A NOTE ON INTERPENETRATING SAMPLES

By V. K. Mokashi

Institute of Plant Industry, Indore

The method of interpenetrating net-works of samples is a particular design in sample surveys in which the sample units are arranged in two or more independent sets of samples within each domain of study. Information for each set of samples supplies an independent estimate of the variate under study. Thus the term interpenetrating samples is synonymous to replicated sampling. The method is due to Mahalanobis (1942), who employed it in the area surveys carried out in Bengal and Bihar. It was intended to secure information on non-sampling errors mainly arising through differential investigator bias. In other words, the interpenetrating samples were employed as a means of control of the quality of information secured through different parties of investigators, since a comparison between the different estimates would show whether there were significant differences between different groups of investigators. This design is advocated by Mahalanobis as an essential feature of sample surveys. Recently the United Nations' Sub-Commission on Statistical Sampling has also recommended the method of interpenetrating samples.

Panse and Sukhatme (1948) have discussed the utility of this method with reference to Mahalanobis's data from the Bihar crop survey and have shown that the method does not work satisfactorily in serving as a useful statistical control, over the reliability of the field work and consequently the extra cost involved might be diverted more profitably towards providing more adequate and active supervision over the field staff. These authors emphasize that an internal agreement between the two samples enumerated by different sets of investigators cannot provide a critical evidence for judging whether the field results are reliable. External evidence entirely independent of the survey is essential for this purpose. For example, in the crop estimating surveys on cotton in C.P. and Berar (Panse and Kalamkar, 1944) arrangements were made to find out the total production of cotton through ginning factory returns. These ginning factory returns when corrected for the import and export of cotton into and out of the area under survey, formed an independent and valid check for verifying the estimates of production derived from the crop estimating surveys. In commenting

on the various uses of interpenetrating samples, Yates (1949) has also recognised that unless each of the interpenetrating samples provides an adequate sample of the material, the comparisons between different samples will be subject to relatively large errors. If, for instance, they are used to test the differences between different investigators, the information obtained will be of insufficient accuracy to be of any real use. He adds that the proper use of interpenetrating samples necessitates increased expenditure on travelling.

In considering the actual method of sampling by which interpenetration of sample units is achieved, at least three different patterns are described by Mahalanobis.

(1) In the jute area survey in Bengal in 1941 (Mahalanobis, 1944), linked pairs of sample units were located at random on maps and one sample unit of each pair was allotted to half-sample A and the other to half-sample B. A constant distance of $\frac{3}{4}$th of a mile was maintained between the sample units that formed a pair. The field data were collected for the two half-samples by separate groups of investigators. The closeness of agreement between the estimates for the two half-samples was believed to supply a good idea of the reliability of the survey.

(2) In the Bihar crop survey (Mahalanobis, 1945) each district was divided into 96 zones and the sub-samples were obtained on the basis of these zones. The 96 zones were arranged in 48 pairs, each of two adjoining zones. The two members of each pair were allotted at random to sub-sample A or to sub-sample B. In this arrangement the method of sub-sampling secures interpenetration of the sample units at the stage of zones instead of at the stage of the sample units themselves.

(3) More recently Mahalanobis (1946) has suggested a different pattern though it is not known whether it has been actually tried in practice by him. In this method, the whole set of randomly located sample units is subsequently divided into two random sub-sets. The odd numbered sample units form one random set and the even numbered units a second set.

Of these three patterns only the first and the third have been considered in the present note. The second pattern is crude and according to Mahalanobis himself, the arrangement did not secure a detailed or fine-grained interpenetration of the sample units which is desirable.

The object of the present note is to examine the statistical efficiency of the method of interpenetrating samples. For convenience, this examination is made with reference to the area survey in which the proportion of area under a given crop to the total area surveyed is required to be estimated, since it is for such material that some actual data are available. The approach and application of the results are, however, perfectly general.

When the sample units are paired, a certain degree of intra-class correlation between the members of a pair is naturally introduced. The estimated variance of the mean value, in this case, the proportion of area under a given crop, based on paired samples would increase as compared to that based on the same number of sample units independently located at random. This increase depends upon the magnitude of the correlation between the sample units. There is thus an obvious loss of statistical precision when sample units are paired. Mahalanobis (1944) has referred to the loss of information due to pairing of sample units. The loss would be further enhanced if considered in relation to the cost of the survey. The ultimate efficiency of a survey design should be studied in the light of both the precision of the estimate and the cost at which this precision has been attained. For maximum efficiency a sampling design should provide maximum amount of information per unit of cost incurred. Applying this test to the method of interpenetrating samples by pairing of sample units, it is found that there is not only a loss of statistical precision, but also the cost of the survey is simultaneously increased. This is due to the fact that in sample surveys where moving investigators are employed, journey time is an important item of cost and in some cases, may form a major component of cost. It is the journey time that is affected with interpenetrating samples, other components, depending upon the number of sampling units in the survey, such as the amount of field enumeration or statistical computation, remaining the same.

