ON THE CHOICE OF SAMPLE CUTS FOR NATIONAL YIELD ESTIMATION SURVEYS FOR CEREAL CROPS

By J. M. SEN GUPTA

Indian Statistical Institute

SUMMARY. There has been a good deal of controversy in India about the choice of a sample cut from which it would be possible to get an unbiased estimate of the yield rate of cereal crops. The object of the present study is mainly (i) to give an account of the attempts made to explain the discrepancies between the estimates of yield obtained by the National Sample Survey (NSS) and those obtained by the States and (ii) to suggest a reliable yard stick against which the NSS estimates could be tested. The results of various type studies carried out at present, show that circular cuts used by NSS and rectangular cuts used by the States are more or less on the same footing so far as bias at the field level is concerned.

Yield rates obtained by harvesting of whole fields show slight deviation from those based on sample cuts. The idea of complete harvesting of fields as a valid and practical basis for testing the sample cut estimates should, therefore, be re-examined. The procedure is too costly and can hardly be executed on any reasonable scale under adequate control to prevent the incidence of large non-sampling errors. An alternative method, which is free from operational bias and which does not employ any configurational unit for sampling, has been suggested here.

1. Historical background of the current developments

Early years. Since the last two decades, there has been a good deal of controversy in India regarding the choice of a sample cut which would give an unbiased estimate of yield rate for the cereal crops. A bias may creep in at any of the stages, right from the selection of the sample units, harvesting of the ultimate sampling unit and finally in the estimation procedure itself. All our attention however has so far remained confined to prescribing the ultimate sampling unit, so as to minimise the ascertainment error, while other aspects of sampling, standardisation of the concepts and specifications have received very little thought.

As a sequel to the pioneering work initiated by Hubback (1927) in the early twenties, followed by Deshmukh in 1928 and Townend in 1938, studies on crop-cutting experiments employing cuts of small size were being continuously carried out by the Indian Statistical Institute (ISI) since the year 1939 on Juto and later, also on paddy crops. The incidence of an overestimating bias in cuts of small size was noticed in the year 1939 itself (Mahalanobis, 1940), but the size of sample in this year was rather small. The experiments in 1940 were on a more extensive scale, when the nature of this bias was put forward (Mahalanobis, 1941) in a published report. Attention to this phenomenon was drawn also in subsequent publications (Mahalanobis, 1944; 1946).

The Imperial Council of Agricultural Research (ICAR) now renamed as the Institute of Agricultural Research Statistics (IARS), also conducted similar experiments with cuts of varying shapes and sizes. The incidence of overestimating bias in small cuts was noticed in the ICAR experiments also. ICAR however got highly biased results in some of its experiments with small cuts of different sizes and shapes (Sukhatme, 1944-45; 1945-46; 1947), but in the ISI, magnitude of this bias was not so high. While ICAR totally rejected the use of small cuts and insisted on taking large rectangular cuts of the order of a tenth of an acre, ISI improved its crop-cutting instruments and was satisfied that cuts of radius 6'-8' were practically free from this over-estimating bias (Mahalanobis and Sengupta, 1951).

Since then, the ISI instrument has been very much improved and provided with auxiliary devices that aid in sharply defining the boundaries of the sample cut. Besides, a special treatment of crop standing just on the perimeter of the cut, has made a circular cut of 4' radius entirely satisfactory (Sengupta, 1961). This was accordingly introduced into the National Sample Survey (NSS) for its crop-cutting experiments since 1957. A brief account of the existing NSS technique for the location and harvesting of a sample cut within the selected field have been given in the Appendix.

The size of the sample cuts adopted by the ICAR has since been considerably reduced from one-tenth of an acre to 1/80 and even 1/100 of an acre in different States. The difference in cut size between the ICAR and ISI has thus considerably narrowed down. The choice of a small cut by the NSS which employs a mobile staff, is bound up with the type of its field organisation. Besides, it was economic and being small the crop could be threshed and weighed on the spot with greatest care. And although the variance within fields is higher for smaller cuts compared to that obtained with larger cuts, the contribution of this component towards the over-all error of the sample becomes unimportant, as the number of sample fields increases. The States, having permanent staff stationed all over it, could easily afford to take up larger cuts without any special difficulty, and were reluctant to break away from the traditional practice of using large size cuts.

Years since the inception of National Sample Survey (NSS). Since its 13-th round in the year 1957-58, the National Sample Surveys are bringing out estimates of production for the important cereal crops for the whole of India. The acreage is obtained through a direct observation of sample plots, while the yield rate is obtained by crop-cutting experiments with circles of 4' radius in randomly selected fields. The outturn thus estimated was found to differ appreciably (by about 33% in 1957-58) from the official figures which are estimated on the basis of acreage mostly compiled from Patwari records, and yield rates usually obtained with sample cuts of a larger size in the shape of rectangles or triangles. In the present days of acute food shortage this uncertainty about our own resources has naturally caused a good deal of concern in all quarters. Instead however of checking up these production figures against the consumption totals objectively determined, or by a critical review of the official figures themselves, doubts have been raised regarding the validity of sampling methods outright.

Object and scope of the present studies. The object of the present studies is to (i) give an account of the attempts so far made in explaining the discrepancies between NSS estimates of outturn and official figures, (ii) enumurate possible sources of defects in the official estimates that have yet to be exlpored, (iii) enumurate the elements of bias in the existing NSS technique, if any, and finally (iv) to suggest a reliable yard-stick against which the NSS estimates could be tested on the basis of specially conducted experiments.

