ON THE PROBLEM OF PLANNING A MULTISTAGE SURVEY FOR MULTIPLE CORRELATED CHARACTERS

By I. M. CHAKRAVARTI

Indian Statistical Institute, Calcutta

1. Introduction

During the last 25 years there has been a rapid development of the theory of multivariate analysis and it has found ready application in biometric problems. Still the lack of a unified approach and of detailed distributions and numerical table has limited its application in other fields like sample surveys, experimental designs etc. Often a bold experimenter trying to take advantage of this theory is confused by too many techniques whose relative importance is little known and fails to choose one which will serve his purpose best. As a result, practical experimenters fight shy of the methods of multivariate analysis and application has failed to keep pace with the development of the theory. In this paper, the possibility of using analysis of dispersion in fields of sample surveys has been sought to be explored.

2. The problem

Suppose that there are p correlated characters $(x_1, x_2, ..., x_p)$. In sampling for the means of these p correlated characters a multistage survey has been decided to be used. For simplicity, a two-stage survey with equal probability of selection of sampling units (with replacement) at each stage is considered.

We consider a linear model (using vector notation)

$$\underset{\sim}{x}_{(ij)} = \mu + \underset{\sim}{\beta}_{(i)} + \underset{\sim}{\epsilon}_{(ij)}$$

where $$\underset{\sim}{x}_{(ij)} = (x_{1(ij)}, ..., x_{p(ij)}), \qquad \underset{\sim}{\mu} = (\mu_1, \mu_2, ..., \mu_p),$$

$$\underset{\sim}{\beta}_{(i)} = (\beta_{1i}, ..., \beta_{pi}), \qquad \underset{\sim}{\epsilon}_{(ij)} = (\epsilon_{1(ij)}, ..., \epsilon_{p(ij)}),$$

with the sample consisting of $i = 1, 2, ..., n_1$ first-stage units and $j = 1, 2, ..., n_2$ second-stage units in each first stage unit.

Further, $$E(\underset{\sim}{\beta}_i) = (0, 0, ..., 0) \quad \text{for all } i,$$

$$E(\underset{\sim}{\epsilon}_{ij}) = (0, 0, ..., 0) \quad \text{for all } i \text{ and } j,$$

and $$E(\beta_i'\beta_i) = \Lambda_1 \,(: p\times p \text{ matrix}) \quad \text{for all } i,$$

$$E(\underset{\sim}{\epsilon}_{ij}\, \underset{\sim}{\epsilon}_{ij}) = \Lambda_2 \,(: p\times p \text{ matrix}) \quad \text{for all } i, j.$$

$$E(\underset{\sim}{\beta}_i'\, \underset{\sim}{\epsilon}_{ij}) = O \,(: p\times p \text{ matrix}) \text{ for all } i, j.$$

The problem is to choose the optimum number of first-stage and second-stage units under given cost restrictions.

Several alternative criteria for allocation of sampling units have been considered below. But which of these criteria is optimum remains yet a subject of investigation.

3. THE USE OF PILOT SURVEY

Informations about Λ_1 and Λ_2 are required to design a survey and such informations will be naturally obtained from pilot surveys.

By carrying out an analysis of dispersion we can find the estimates of Λ_1 and Λ_2, and also test the hypothesis $\Lambda_1 = O_{(:p\times p)}$.

From a pilot survey conducted, we obtain an analysis of dispersion like the following :

dispersion due to	d.f.	s.p. matrix	covariance matrix	expectation of covariance matrix
between first stage	n_1-1	$S_{1(:p\times p)}$	$C_1 = (n_1-1)^{-1}S_1$	$\Lambda_2+n_2\Lambda_1$
between second stage	$(n_2-1)n_1$	$S_{2(:p\times p)}$	$C_2 = [n_1(n_2-1)]^{-1}S_2$	Λ_2
total	n_1n_2-1	$S_{(:p\times p)}$		

For Λ_2 and Λ_1 we can find estimates in terms of C_2 and C_1 (Chakravarti, 1952). To test whether there are real differences between the first-stage units i.e. to test whether $\Lambda_1 = O$, Wilks' Λ-test may be used. The test is

$$\Lambda = \frac{|S_1|}{|S_1+S_2|},$$

$$V = -m \log_e \Lambda,$$

$$m = (n_1n_2-1) - \frac{p+n_1}{2}.$$

Then V can be used as χ^2 with $p(n_1-1)$ d.f. as a first approximation. The approximation can be made closer by considering second and third terms of the asymptotic expansion of the distribution of V (Rao, 1952).

4. Planning of the survey

In the survey to be planned, suppose $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ are the means of p characters based on N_1 first-stage and N_2 second-stage per first-stage sampling units.