Expressions for the loss of information in relation to cost resulting from the method of interpenetrating samples are derived below:

In relation to an area survey, let

A = Geographical area (in square miles) under the survey.

p = Proportion of land under a given crop.

Then

Ap = The estimated area in square miles under the crop.

Also let,

V = The variance of the estimated area (Ap) under the crop (without interpenetration of sample units).

T = The total cost of the survey in rupees per square mile.

$2n$ = Total number of sample units enumerated in the survey.

The total cost may be split into components due to (1) Journey, (2) Field Enumeration, (3) Miscellaneous work, and (4) Statistical work. If c_j, c_e, c_m and c_s represent these components per square mile,

$$T = c_j + c_e + c_m + c_s$$

or

$$T = c_j + c_k, \text{ where } c_k = c_e + c_m + c_s$$

which may be considered to be constant for a given number of sample units in the survey.

Now the amount of information in respect of the estimated proportion of area under the crop is given by

$$I = \frac{A^2}{V} \qquad (1)$$

the total cost of the survey is, $A(c_j + c_k)$. Therefore, the amount of information per unit of cost may be shown as

$$I_c = \frac{A}{V(c_j + c_k)}. \qquad (2)$$

This is the expression appropriate when the sample units are independently located at random.

When the sample units are paired and the correlation coefficient between the units of a pair is ρ, the amount of information in respect of the estimated proportion of area under the crop will be

$$I' = \frac{A^2}{V(1+\rho)}. \qquad (3)^*$$

Comparing (1) and (3), it is evident that there is a loss of information in the estimation of crop area, when units are paired, the total number of sample units remaining the same. Consideration of the journey cost involved in this particular arrangement of the sample units shows that this loss is increased when calculated per unit of cost. If $2n$

* Following Mahalanobis (1944) the value of V has been assumed in this paper to remain unchanged with interpenetration of samples. The influence of interpenetration with two or more observers on V is being investigated.

sample units are located independently at random, the cost of journey would be approximately proportional to $\sqrt{2n}$. But if this number of sample units is distributed as n pairs randomly located at n points and two separate parties of investigators are required to collect the information for each member of a pair, the cost of journey would be proportional to $2\sqrt{n}$, since the n pairs of sample units are traversed twice. Thus the total cost of journey would be $\sqrt{2}$ times greater when the sample units are paired than when they are independently located at random and only one party of investigators is employed to collect the information.

Therefore, when the sample units are paired, the total cost per square mile would become $\sqrt{2}c_j + c_k$ and for the entire survey, $A(\sqrt{2}c_j + c_k)$. Thus, the information per unit of cost will be,

$$I_c' = \frac{A}{V(1+\rho)(\sqrt{2}c_j + c_k)} \tag{4}$$

From (2) and (4), we can calculate the percentage loss of information per unit of cost resulting from pairing. It is given by

$$\left\{1 - \frac{c_j + c_k}{(1+\rho)(\sqrt{2}c_j + c_k)}\right\} \times 100 \tag{5}$$

From (5), we see that the loss of information per unit of cost depends upon the value of ρ and c_j. Table I has been constructed to show the

TABLE I

The percentage loss of information per unit of cost in the design of interpenetrating samples

c_j as fraction of c_k \ ρ	0	·13	·2	·3	·4	·5
1	17	27	31	36	41	44
3/4	15	25	29	35	39	43
2/3	14	24	29	34	39	43
1/2	12	22	27	32	37	41
1/3	10	20	24	30	35	40
1/4	8	18	23	29	34	38

percentage loss of information per unit of cost for values of ρ ranging from 0 to ·5 and values of c_j, expressed as a fraction of c_k, ranging from $\frac{1}{4}$ to 1.

In the Jute area survey in 1941, Mahalanobis (1944) estimated the value of ρ as ·13. From the data of the crop estimating surveys on cotton carried out in the Central Provinces, the correlation between fields within a village, which are analogous to sample units of a pair was found to be ·23. From a survey on wheat in Delhi Province, a correlation ranging from ·39 to ·56 was found (Sukhatme, 1949). As regards journey cost, c_j was computed as being ·43 times c_k from the jute area survey in 1941. It is interesting to note from Table I that for the particular values of ρ and c_j found for the jute area survey the loss of information in relation to cost was roughly 21 per cent. In the extreme case, considered in this table, when the value of ρ is ·5 and the journey cost as high as the cost for other items, almost half of the information is lost by the interpenetration of samples by pairing.

When the sample units are not paired but are independently located and then divided into two sub-sets as in pattern (3) described above, the value of ρ would be zero. There is consequently no loss of precision of the estimate through this kind of interpenetration of samples. But the cost of journey is affected as in the case of pairing. Here also each set of investigators has to travel to n random points scattered over the area under survey and consequently the journey cost is increased by $\sqrt{2}$ times the cost required without interpenetration of samples. From the first column of Table I, where $\rho = 0$, it is seen that this method of obtaining the interpenetration of the sample units leads to a loss of 8 to 17 per cent. in the amount of information per unit of cost.