This paper will however deal only with problems of locating and demarcating a sample cut, in so far as they affect an unbiased estimation of yield rate at the field level. Bias introduced through the process of selecting the fields or of the higher stage units and in the estimating procedure itself, will not be considered here. The discussion in the earlier paragraphs is intended merely as a general background of the current problem and to give an idea of the urgency and relative emphasis due to its different aspects.

2. Recent attempts for an appraisal of the NSS crop-cutting technique

Joint studies on the performance of sample cuts employed by the NSS and by the States. In the year 1960, a Technical Committee (TC) was set up by the Central Statistical Organisation (CSO) to investigate into the reasons for a large divergence between the NSS estimates and the official figures. Special type studies were recommended by this committee to ascertain how far the sample cuts used by NSS and States were responsible for this. On these recommendations the CSO drew up a scheme of crop-cutting experiments to be conducted under the joint supervision of the NSS, State and ISI. Accordingly Type-I studies, in which both the cuts were to be harvested in common fields under the closest supervision and control have been carried out at Bundi (Rajasthan, 1960) and Barh (Bihar, 1961). These studies did not however reveal any significant divergence between the two types of cut. In 1962, the Planning Commission set up another high power Technical Committee for further investigations. A second series of studies called Type-II exporiments were accordingly conducted on an extensive State-wise scale under more or less normal working conditions, both the cuts being however taken from a set of common fields. These schemes included also a programme of having as many of the sample fields wholly harvested as possible, so as to provide with an ultimate check on the sample cut estimates. In these experiments also, the two types of cuts did not show discrepancies which would explain the divergence between Official figures and NSS estimates.

A small divergence of the whole field rates from sample cut estimates so far as the circle is concerned, is not an unexpected feature in the experiences of the ISI. In fact, validity of the whole field rates being accepted as an yard-stick for the verification of the sample cut estimates has been questioned and discussed in subsequent paragraphs. A summary of results obtained from these Type-I and Type-II experiments (Report on the Type studies prepared by ISI)[0,1,2] has been given in Tables A.1 and A.2 in the Appendix.

In retrospect, such a joint study for the resolution of the differences between ICAR and ISI was proposed by the Indian Statistical Institute as early as 1956, and a concrete scheme of joint crop-cutting experiments in U.P. on wheat was drawn up jointly by the ISI and Dr. P. V. Sukhatme of the ICAR. The project however did not ultimately materialise and this important problem remained shelved for quite a long time. In Feburary 1947, at the suggestion of Dr. W. E. Deming of the US Bureau of Census then visiting India, ICAR was invited to join in a round table conference at Calcutta to discuss the existing differences. No response was however received from ICAR. Finally, in its meeting on 25 October 1948, ICAR Statistical Committee decided that a joint investigation was not any more necessary. In February 1949, Professor P. C. Mahalanobis from the Indian Statistical Institute made a last effort and urged ICAR to conduct a joint study, but this also met with the same fate. We are now at exactly the same spot as we were some nineteen years back.

---

[0] A report on the joint crop-cutting experiments on kharif Jowar in Bundi (Rajasthan) in November-December 1960 and on rabi wheat in Barh (Bihar) in February-April 1961 (mimeographed 1964).

[1] Some results of the Type-II studies on maize in Bihar, August-September 1963 and paddy in Andhra Pradesh, November-December 1963 (mimeographed 1964).

[2] Report on Typo-II studies on wheat in Uttar Pradesh, February-May 1965 (mimeographed 1965).

Special studies conducted by the Agricultural Statistics Division of the NSS Directorate. The Agricultural Statistics Division (ASD) of the NSS Directorate had also been conducting a parallel set of experiments since 1960 in these States, where the State experiments employing rectangular or triangular cuts are carried out under its technical supervision. The ASD supervisors were asked to take one circular cut of 4' radius from every field inspected by them to match against the State cut in the shape of a rectangle or a triangle. This incidentally provides with an additional material for comparing the NSS circles with the State cuts. Table A.3 in the Appendix gives a summarised account of the results compiled from the Report by the NSS Directorate[3, 4] relating to the years 1960-61, 1961-62 and 1962-63 which shows a very satisfactory agreement between the NSS cuts and State cuts. It may be noted that the circular cuts were harvested by the supervising staff and not by the primary staff normally employed. Nevertheless, the general agreement in the results is remarkable and at least reveals that there was no inherent difference between the two types of cuts.

Whole field rate as a reliable yard-stick for appraisal. In the above type studies, it has been taken for granted that the whole field rate obtained by dividing the whole field produce with its total crop area would provide the most valid basis for an appraisal of the sample cut estimates. The quantitative difference in the scale of operations involved in the two cases, leading almost to a qualitative departure, should not however be lost sight of. While the sample cuts measure the yield potential as standing on the ground per unit of acre, permitting little loss in the stages of harvesting, a whole field harvesting is altogether of a different dimension with uncontrolled losses in all the stages of carrying, storing, threshing and weighing, and represents the 'exploited total' *under the existing technique of exploitation*. An improvement in this technique, mechanisation in harvesting for instance, may altogether change this technological ratio in large scale exploitations.