Then covariance matrix of $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ is given by

$$Var(\bar{x}_1, \ldots, \;_p) = \frac{\Lambda_1}{N_1} + \frac{\Lambda_2}{N_1 N_2} \; (: p \times p).$$

Let the cost function be of the form $T = a + b_1 N_1 + b_2 N_1 N_2$. Now for a fixed cost $T = T_0$, the problem is to obtain a suitable allocation of the sampling units to different stages so that maximum precision is attained in the estimated means of the different characters. If the Wilks' idea of a generalised variance for p characters (which is the determinant of the dispersion matrix) is accepted, then one method of allocation may be to minimise the determinant of the dispersion matrix subject to cost restrictions. So if the quantity

$$L = \left| \frac{\Lambda_1}{N_1} + \frac{\Lambda_2}{N_1 N_2} \right| + \lambda (T_0 - a - b_1 N_1 - b_2 N_1 N_2)$$

is minimised with respect to N_1 and N_2 we get

$$N_2 = \frac{|\Lambda_2|}{|\Lambda_2 + N_2 \Lambda_1| - |\Lambda_2|} \cdot \frac{b_1}{b_2}.$$

An explicit solution may be difficult to obtain (since this involves the solution of a $(p+1)$th order equation) except in specific cases. But an approximate solution for N_2 will be given by

$$N_2 \simeq \frac{|\Lambda_2|}{|\Lambda_2| + N_2^p |\Lambda_1| - \Lambda_2} \cdot \frac{b_1}{b_2} = \frac{|\Lambda_2|}{N_2^p |\Lambda_1|} \cdot \frac{b_1}{b_2},$$

that is
$$N_2^{p+1} = \frac{|\Lambda_2|}{|\Lambda_1|} \frac{b_1}{b_2},$$

or
$$N_2 = \left[\frac{|\Lambda_2|}{|\Lambda_1|} \cdot \frac{b_1}{b_2} \right]^{1/(p+1)}$$

For Λ_1 and Λ_2 we have to use their estimates in terms of C_1 and C_2 ; and the corresponding value of N_1 is easily obtained from the cost restraint.

Another approximate method may be to consider a linear compound of the sample means $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ and so choose the coefficients that the variance of the linear compound is maximised. Then with the usual cost-restraint, this maximum variance may be minimised to derive a logical allocation of the sampling units.

Still another rough but less troublesome method may be to neglect all the elements except the diagonal ones of the matrix $\frac{\Lambda_1}{N_1}+\frac{\Lambda_2}{N_1N_2}$ i.e. to consider the diagonal matrix $\frac{D_1}{N_1}+\frac{D_2}{N_1N_2}$ where D_1 has as its diagonal elements the variances of the first-stage means and D_2 has the variances of the second-stage means as its diagonal elements. Then minimising

$$\left|\left(\frac{D_1}{N_1}+\frac{D_2}{N_1N_2}\right)\right|+\lambda(T_0a-b_1N_1-b_2N_1N_2),$$

we obtain
$$N_2=\frac{|D_2|}{|D_2+N_2D_1|-|D_2|}\cdot\frac{b_1}{b_2},$$

that is
$$N_2^{p+1}\simeq\frac{|D_2|}{|D_1|}\cdot\frac{b_1}{b_2},$$

or
$$\mathrm{N}_2\simeq\left[\frac{|D_2|}{|D_1|}\frac{b_1}{b_2}\right]^{1/(p+1)}$$

As usual, $|D_2|$ and $|D_1|$ are to be replaced by their estimates from a preliminary survey.

5. NUMERICAL ILLUSTRATION

Here the data collected in a sample survey in 1951-52 by the Indian Statistical Institute for the determination of the acreage under the two crops jute and *aus* paddy are used to plan for a future survey. The survey was conducted in the State of West Bengal and for the purpose, the entire State was divided into a number of strata. Within each stratum, Unions which have on an average an area of 13 sq. miles, were treated as first-stage units, within each union clusters of about 40 plots formed the second-stage units. For our purpose clusters will be treated as sampling units and analysis is restricted to only one stratum. The actual design and data of the survey were adjusted a little to gain in simplicity so that the example may be regarded as a sort of abstraction from reality. The particular stratum chosen here is Cooch Behar. From this stratum, 20 first-stage units and 8 second-stage units in each first-stage units were drawn, selection being with equal probability at each stage. The two characters enumerated in a second-stage unit were the proportion of the area under jute (x_1) and the proportion under *aus* paddy (x_2). An approximate cost function was obtained on the basis of investigator time-record analysis. Activities of a field-investigator can be conveniently brought under three mutually exclusive categories: office-work and other miscellaneous work volume of which is more or less independent of the number of first-stage or number of second-stage units, camp-setting and similar type of work whose volume depend only on the number of first-stage units and journey from cluster to cluster and enumeration in each cluster wherevolume of work depends on the number of clusters and size of clusters. Total time spent on the first type of activity gives the fixed part of the cost function and average time spent on the rest of the two types determine the remaining two coefficients of the cost function. Cost function so obtained and as considered here is in unit of investigator-hour. In order to bring it to monetary unit, it should be multiplied by a factor, i.e. cost perinvestigator-hour. This, however, does not affect the determination of the optimum number of units. Results of the analysis are given below.