The relative inefficiency of the interpenetrating design is an inherent weakness of the method as demonstrated above. With the additional consideration that it is of doubtful validity as a means of checking the reliability of the field investigators, there does not appear any justification for recommending the method for this purpose. Where, however, there is no risk of lowering the efficiency of the survey through interpenetration or replication of samples, it may be desirable to obtain information from independent sets of samples and the survey should then be designed to provide for such information being secured. This plan has already been adopted in the crop yield surveys conducted by the Indian Council of Agricultural Research. In areas where two field agencies are available, *viz.*, the field staff of the revenue or the

land records and agricultural departments, the total number of villages selected for sampling are divided into two random groups, one group being assigned to the staff of one department and the other to the staff of the other department. As in both cases the staff works within the area under their normal jurisdiction, no special travelling is involved and the cost of the survey is not affected. Without, therefore, affecting the efficiency of the survey, this sort of replication can provide information on the relative efficiency with which the two agencies carry out the field work.

SUMMARY

The method of interpenetrating samples is a design for the sample surveys in which the sample units are arranged in sets of two or more interpenetrating samples and the information for each set is collected in an independent manner. Mahalanobis has used this design in the area surveys he carried out in Bengal and Bihar as a means of controlling the reliability of field work. The statistical efficiency of the design in relation to the precision of the estimate and the cost involved, has been examined in the present note. It has been shown that the method leads to an appreciable loss of information per unit of cost. This loss for the jute area survey in Bengal is computed at 21 per cent. In more extreme cases nearly half of the information may be lost. In the case where the sample units are independently located at random and are then grouped into two sub-samples, the loss of information per unit of cost would still be 8 to 17 per cent.

REFERENCES

1. Mahalanobis, P. C. .. "Presidential Address, Section of Statistics," *Indian Science Congress*, Baroda, 1942.
2. ——— .. "On Large-scale Sample Surveys," *Phil. Trans. Royal. Soc. London*, 1944, **231**, 41.
3. ——— .. "Bihar Crop Survey," *Sankhya*, 1945, **7**, 29.
4. ——— .. "Sample Surveys of Crop Yields in India," *ibid.*, 1946, **7**, 269.
5. Panse, V. G. and Sukhatme, P. V. .. "Crop Surveys in India—I," *Journ. Ind. Soc. Agri. Stat.*, 1948, **1**, 34.
6. ——— and Kalamkar, R. J. .. "Estimation of Crop Yields," *Curr. Sci.*, 1944, **9**, 223.
7. Sukhatme, P. V. .. "Efficiency of Stratification and Size of sampling unit in a sub-sampling design in yield surveys." *Journ. Ind. Soc. Agric. Stat.*, 1949, (In press).
8. Yates, F. .. *Sampling Methods for Censuses and Sample Surveys*, Charles, Griffin & Co., London, 1949.

EXPRESSION OF UNITARY COMPONENTS OF THE HIGHEST ORDER INTERACTIONS IN 3^5, 3^6, 4^4 AND 5^3 DESIGNS IN TERMS OF SETS FOR THESE INTERACTIONS

By K. Kishen

Chief Statistician, Department of Agriculture, Uttar Pradesh, Lucknow

1. Introduction

In a previous paper (Kishen, 1942), a general method was developed for expressing any single degree of freedom for treatments in the case of the general symmetrical factorial design s^m, s being a prime positive integer or a power of a prime and m any positive integer, in terms of its sets for main effects and interactions, and was utilized for obtaining expressions for the unitary components of the third order interaction in a 3^4 design and of the second order interaction in a 4^3 design. When, however, the single degree of freedom belongs to a $(k-1)$-th order interaction (k varying from 1 to m), a simplified and short-cut method of deriving these expressions has been described in the present paper and has been employed for deriving expressions for the unitary components of the highest order interactions in the 3^5, 3^6, 4^4 and 5^3 designs. Throughout this paper, when dealing with the finite elements of the m-dimensional finite projective geometry $PG(m, s)$, we shall as usual write their co-ordinates, equations, etc., as if they belonged to the m-dimensional finite Euclidean geometry $EG(m, s)$ immersed in the projective geometry (Bose and Kishen, 1940).

2. Method of obtaining Expressions for any Single Degree of Freedom belonging to the $(k-1)$-th Order Interaction in an s^m Design

In an s^m design, let any treatment combination (or the quantitative measure of the result of application of the treatment combination) be represented by the symbol $a_1^{i_1} a_2^{i_2} \ldots a_m^{i_m}$, where $a_r^{i_r}$ denotes the i_r-th level of the r-th factor (i_r varying from 0 to $s-1$, and r varying from 1 to m). Then any single degree of freedom belonging to treatments may be written as

$$L = \Sigma\, l_{i_1 i_2 \ldots i_m}\, a_1^{i_1} a_2^{i_2} \ldots a_m^{i_m} \quad (i_1, i_2, \ldots, i_m \text{ varying from } 0 \text{ to } s-1),$$

where $l_{i_1 i_2 \ldots i_m}$ is a constant coefficient such that $\Sigma l_{i_1 i_2 \ldots i_m} = 0$.