Ordinarily, it should be enough to check up the findings of one sample by another independent sample, perhaps with variations in the sampling technique (one such scheme for estimating yield rates has been discussed later). As for the residual sources of bias that may remain undetected, it would be through a thorough checking at every stage of the operations, that the weaknesses will have ultimately to be spotted. There is no escape from this ultimate responsibility. If for instance, sample estimates continue to differ from the so-called population values for some unknown reason especially in the same direction, this task of spotting out of the defects by all sorts of measures, technical as well as administrative, will have to be undertaken any way, unless sampling as an estimating technique is written off.

In the joint studies discussed earlier, whole field harvests were carried out on an extensive scale with a maximum of supervision that the circumstances permitted. Total yield thus obtained is therefore not exactly what a cultivator would get ordinarily, nor what would be obtained through sample cuts taken with the usual care. Besides, it is not always so easy to get the correct measurements of crop area sampled, unless done by really skilled Amins and perforably with independent replications. During the investigations made at Giridih in 1965 (results given in the Appendix), considerable discrepancies between replicated measurements relating to the same set of fields, have sometimes been observed.

[3]Report on circular cuts conducted by the NSS Supervisory Staff in the Agricultural Statistical Division 1960-61 and 1961-62 (mimeographed 1962).

[4]Report on circular cuts conducted by the NSS Supervisory Staff in the Agricultural Statistical Division 1962-63 (mimeographed 1963).

3. Certain weaknesses in the official methods of estimation that have yet to be removed

From the special studies discussed above, it appears that the large divergence of NSS estimates from the official figures can hardly be explained by the 'mis-demeanour' of the sample cuts, rectangular or circular, i.e. by the small discrepancies at the 'cut' level. To isolate the major factors causing this divergence, we must have to look up elsewhere. Official estimates of crop acreage are based on Patwari returns (village level official, among the multifarious duties loaded into whom, preparation of the crop returns is merely one) for the temporarily settled provinces, while in the permanently settled States like West Bengal, other Agencies (Sample Survey in W.B.) furnish the estimates. Upto the year 1948-49, no regular agency for reporting crop statistics existed for about one-fifth of India, known as the 'non-reporting' areas. A substantial portion of this area has since been brought under the system of 'conventional reporting' (NSAIB Report),[8] the exact nature of which is not very clear. This leaves a still non-reporting area of the order of 11% of the Indian Union. From the above Report it also appears that in a large number of States, the final area estimates have to be drawn up without utilising the actual acreage based on crop inspection (NSAIB Report, 16) and due to delay in submission of crop-cutting returns, "a large number of these returns cannot be utilised for framing estimates of yield for purposes of the various forecasts" (NSAIB Report, 25).

Estimates for these 'conventional' and 'non-reporting' regions are presumably built up on *ad-hoc* reports received from indefinite sources. The non-reporting areas (possibly the relatively under-developed tracts) must be developing fast and the *ad-hoc* system of reporting for these areas should be replaced forthwith by suitably regularised agencies.

It is well known that no uniform procedure is adopted for the whole of India on the method of crop enumeration which varies from State to State and even within the same State. Area under component crops sown in a mixture are more or less arbitrarily arrived at. The basis of allocation also differs from State to State, sometimes carried out at the plot level and sometimes at the district or State level. Some of the minor components in crop mixture are also arbitrarily kept out of the routine crop returns, which taken over the entire State might assume considerable proportions. All these render the official estimates of crop acreage open to much doubts.

NSS, on the other hand, does not attempt to allocate a so-called net area to the component crops when sown as a mixture. It credits the entire geographical area to each of the components, carries out crop-cutting experiments in all kinds of plots sown singly or in mixture and relates all yields to the corresponding geographical or 'gross' acreages. In its 13-th round, NSS made an artificial allocation of the component crops according to the eye-estimated proportions of area occupied by each at the field level, and thus an estimate of the 'effective' or 'net area' was also worked out tentatively. This so-called 'notarea' was more or less in nominal agreement with the official 'net area' figures, and since the outturn figures were widely divergent, the entire divergence was attributed to the defect in the estimation of the yield rates.

[8] Report of the National State Agricultural Intelligence Board, April 1961, on Crop Statistics in India.

In mixed cultivation, the yield rate of each individual component is proportionately much higher than the fields cultivating them singly (Mahalanobis, 1945, also NSS Report No. 38). In other words, with a relative intensity of say 50 : 50 in a wheat-barley mixture, the yield rate of wheat is found to be much above one half the yield rate observed in fields where wheat is grown by itself. Thus, if the official yield rates of wheat refer merely to fields cultivating wheat singly but wheat-barley mixtures are left out of the experiments (as in U.P.) or if the selection of wheat fields is otherwise biased, then the official estimates of wheat outturn might be appreciably affected in spite of the nominal agreement between the estimates of a so-called 'net acreage' under wheat. This calls for a careful review. In fact, the need for an appropriate weighting of the specific yield rates in pure and mixed plots, by their respective acreages have been particularly stressed in the NSAIB Report, p. 24.

Finally, the estimation procedure adopted by the NSS and by the States in arriving at the outturn figures should also be carefully reviewed for an exhaustive investigation into the sources of disagreement.

