COST AND ACCURACY OF RESULTS IN SAMPLING AND COMPLETE ENUMERATION

by P. C. Mahalanobis

Statistical adviser to the Cabinet, Government of India, Calcutta

The increasing use of statistical sampling is primarily due to the great economy of costs in terms of both money and labour. Actual experience has also shown that results based on sampling are often more reliable than results obtained from complete enumerations attempted at many times greater cost.

Cost in relation to precision of results was deliberately adopted as the theoretical foundation for the development of large scale sample surveys in the Indian Statistical Institute from 1937, and special efforts have been made to study both costs and precision. Numerical examples are given in the paper out of which a few illustrative tables have been selected for the present summary.

Margin of error of complete enumeration. One way of investigating the reliability of complete enumeration is to have the work done independently by two (or more) different sets of investigators. Extensive studies of this kind have been made by the «plot by plot» enumeration of agricultural crops.

Table 1. Discrepancy in «plot by plot» crop enumeration

Name of crop	Province and year of survey	Approximate total area covered in survey in sq.miles	Area under crop (in acres) as enumerated by party of investigators		Discrepancy (in acres) between A and B based on 'plot by plot' comparison		Sum of discrepancies (in acres)		Percentage discrepancy	
			A	B	positive	negative	total = col. 6+7	net = col. 4—5	total = from col. 8	net = from col. 9
1	2	3	4	5	6	7	8	9	10	11
Jute	Bengal 1937	200	355	385	103	133	236	— 30	66.5	— 8.5
Wheat	Bengal 1944	600	286	298	100	112	212	— 12	74.6	— 4.3
Gram	Bihar 1944	800	730	636	244	150	394	+ 94	53.9	+ 12.9
Rice	Bihar 1944	1800	6715	6481	1658	1424	3082	+234	45.9	+ 3.5

Discrepancies between the two independent A and B enumerations are based on a «plot by plot» comparison; when a plot occurs in the A-enumeration but not in the B-enumeration, it is counted as a positive discrepancy.

Sample survey of agricultural crops. Specifications of the design and the components of cost are given below for the sample survey of agricultural crops in Bengal in 1947/48. The survey was designed to supply an estimate of the total production of rice with a standard error of two per cent (which is believed to have been attained in practice). The survey was repeated in three crop-seasons in the same year.

Comparative cost of complete enumeration. A complete enumeration would require a field staff of about 25 000 and a statistical staff of about 5000 or

Table 2. Components of cost: Agricultural crops in Bengal, 1947–1948

Components of cost	Man years		Cost	
	total	per cent	per cent	per person per year (rupees)
1	2	3	4	5
Field section				
1. Investigation	352	47.7	24.6	586
2. Other services	155	21.0	12.0	647
3. Inspection	52	7.0	7.5	1218
4. Supervision	25	3.4	6.6	2217
5. Field staff	584	79.1	50.7	
6. Other expenses			17.4	
7. Total field			68.1	
Statistical				
8. Computation	73	9.9	10.2	1171
9. Other services	58	7.9	8.1	1177
10. Inspection	12	1.6	2.5	1782
11. Supervision	11	1.5	4.2	3169
12. Statistical staff	154	20.9	25.0	
13. Other expenses			6.9	
14. Total statistical			31.9	
15. Grand total		100.0	100.0	

Design of the survey

Stratification: geographical cells (64 sq. miles).
Sample-units of same size (2.25 acres) marked on maps (scale: 8 or 16 inches = 1 mile).
One stage randomization of same number of sample-units within each cell.
Two inter-penetrating samples (of equal size) randomized within each cell.
Total sampling fraction: about 1 in 167 in rice season.

Survey repeated in three crop-seasons

Season	Total area surveyed (sq. miles)	Number of sample-units
1. Jute-Rice	64 325	108 104
2. Rice	24 952	42 992
3. Winter crops	24 952	27 105
Total (3 seasons)	114 229	178 201

Total cost = Rs. 838 797.
Cost per sq. mile = Rs. 7.34.
Cost per sample-unit = Rs. 4.71.

a total staff of about 30 000 (forty times greater than that required in the sample survey). The cost would be about nine or ten times greater. The complete enumeration would give information for one season (or at most, partially for two seasons) but not for three seasons as in the sample survey. The reliability of the results of complete enumeration would be almost certainly less.

Cost in relation to precision in sample surveys. The higher the precision desired to be attained, the greater must be the cost of a sample survey. One great advantage is that the relation of cost to precision of results can be studied on a scientific basis. I am giving illustrative figures for the sample survey of early (Avs) rice in West Bengal (total area = 24 952 sq. miles) in 1948–1949. The estimated area under Avs (early) rice was about 1 240 000 acres (or about 7.8 % of the geographical area). The cost and variance functions used were derived from data for earlier years.

Table 3. Relation of cost to precision of results

E = per cent standard error of estimated area under rice	C = cost in rupees (= 0.3 US $) per square mile	Marginal cost $= -\frac{dC}{dE}$	$-\frac{1}{C} \cdot \frac{dC}{dE}$	$-\frac{E}{c} \cdot \frac{dc}{dE}$
1	2	3	4	5
12.84	0.50	0.0018		0.047
4.00	0.60	0.05	0.08	0.32
3.50	0.64	0.09	0.14	0.49
3.00	0.70	0.14	0.20	0.60
2.75	0.74	0.18	0.24	0.66
2.50	0.80	0.23	0.29	0.73
2.25	0.88	0.32	0.36	0.81
2.00	0.97	0.45	0.46	0.92
1.90	1.01	0.52	0.51	0.97
1.80	1.06	0.62	0.57	1.02
1.70	1.13	0.72	0.63	1.07
1.60	1.22	0.85	0.70	1.12
1.50	1.32	1.03	0.78	1.17
1.40	1.44	1.25	0.87	1.22
1.30	1.59	1.54	0.97	1.26
1.20	1.78	1.89	1.06	1.27

In the present case, the cost for reducing the standard error below two per cent or so becomes increasingly and relatively large. Fortunately, a standard error of two per cent is sufficient for most practical purposes. Also, and this point is important, even in a «good» complete enumeration, it would be probably impossible to reduce the margin of uncertainty to less than two per cent; in actual practice, the margin of uncertainty of a complete enumeration would almost certainly be much larger. In other words, the sample survey in practice would be not only much more economical but also more reliable than a complete enumeration in the present case.

Additional cost of using inter-penetrating samples. I shall make a few brief observations on the cost of using inter-penetrating samples. The only

addition will be in the cost of the field survey; and this also, only in that portion of the cost which is incurred for «journey» by the investigators from one sample-unit to another. In large scale crop-surveys, the journey portion may amount to half (but would be usually less) of the total work done by the field investigators and inspectors. In table 2, this amounts to about 32 %, so that the additional cost of using two inter-penetrating samples can, at the most, come to about 16 %. This, however, is practically an upper limit. In agricultural surveys, the additional cost is often of the order of 8 % or 10 %. In more localized surveys (such as, family budgets, labour enquiries, etc.) in which the «journey» time is small, the additional cost of using inter-penetrating samples would be appreciably less, and quite often not more than 3 % or 4 % of the total cost. The additional control secured, and the possibility of estimating (and eliminating) fluctuations arising from investigator bias and other factors would appear to be fully worth the additional price.