dispersion due to	d.f.	s.p. matrix x_1 x_2	covariance matrix
between first-stage units	19	$S_1 = \begin{array}{c} \begin{bmatrix} 0.5502 & 0.2093 \\ .. & 1.1026 \end{bmatrix} \end{array}\begin{array}{c} x_1 \\ x_2 \end{array}$	$\begin{bmatrix} 0.0294 & 0.0157 \\ .. & 0.0580 \end{bmatrix}$
within first-stage units between second-stage units	140	$S_2 = \begin{bmatrix} 1.0872 & 0.3568 \\ .. & 3.4041 \end{bmatrix}$	$\begin{bmatrix} 0.0078 & 0.0025 \\ .. & 0.0243 \end{bmatrix}$
total	159	$S = \begin{bmatrix} 1.6464 & 0.6561 \\ .. & 4.5067 \end{bmatrix}$	

In order to test $\Lambda_1 = O$ i.e. whether staging has been effective the statistic $V= -m \log_e \Delta$ was calculated

where $$\Delta = \frac{|S_2|}{|S_1+S_2|} = 0.5113$$

and $$m = (n-1)-\tfrac{1}{2}[p+(n_1-1)+1] = 159-\tfrac{1}{2}(2+19+1) = 148.$$

V is distributed in large samples as χ^2 with $p(n_1-1) = 2\times 19 = 38$ degrees of freedom and here V comes out as 99.278 and is highly significant.

Next the same hypothesis was also tested by the variance ratio method (Rao, 1952). Here the statistic used is $v=\frac{1-\Delta^{1/s}}{\Delta^{1/s}}\frac{(ms+2\lambda)}{2r}$ which is distributed as an F with with $2r$ and $ms+2\lambda$ degrees of freedom. Here

$$m = (n-1)-\tfrac{1}{2}[p+(n_1-1)+1] = 148,$$

$$s = \sqrt{\frac{p^2(n_1-1)^2-4}{p^2+(n_1-1)^2-5}} = \sqrt{\frac{4.19^2-4}{4+19^2-5}} = \sqrt{\frac{1440}{360}} = 2,$$

$$r = \frac{p\times(n_1-1)}{2} = 19, \lambda = -\frac{p(n_1-1)-2}{4} = -9;$$

therefore $$v = \frac{1-\sqrt{\Delta}}{\sqrt{\Delta}} \cdot \frac{148\times 2-2\times 9}{2\times 19} = \frac{1-0.7144}{0.7144} \times \frac{139}{19} = 2.925.$$

As seen from the F-table, this value of v with 38 and 278 degrees of freedom is also highly significant.

The estimates of Λ_1 and Λ_2 in terms of S_1 and S_2 come out as

$$\Lambda_1 \simeq \begin{bmatrix} 0.0027 & 0.0025 \\ \dots & 0.0042 \end{bmatrix},$$

$$\Lambda_2 = \begin{bmatrix} 0.0078 & 0.0025 \\ \dots & 0.0243 \end{bmatrix}$$

The empirical cost (in investigator hour) function obtained is

$$T = 2250+8.7N_1+2.5N_1N_2.$$

On optimising with total cost fixed at T_0 (say), N_2 is given by

$$N_2 = \frac{|\Lambda_2|}{|\Lambda_2+N_2\Lambda_1|-|\Lambda_2|}\cdot\frac{8.7}{2.5}.$$

Substituting sample estimates for Λ_1 and Λ_2 we get

$$N_2^2 = \frac{0.6345}{0.0902+N_2\,.0087}.$$

So N_2 is to be obtained as a root of the equation

$$0.0087\,N_2^3+0.0902\,N_2^2-0.6345 = 0.$$

It is seen that the root lies between 2 and 3. So N_2 can be taken as 3.

Using the approximation suggested i.e.

$$N_2 \simeq \left[\frac{|\Lambda_2|}{|\Lambda_1|}\,\frac{b_1}{b_2}\right]^{1/(p+1)} = \left[\frac{|\Lambda_2|}{|\Lambda_1|}\cdot\frac{8.7}{2.5}\right]^{\frac{1}{3}}.$$

N_2 is found to be 4.18. Whether the approximation is good enough is a subject matter of further investigation.

References

Chakravarti, I. M. (1952): Use of the analysis of covariance in two-stage sampling. *Cal. Stat. Ass. Bull.*, **4**, 127.

Cochran, W. G. (1939): The use of the analysis of variance in enumeration by sampling. *J. Amer. Stat. Ass.*, **34**, 492.

Rao, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, John Wiley & Sons, New York.

Paper received: September, 1953.