4. Elements of bias in the existing NSS technique

It has already been pointed out that in NSS cuts, a bias upto the field level, if any, does not explain the large discrepancy between sample estimates of outturn and the corresponding official figures. It is quite possible of course, that the NSS technique for crop-cutting experiments still suffers from some elements of bias, which although small could be further reduced. The sources from which a sample estimates for selected fields may get biased may be enumerated as: (1) bias at the cut level i.e. in assessing the correct yield due to a demarcated cut; (2) bias at the field level, i.e. in arriving at the yield rate of a selected field.

Bias at the cut level. A bias at the cut level may be induced either by instrumental errors or in the identification of the crop-plants falling within the cut enclosure. So far as the circle is concerned, the tendency of over/under inclusion of plants, if any, is largely neutralised by the special procedure laid down for the treatment of border plants see Appendix). Besides, with the improved apparatus now in use, the chances of over-stopping the circumference or falling short of it is very much reduced. For larger cuts in the shape of rectangles or triangles, the perimeter is proportionately very small and thus a perimeter bias is unlikely to affect the yield rate seriously.

Bias at the field level. There will be a bias if all units constituting the total field do not get an equal chance in selection. This may arise from : (1) an operational defect in the act of location, leading to a discontinuity in selection; or an under-representation of particular strips; (2) configurational restrictions imposed by the size and shape of the sample cut vis-a-vis that of the sample field, leading to an under-sampling of the field borders.

Existing method of location in integral steps. The location of random points within a selected field at discrete intervals of integral steps is a defect, perhaps a minor one, since it precludes the chance of locating a point in between two integral steps and thus leads to an under-sampling of alternate strips in a systematic pattern. Besides, it causes total non-sampling of a strip nearly one foot wide, all round the border, a minimum of two steps (2'—6" each) being needed for the location of a full cut of four feet radius.

In order to study the effect of varying units of measurement on the location pattern, a small experiment was conducted at Giridih in 1964 and the results, although inconclusive, are recorded here as an item of general interest. The details of this experiment and its findings have been furnished in the Appendix.

Under-sampling of the field borders and rejection of partially included cuts. This is a phenomenon which affects the larger cuts more seriously than cuts of a smaller size. For circular cuts, a strip 4' wide, all around the field is under-sampled. In the absence of any differential in yield rate towards the borders compared to the central parts, this may not seriously affect the results. It would be desirable however to investigate this question more closely, so as to eliminate this source altogether.

In NSS crop-cutting experiments, cuts falling near the field borders, of which only a part falls inside the field and hence called 'partial' cuts, are harvested and an eye-estimation is made of the percentage of cut area actually included within the field. But another full cut in substitution of such 'partial' cuts is also taken from the same field. In the estimation of yield rates, the 'partial' cuts are however excluded. It will be interesting to find out how the estimates are affected if the 'partial' cuts were accepted as such. We may, for instance, work out the yield rates based on 'partial' cuts and compare them with the same obtained from their full-cut substitutes. For estimating these yield rates, total yield obtained from the 'partial' cuts will have to be inflated by the inverse of the eye-estimated proportions of cut area included within the field.

The following table gives the unweighted yield rates in tolas per circle* of 4' radius based on 'partial' cuts and their substitutes as in the 13-th and 14-th rounds of the NSS.

TABLE 1. UNWEIGHTED MEAN YIELD IN *TOLAS* PER CIRCLE OF 4' RADIUS, AS OBTAINED FROM 'PARTIAL' CUTS AND THEIR FULL-CUT SUBSTITUTES IN THE 13-TH AND 14-TH ROUNDS OF THE NATIONAL SAMPLE SURVEY

crop state	NSS rounds	number of cuts		totals per circle of 4' radius based on		difference cols. (5)-(6) with s.e.	students *t*	p.c. of fields with	
		all	'partial' with substitutes	full cut	partial cut			'partial' cuts under estimating	'partial' cuts to all fields
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
crop : winter paddy									
Uttar Pradesh	14	198	12	67.8	62.7	+5.1±6.9	0.73	50.0	6.1
Orissa	14	194	5	66.5	67.0	−0.5±3.0	0.17	20.0	2.6
West Bengal	13	369	21	67.5	70.9	−3.4±6.5	0.53	42.9	5.7
	14	213	17	69.5	66.2	+3.3±6.1	0.54	47.1	8.0
Andhra Pradesh	14	262	13	98.1	102.0	−3.9±6.7	0.58	46.2	5.0
Madras	14	109	5	95.1	94.0	+1.2±9.2	0.13	40.0	4.6
Bihar	13	334	19	57.6	48.7	+8.9±5.6	1.58	73.7	5.7
	14	201	7	47.0	39.4	7.6±6.9	1.10	42.9	3.5
crop : spring wheat									
Bihar	14	158	8	41.8	33.0	+8.8±9.5	0.93	50.0	5.1
Uttar Pradesh	13	431	23	48.5	45.2	+3.3±5.9	0.56	56.5	5.3
crop : winter jowar									
Uttar Pradesh	14	147	5	16.1	16.0	+0.1±3.6	0.03	60.0	3.4

* Tola = 11.7 grams (approximate)

The 'partial' cuts, i.e. rejected border cuts, perhaps show a slightly lower yield rate compared to their substitutes, the differences being however insignificant in all cases and the proportion of the fields in which 'partial' cuts have occurred is again very small. The resultant effect on the overall yield rate even by accepting the 'partial' cuts would thus be negligible. The number of cuts considered here are of course too small ; planned experiments to bring out this trend would be really interesting.

