat slower than tailor-made programs written for a particular survey. Our experience clearly indicates that the cost of the additional machine time is more than offset by the programming effort that would be required to write special-purpose programs, except possibly in the case of very large surveys which are in continuous use.

Market Research Applications on LEO

By J. A. Gosden

This paper was presented at the Harrogate Conference of The British Computer Society on 5 July 1960.

Introduction

The Market Research work carried out by LEO Computers Ltd. has, to a very large extent, been carried out for the Market Research section of their parent company, J. Lyons and Company Ltd., who are engaged in marketing a wide range of consumer products. Three basic applications have been carried out: first an application producing concomitants; second, multiple-regression analysis; and third, an application producing cross-tabulations of answers to questionnaires. The first two are straightforward uses of standard techniques, but the third is a special program designed to meet the needs of the particular types of survey carried out.

Concomitants

The concomitants application takes the answers to questionnaires and seeks out the dominant attitudes of consumers with respect to their eating habits, as affected by such factors as the food itself or the environment in which it is consumed.

The questionnaires may be completed by interviewers in the field, or by observation and interview during experiments in tasting laboratories. The number of questions answered varies between 20 and 45, and the sample sizes vary between 100 and 1,000.

In this application the answer to any question can only be made from a choice of two. For example, sex is either male or female. Other questions are framed so that there are only two possible answers; for example,

“Are you over or under 35 years of age?” Wherever possible, questions are framed so that the number of replies of each type is nearly equal.

The procedure to compute the dominant attitudes is a series of steps. In the first step the answers are counted and the correlation between each pair of answers is calculated. In the second step the sets of answers of a small part of the sample are searched for the most frequent patterns of inter-correlated answers. From these totals it is possible to reduce the range of answers which need to be considered in the last step. The last step assembles the correlations between the remaining more highly inter-correlated answers into a matrix, and then extracts the principal latent vectors. These vectors define the dominant attitudes of the sample. An alternative procedure, when there are few questions, is to take the most frequent patterns discovered in step 2 and count their occurrences in the full sample, providing a profile analysis of the sample.

Regression

The second application is the straightforward calculation of multiple regression coefficients. This program is used to monitor the relationships between sales of various commodities and variables which influence sales. It is in two parts: the first calculates the regression constants from actual sales, and the second uses these coefficients to estimate future sales. These assist in the compiling of production plans both for day-to-day and long-term

projects. The data for this work, unlike the other applications, is not taken from field surveys but extracted from invoices in the company's files, and represents a one-hundred-per-cent sample.

Field Surveys

The third application arose in connection with the extensive field surveys carried out by the Market Research section over samples of between one and two thousand householders. The questionnaires contain answers to between twenty and fifty questions, and the labour of analysis is quite large. Such surveys were begun in 1953 and the analysis carried out on punched cards. The main drawbacks to this were, first, the long lapse in time between completion of any survey and the completion of the basic analysis; second, the labour of percentageing the basic analysis; and third, the inability to weight items. A LEO I program was therefore prepared to do this analysis. It has been sufficiently general to be still in use after six years, having had only minor changes made to its structure. Although this program has been used mainly for field surveys, its facilities have permitted its use for quite varied applications.

The surveys are concerned with gathering and tabulating information about the knowledge, behaviour, habits, or beliefs of different groups of the population, for a selected commodity or commodities.

The information is gathered from a scientifically chosen sample of the population, on schedules filled in by interviewers on the basis of verbal answers to questions. The questions are of two types: some are intended to fix the general circumstances of the person and his or her household; others probe in a detailed way into the relationship between the person and the commodity that is the subject of the survey.

Among the information included under the first head will always be the area in which the person lives. There will generally also be included the person's sex, age and marital status, the number of people in the household, and the income group in which it falls. There may be up to 29 such questions. Then there are up to eight questions for each of the various items that the household has consumed. The more specialized questions will themselves often be of two sorts, some aiming to obtain a broad picture of preference and buying habits, and others testing reaction to a particular sales campaign.

Two programs are used to analyse the answers. The first is an editing and transcribing program which reads the answers from paper tape, checks them for consistency, and records them on to binary punched cards. These cards carry check totals to permit easy and rapid verification of the data when used in the second program. The

second program reads the binary punched cards and produces cross-tabulations of answers to all pairs of questions specified.

Each cross-tabulation requested is first formed by counting the various answers to the pair of questions. For example, the two questions might concern sex and smoking, and the answers from a sample of 4,000 might be tabulated as follows:

	MALE	FEMALE	ROW TOTAL
Smokers	1,397	1,731	3,128
Non-Smokers	360	512	872
Column total	1,757	2,243	4,000

To facilitate the interpretation of the statistics, most of the figures given are actually expressed in terms of per 1,000 of a suitable base. This base may be the column total, the row total, or the whole sample. Where the base is the whole sample, the column totals, too, are expressed not only as substantive figures but also as per 1,000 of the whole sample.

In some tabulations all answers do not have equal significance. For instance, in some cases it is appropriate to count items, but in others the questions may relate to different qualities of a commodity. This difference in quality is allowed for by assigning one of two types of weighting to answers in a tabulation. In one case the first weighting could be used to turn the sample into the value of purchase, or the other used to turn the sample into the purchasing power of consumers.

Where a survey is carried out for more than one group of commodities, cross-tabulations may be required using items of each group in turn. Provision is made in the program for twelve such commodity groups, and the data (in punched-card form) is sorted into these commodity groups. While several cross-tabulations of the same questions are produced for each commodity group in turn, totals cards are carried forward for grand-total tabulations over all groups.

The cross-tabulation programs were put on to the pioneer computer LEO I in 1954, and are still carried out on that computer. LEO I has an effective arithmetic speed of approximately 600 instructions per second, and this program can produce up to five tables in each run. The time required is approximately half a minute per table, per 1,000 households or items. But more important is the improvement in timetable. On the first production run in 1954, for a major survey, the last of the data was received on a Friday afternoon, the computer work was completed before Saturday morning, and a report, complete with selected tables, was ready for management first thing Monday morning.