USE OF PROGPAMMING METHODS IN SAMPLE SURVEYS

By DES RAJ
United Nations Expert, Greece

1. During the past few years great progress has been made in the area of programming methods. These methods have been used with advantage in solving practical problems arising in many different fields. In this paper a brief discussion is given of some problems in sample surveys whose solution has been made possible by the use of linear programming.

2. The basic requirement in the area of sampling is to design the survey in such a manner that for a specified budget the accuracy of the estimates made is maximised, or that for a desired level of accuracy the cost of the survey is minimised. For this reason programming methods have proved invaluable in this field. Since the quantities to be maximised (or minimised) have been generally found to the linear in the variables involved and the constraints to which these variables are subject are also linear, actually the techniques of linear programming have been used in practice.

3. Consider the following problem arising in the National Sample Survey of India. Information is to be collected from households for economic statistics and from plots for agricultural statistics for the villages in the sample. It is desirable to select the villages for the household enquiry with probabilities proportionate to population and the villages for the agricultural enquiry with probabilities proportionate to area. If the selection is made independently for the two enquiries, we are likely to get different villages and this will add heavily to the cost of field work. However, if the two enquiries could be so integrated that the vil-

---

Ὁ Δρ. Des Raj εἶναι στατιστικὸς ἐμπειρογνώμων τῶν Ἡνωμένων Ἐθνῶν παρὰ τῇ Ἐθνικῇ Στατιστικῇ Ὑπηρεσίᾳ τῆς Ἑλλάδος. Ἔχει λάβει τὸ δίπλωμα τοῦ Master εἰς τὰ Μαθηματικὰ καὶ τὸ δίπλωμα τοῦ διδάκτορος τῆς Στατιστικῆς τοῦ Πανεπιστημίου τῆς Καλκούτας. Ἔχει διδάξει εἰς τὰ ἰνδικὰ πανεπιστήμια τῆς Luchnow καὶ Agra, εἰς τὸ Ἰνδικὸν Στατιστικὸν Ἰνστιτοῦτον τῆς Καλκούτας καὶ τὸ Ἀμερικανικὸν Πανεπιστήμιον τῆς Beirut. Ἐδημοσίευσε πλείστας ἐπιστημονικὰς ἐργασίας ἐπὶ θεμάτων στατιστικῶν καὶ οἰκονομικῶν.

lages for the two enquiries are identical or very near to each other, the cost of operations would be greatly reduced. This problem could be stated mathematically in the following way. Let a stratum consist of n villages with populations proportionate to

$$\frac{a_1}{G}, \frac{a_2}{G}, \ldots\ldots, \frac{a_n}{G}$$

and areas proportionate to

$$\frac{b_1}{G}, \frac{b_2}{G}, \ldots\ldots, \frac{b_n}{G}.$$

Let c_{ij} be the cost of journey between the ith. population village and the jth area village. Further, the probability with which the corresponding pair of villages is selected may be denoted by $\frac{x_{ij}}{G}$. The problem is then to determine x_{ij} such that

$$\sum_{j=1}^{n} x_{ij} = b_i, \qquad \sum_{i=1}^{n} x_{ij} = a_j$$

$$\sum_j a_j = \sum_i b_i = G, \qquad x_{ij} \geqslant 0$$

and

$$Z = \sum\sum c_{ij}\, x_{ij}$$

is minimised.

In this form, this is the familiar transportation problem in linear programming. Its solution by the simplex method due to Dantzig may be found in the book edited by Koopmans (1951).

4. The following problem arose in a Canadian population survey. First stage units are to be selected from within strata with probabilities proporsionate to 1950 populations. The survey is, however, to be repeated at another occasion when new population figures would be available. It is desired that, as far as possible, the new sample at the second occasion be identical with the old sample. This is so since it is very costly to make lists of second —and subsequent— stage units from selected first-stage units. One would not like to change these units every time to avoid the cost of fresh listing. The mathematical formulation of this problem would be the following. Let the 1950 populations of first stage units be proportionate to

$$\frac{a_1}{G}, \frac{a_2}{G}, \ldots\ldots, \frac{a_n}{G}$$

while the 1951 populations are proportionate to

$$\frac{b_1}{G}, \frac{b_2}{G}, \ldots\ldots, \frac{b_n}{G}$$

We shall denote by $\frac{x_{ij}}{G}$ the probability that the ith and jth first stage units are selected at the first and second occasion vespectively. The probability of geting non-identical units at the two occasions is given by

$$Z = \frac{1}{G} \Sigma \Sigma c_{ij} x_{ij}$$

where the cost matrix (c_{ij}) is given below.

Table I. Cost Matrix

unit no.	1	2	3	.....	n
1	0	1	1	.....	1
2	1	0	1	.....	1
3	1	1	0	.....	1
.					
.					
.					
n	1	1	1	.....	0

The problem is to determine x_{ij} such that

$$x_{ij} \geqslant 0, \qquad \sum_j x_{ij} = b_1, \qquad \sum_i x_{ij} = a_j, \qquad \Sigma a_j = \Sigma b_i = G$$

and Z is minimised. Mathematically this problem is equivalent to the previous one.

5. Suppose a population is divided up into two strata, each containing rural and urban primary units. We want to select two units, one from each stratum, with probabilities given in the margins of Table 3 below. The selection is to be so made that the chance of getting one urban and one rural unit is maximised. The cost matrix would, then, be the following :

Table 2. Cost Matrix

Units		Stratum 1					
		U_1	U_2	U_3	R_1	R_2	R_3
stratum 2	Y_1	0	0	0	1	1	1
	Y_2	0	0	0	1	1	1
	Y_3	0	0	0	1	1	1
	Y_4	0	0	0	1	1	1
	u_1	1	1	1	0	0	0

Mathematically formulated the problem is equivalent to those discussed before. The optimum solution, giving the probabilities with which pairs of units be selected, is given in Table 3 below.

Table 3. Optimum Solution

Units	U_1	U_2	U_3	R_1	R_2	R_3	Total
Y_1	15						15
Y_2	0	10	20	0			30
Y_3				10	0		10
Y_4					20	5	25
u_1						20	20
Total	15	10	20	10	20	25	100

6. One final result which will be presented relates to sampling without replacement with varying probabilities. Two units are to be selected from a stratum containing N units such that the probability that a specified unit U_i is selected in the sample is π_i (known). The problem is to determine π_{ij}, the probability that the pair (U_i, U_j) is selected in the sample, so that the variance of the estimate of the stratum total is minimised. If the π_i are based on x_i, which are known measures of size of the units, and if the relation between y, the character under study, and x is linear, it can be easily shown that the problem is to minimise

$$Z = \Sigma \frac{\pi_{ij}}{\pi_i \pi_j}$$

such that

$$\sum_{j=i} \pi_{ij} = \pi_i \qquad (i = 1, 2, \ldots\ldots, N)$$

and

$$\pi_{ij} \geqslant 0 .$$

Stated in this form, this is a familiar problem in linear programming and can be solved by the simplex method given in Charnes and others (1953).

REFERENCES

Charnes, A., Cooper, W.W. and Henderson, A. (1953): «An introduction to linear programming», John Wiley and Sons, New York.

Koopmans, T. C. (1951): «Activity analysis of production and allocation, John Wiley and Sons. New York.

Raj, D. (1956): On the method of overlapping maps in sample surveys, Sankhya, Vol. 17.

Raj, D. (1956): A note on the determination of optimum probabilities in sampling without replacement, Sankhya, Vol. 17.