5. AN YARD-STICK FOR THE APPRAISAL OF CONVENTIONAL CUTS

Non-configurational units for sampling.[6] As discussed earlier, complete harvesting of whole fields is neither a practical nor a suitable basis for testing the merits of sample cuts, circular or rectangular. Obviously, a technique unaffected by the size or shape of the ultimate sampling unit, would serve us best as an yard-stick for appraising the conventional cuts, which as we suspect are affected by their size and shape. The following prerequisites should also be fulfilled as nearly as possible :

(i) quantity, i.e. bulk of the crop contents in the units employed should be comparable with that obtained from the conventional cuts ;

(ii) there should be no ambiguity or bias either in defining and identifying the unit or in its selection from the total field;

(iii) the operational procedure should be simple, so that a few investigators could guide a large team of labourers without undue confusion.

All these conditions, can very well be satisfied, if a selected field is harvested in small bundles of nearly uniform size, numbering them serially and then selecting a number of sample bundles either at random with replacements or systematically with a random start. The selected units would then be threshed and weighed separately. Obviously, lack of uniformity in the size of bundles would not introduce any bias in the estimated yield rate, apart from an increase in variability. If the variability in the yield rate per bundle unit happens to be reasonably small, the number of such units required to give a desired precision should be a mere fraction of the total produce. The subsequent stages of threshing, drying and weighing operations would thus be performed under conditions comparable to those in harvesting the conventional cuts.

In harvesting, a labourer goes on gathering plant after plant, until his hand i.e. fist, is full, which he lays down somewhere on the ground, so that,a headload of adequate size is built up. In some areas, the labourer ties up each handful with the flexible stalk of one of the plants constituting the handful. Such handfuls therefore retain their identity and ordinary handling does not undo these ties. It is evident, that a handful measure will vary from person to person, the fists of an adult being bigger than that of a boy's, a man's from a woman. Notwithstanding this, variability in the weight of paddy in single handfuls (h.f.) or per headload (h.l.) of a specified size, has been found to be reasonably low in the try-out experiments conducted at Giridih in the years 1963 and 1964, as will be seen from the results given below. The particulars of the experimental procedure and detailed results have been given in the Appendix.

[6] V. Nomchinov (USSR, 1932) has described a method of "sheaf-sampling" or "heap-sampling" in his paper "Measurement of crop through sampling" conceptually on the same lines, although the operational procedure is different.

CHOICE OF SAMPLE CUTS FOR NATIONAL YIELD ESTIMATION SURVEYS

Basic results. Table 2 below gives the mean yield rates of paddy in grammes per unit, based on the different sizes of sampling unit along with their coefficients of variation. In 1963 single handfuls in Field 1 and 'DBL' units[7] in Field 3 were drawn in a second stage from each headload of 32 and 16 handfuls respectively. In Field 2, the whole crop was harvested in headloads of 16 handfuls only. In 1964, subsamples in bundles of 8 handfuls were drawn both at random and systematically for the Fields 1–4 and in handful units selected systematically in Field 5. But in this table, results based only on the random samples have been shown except for Field 5 in 1964, which was systematic.

Although the sub-headload units (1 h.f. and 2 h.f.) in 1963 were drawn from within each headload (into which the whole crop had in effect been stratified), they are here considered to have been drawn unistage and the coefficients of variation shown in col. (5) were computed accordingly.

It will be seen that except in Field 1 of 1964, which was a partly inundated plot with extreme heterogeneity, the coefficients of variation in bundles of 8 handfuls to 32 handfuls range from 17% to 20%, while the same between single handfuls range between 25%–28%. The variability between bundle units of different sizes thus seems to be of a similar order as that obtained with conventional cuts. The results in greater details have been given in Appendix Tables A.8 and A.9.

From these results, it appears that a sample of 25 bundles of say 8 handfuls per field could estimate the field total with an error of 4% or so, while remaining perfectly unbiased. This method can therefore be profitably utilised for an unbiased estimation of the field total and hence furnishes a practical basis for testing the performances of other methods.

TABLE 2. MEAN YIELD IN GRAMMES PER UNIT WITH THEIR COEFFICIENTS OF VARIATION, GIRIDIH 1963 AND 1964

units	year/field number	number of units	yield in gms. of paddy per unit	
			mean	c.v.
(1)	(2)	(3)	(4)	(5)
1. sample handfuls (1 h.f.)	63/1	450	82.5	28.6
	64/5	74	78.0	25.8
2. 'DBL' units (2 h.f.)	63/3	450	226.6	25.5
3. sample bundles of 8 h.f.	64/1	32	403.0	39.0
	64/2	50	664.0	17.2
	64/3	50	733.0	17.4
	64/4	100	644.0	18.8
4. headloads of 16 h.f. (population)	63/2	411	1661.0	20.6
	63/3	225	1838.0	18.4
5. headloads of 32 h.f. (population)	63/1	225	2627.0	20.3

[7]Binary units first introduced by Professor D. B. Lahiri of the Indian Statistical Institute in 1952 for sampling and hence named as 'DBL' unit after him.

It would be noted here that the labourers were expected to harvest in bundles of uniform volume and the resultant uniformity in the grain contents per bundle is incidental. It will be interesting to conduct fresh experiments in which the labourers will be asked to harvest in bundles containing say one kg. of paddy. With the experienced eyes of a cultivator this may be quite effective in bringing down the coefficients of variations to a still lower level.

Operational procedure for an unistage selection of small bundles. The most efficient arrangement would be to sample unistage, employing sampling units of a very small size. One way to simplify the selection procedure would be to harvest in medium-size bundles of say 10 handfuls, but to make an unistage selection of smaller units of say, single handfuls by arranging the constituent handfuls in any order (or by a fresh splitting into ten parts by an eye-estimation) for only those bundles within which a selected handful chances to fall. The range of the random numbers would obviously be ten times the number of bundles harvested, and to ensure an unbiased selection, arrangement of the handfuls within the relevant bundles must be completed before the actual selection is made. The same procedure may be adopted for a systematic selection as well.

Operational cost. As harvesting progresses, more and more of the harvesting hands can be released and diverted for carrying the headloads to the central yard for threshing, clearing and winnowing. Since the scheme aims at a sampling of fields by threshing and weighing of only a limited number of headloads or bundles, although harvesting will have to be done for the entire field, the subsequent operations would be reduced to a mere fraction.

6. SUMMARY OF FINDINGS

From our past and recent experiences it appears that both the circles and the rectangles behave more or less similarly.

Yield rates obtained by harvesting of whole fields show a slight deviation although insignificant from those based on sample cuts. This may however be ascribed to uncontrolled losses in whole field harvesting leading to an underestimation, uncertainties in area measurement and finally to the sampling errors in sample cuts apart from a small residual bias, if any.

The sample cuts employed by the NSS may be considered to be free from bias at the cut level, but other sources of bias may still linger to affect the estimates at the field level. Although the magnitude of any such residual bias appears to be very small, in fact negligible, certain steps to reduce them are suggested for further investigations.

(i) *Under-sampling of the field borders (which in case of smaller cuts is a very small fraction of the total field) due to the rejection of partially included cuts.* The data collected in the 13-th and 14-th rounds of the NSS, do not bring out any differential in yield rate within 4′ of the field borders, compared to the rest of the field. Nevertheless, it may be advisable to take the smaller cuts of 2′–3″ radius or to admit the 'partial' cuts instead of substitutions. The effective size of the partial cuts may be read off with the help of a suitably calibrated scale.

(ii) *Non-continuous location of random points at discrete intervals of integral steps.* This can be improved by measuring the coordinates in tenths of a step, the tens being counted in full steps and the tenths being read off with a small scale of 2′·6″ divided into ten equal parts.

CHOICE OF SAMPLE CUTS FOR NATIONAL YIELD ESTIMATION SURVEYS

The idea of having fields harvested wholesale, as a valid and practical basis for testing the sample cut estimates should be re-examined. The procedure is too costly and can hardly be executed on any reasonable scale under adequate control to prevent incidence of large non-sampling errors.

Independent replication of field area measurements through trained Amins has revealed serious discrepancies in a recent experiment. Whole field yield rates obtained by dividing the whole field produce with the field area is thus liable to appreciable uncertainties.

Instead of searching for a 'population value' to test the merits of a sample cut of any particular size or shape, an alternative method free from operational bias and which does not employ any configurational unit for sampling, has been suggested as follows :

A selected field may be harvested wholly in more or less uniform bundles and out of these, a selection of bundles, either at random or systematically may be made, which would give an unbiased estimate of yield per bundle and hence total outturn for the field can be obtained. The scale of operations involved in dealing with these sample bundles will be similar to that employed in taking conventional cuts, threshing and weighing being completed on the same day and thus will be on a comparable footing in all respects.

Uniformity in the size of bundles is desirable, but not essential. A lack of uniformity may increase the variability but would not introduce any bias. The uniformity in practice has however been observed to be remarkable in the present experiments. A unistage selection in bundles or even of handfuls is feasible if the harvested sheaves are arranged on the ground in suitable groups. The variability in the yield of single handfuls is found to be of the order of 30%, which with bundles of 8 h.f. comes to 20%. Thus with a sample of 100 bundles it should be possible to estimate within 2% error for an individual field.

It is not however suggested here that the conventional sample cuts be replaced by bundle samples in yield estimating surveys. It is perhaps costlier and cannot be left to the sole care of a primary investigator. An adoption of this method of sampling in a sub-sample of field where ordinary cuts have been taken should prove us with a very good check over the latter. It need hardly be said, that an unbiased sample estimate, albeit with a small sampling error, is much to be preferred to a so-called 'population' figure independent of sampling errors but liable to uncertain non-sampling errors. As a matter of fact, whole field enumeration in future type studies can easily be replaced by a threshing and weighing of only a sample of bundles or bagfuls although the harvesting may be carried out in full.

Acknowledgement

For this study, I have to acknowledge the ungrudging help that I received from my colleagues in all stages.

Sri J. N. Taluqdar, Sri T. Sen and Sri P. Jacob have taken the full initiative in all phases of the try-out experiments conducted at Giridih, and I am particularly grateful to Sri G. Venkateswarlu, Statistician-in-Charge of our Giridih branch for his valuable suggestions and encouragement.

For the analysis of data, I wish to mention specially of Sri M. Chanda. Sri M. Ganguli and Sri M. N. Murthi have kindly helped me with their critical reviews.

Appendix

1. NSS TECHNIQUE OF CROP-CUTTING WITHIN A SELECTED FIELD

Definition of crop-area. In speaking of the crop area constituting a crop field to which the yield rates refer, it must be made clear as to how the *ail* areas, i.e. the raised foot-paths or bunds demarcating its boundaries are to be allocated. The conventional practice with some of the States and even with the NSS in its earlier rounds, was to consider a field extending up to the mid-*ail* line, such that all fields taken together would account for the entire geographical area. The NSS has since modified this practice and defines the fields as extending up to the inside foot of the *ails*, the *ail* area being reported separately. Whatever be the practice, in making comparisons between two types of sample cuts among themselves or between a sample cut and whole field estimates, the definition of the crop area sampled must be kept indentical in both.

Location and demarcation of sample cuts. Within a selected field, the centre of the sample cut is located with the help of a pair of random numbers representing the two coordinates which are mutually at right angles and more or less parallel to the sides of the field. The full ranges along the X-axis and Y-axis are first measured out in terms of the investigator's own steps, such that the rectangle formed by them encloses the entire field. Taking x-steps along the X-axis and y steps along the Y-axis the investigator reaches a point x, y inside the field at which he fixes the central peg of the crop-cutting instrument. Once the centre is fixed, the vertical screw is driven into the ground and the moving horizontal arm is extended to its full length of 4 feet and rotated, progressively harvesting all the crop lying fully inside the circle. The rest of the crop standing on the circumference, i.e. intercepted by it, is then separately harvested. The weight of paddy obtained from the border plants is also separately recorded, for an allocation at the stage of statistical analysis. Although the contribution of the border plants is extremely small the procedure is considered to exercise an effective control on the performance of the investigator, whose critical attention is arrested by this delicate act of discrimination.

2. RESULTS OF THE RECENT TYPE STUDIES ON NSS STATE CUTS (1960-65)

TABLE A.1. MEAN YIELD IN KG. PER ACRE AS OBTAINED FROM NSS (CIRCLE) AND STATE (RECTANGLE/TRIANGLE) CUTS IN DIFFERENT STATES FOR DIFFERENT CROPS

study type	state/centre and crop	size of state cut	number of fields	mean (un-weighted) yield in kg. per acre as per		mean difference [cols. (5)-(6)] ±s.e.
				circle	rectangle/ triangle	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
I	Bundi, jowar (1960)	33′×16½′	58	68.4	67.7	0.7± 5.2
II	Barh, wheat (1961)	,,	94	278.1	277.6	0.5± 8.4
II	Bihar, maize (1962)	,,	284	456.9	478.9	−22.0±15.3
II	Andhra, paddy (1963)	33′×13.2′	103	865.3	832.7	32.6±18.5
II	U.P., jowar (1965)	10 M. sq. Δ	144	480.0	478.8	1.2±13.9

TABLE A.2. MEAN YIELD IN KG. PER ACRE AS OBTAINED FROM NSS (CIRCLE) AND STATE (RECTANGLE/TRIANGLE) CUTS COMPARED WITH THE CORRESPONDING WHOLE FIELD HARVEST RATES IN BIHAR AND ANDHRA

study type	state and crop	number of fields	mean (unweighted) yield in kg. per acre as per			mean difference ±s.e.	
			circle	rectangle/ triangle	whole field	circle minus whole field	rectangle minus whole field
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I	Barh, wheat (1961)	9	234.6	219.0	224.0	10.6	− 5.0
II	Bihar, maize (1962)	200	448.6	459.0	412.3	36.3±15.7	46.7±11.0
II	Andhra, paddy (1963)	112	931.1	881.4	893.5	37.6±20.1	−12.1±10.6
II	U.P., jowar (1965)	84	498.0	501.8	483.7	15.2±12.6	23.1±12.8

TABLE A.3. COMPARISONS OF THE GROSS YIELD RATES BASED ON CIRCULAR CUTS OF 4' RADIUS AND STATE CUTS, AS OBTAINED IN THE SPECIAL STUDIES CONDUCTED BY THE AGRICULTURAL STATISTICS DIVISION OF THE NSS DIRECTORATE 1960-61 TO 1962-63

year	crop		n	yield of dry grain in lbs. per acre		difference col. (4) − col. (5)	p.c. of col. (6) to col. (5)
				circular cuts of 4' radius	state cut		
(1)	(2)		(3)	(4)	(5)	(6)	(7)
1960-61	kharif	paddy	20	1387	1399	− 12	− 0.9
		jowar	40	530	532	− 2	− 0.4
		maize	2	1127	894	233	26.1
		ragi	22	574	621	− 47	− 7.6
		all	84	760	770	− 10	− 1.3
	rabi	wheat	80	880	813	67	8.2
		gram	24	624	651	− 27	− 4.1
		jowar	80	483	438	45	10.3
		all	184	674	629	45	7.2
1961-62	kharif	paddy	16	1680	1943	−263	−13.5
		jowar	135	320	318	2	0.6
		bajra	85	228	231	− 3	− 1.3
		maize	99	900	881	19	2.2
		ragi	64	658	645	13	2.0
		all	399	553	557	− 4	− 0.7
	rabi	wheat	70	698	611	87	14.2
		gram	31	318	308	10	3.2
		barley	25	482	512	− 30	− 5.9
		jowar	62	368	371	− 3	− 0.8
		all	188	498	469	29	6.2
1962-63	kharif	paddy	30	1650	1840	−190	−10.3
		jowar	132	322	369	− 47	−12.7
		bajra	109	259	226	33	14.6
		maize	113	815	827	−12	− 1.5
		ragi	51	821	844	− 23	− 2.7
		all	435	584	609	− 25	− 4.1
	rabi	wheat	153	580	570	− 1	− 0.2
		gram	66	287	279	8	2.9
		barley	57	573	557	16	2.9
		jowar	61	301	293	8	2.7
		all	337	466	461	5	1.1

3. MISCELLANEOUS EXPERIMENTS CONDUCTED IN 1965

3.1. *Special experiments at Giridih in 1965 on the location of sample points by employing varying units of measurements.* The experiment consisted of actual locations of a large number of points in terms of full steps and half steps as two different units of measure with a batch of eight different investigators in two teams of four each. Four paddy fields were chosen, in two of which mid-*ail* was treated as the field boundary, and in two others, the inside foot of *ail* was treated as the field boundary. In all, a total of 576 points,

72 by each of 8 investigators, were located using full step units. In order to simplify the operations, twelve different points, six of them in fields with mid-*ail* border and six others in fields where foot of the *ail* represented the border, were located in advance. Since the investigators were now required to locate all the points along the y-axis only, the operation was reduced to a simple one-way location. The points located with 4' of the border at either end of these strips, were marked with pegs and the number of such points located by each team of 4 workers was counted up separately. The total length of each strip was then measured in feet and inches with the help of measuring tapes. For a given set of random numbers it is possible theoretically to work out the number of points that would have fallen within 4 feet of the border, if the 'steppings' were correctly executed, true to their original measure in terms of which the full range was determined.

Table A.4 gives the 'actual and correct' number of points due, to this border region totalled over the six strips in each border type employing either the full step or the half step units for location. Curiously enough, there has been a considerable excess of 'actual' locations over the 'correct' number in fields with mid-*ail* borders specially when using full step units. It may be noted that *ails* or bunds are sometimes raised considerably with an abrupt drop on either side. When mid-*ail* is treated as the field border, the first step taken from such borders must tend to be somewhat abnormal, the effective horizontal sweep being automatically shorter. It was also noticed that when the coordinate value exceeded half the range, the investigator usually went over to the other end of the strip and measured the difference in steps from that end. A shorter traverse from the immediate border may have also affected the measure of his steps. But the position remarkably improves as soon as we change over to half step units. The agreement is perfect, when half step units are applied in fields with the foot of *ail* as their borders.

There is thus a clear indication that the location improves as the unit of measure is made finer. Perhaps, the need for a finer judgement alerts the investigator to much greater attention than for a cruder one. While it will be impractical to resort to the use of measuring tapes in the location of points on an extensive scale, some improvement in this direction should be thought out.

TABLE A.4. NUMBER OF POINTS LOCATED WITHIN 4' OF EITHER BORDER OUT OF A TOTAL OF 96 LOCATIONS IN EACH EXPERIMENT USING FULL STEP AND HALF STEP UNITS FOR MEASUREMENT; WITH DIFFERENT CONVENTIONS FOR FIELD BORDERS

field border	teams of 4	full-step units		half-step units	
		actual	correct	actual	correct
(1)	(2)	(3)	(4)	(5)	(6)
mid-ail	1	40	29	31	33
	2	45	36	39	33
	total	85	65	70	66
foot of ail	1	33	28	32	30
	2	27	26	29	31
	total	60	54	61	61

3.2. *Experimental investigation into the accuracy of area measurements, Giridih*, 1965. An investigation into the accuracy of crop area measurements through qualified Amins was taken up at Giridih in 1965. The services of three registered Amins (two of them holding the "Amanat Diploma" while one with twelve years of services behind him) were made available through the courtesy of the Divisional Forest Office at Giridih, where they are employed. Nine paddy fields including these on which crop-cutting experiments were carried out in the years 1963 and 1964 were selected for this purpose. Three independent measurements were taken by these three surveyors on three different dates for each of these fields. Relevant working sheets and all sketches were taken away from each as soon as his work was completed. None of them

were told that the same fields measured on the day will be measured again for verifications. Table A.5 gives the area of each field in acres as returned by the three Amins independently. It will be seen that the mutual discrepancies were often considerable. Half of the range over which the three assessments are dispersed for an individual plot, have been expressed as a p.c. to their average and given in col. (8). Amin-3 seems to be unusually off from the other two in his calculations.

TABLE A.5. CONSISTENCY IN THE REPLICATED MEASUREMENTS OF THE SAME SET OF FIELDS INDEPENDENTLY BY THREE DIFFERENT AMINS, GIRIDIH 1965

field no.	owner	area in acres				range of dispersion	half range as % of average
		Amin-1	Amin-2	Amin-3	average		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	ISI	0.730	0.740	0.580	.683	.080	11.7
2	,,	.415	.395	.290	.367	.062	16.9
3	,,	.385	.265	.230	.293	.078	26.6
4	,,	.335	.260	.240	.278	.048	17.3
5	,,	.520	.480	.380	.460	.070	15.2
6	Chattu Gope	.595	.600	.520	.572	.040	0.7
7	,,	.675	.520	.480	.558	.098	17.6
8	,,	.615	.560	.510	.562	.052	9.3
9	Gairb Miah	.630	.630	.640	.633	.005	0.8

This reveals the danger of possible inaccuracies in area measurement even with trained Amins, of an order comparable with the discrepancies observed between sample cut estimates and whole field rates. The number of field must be made reasonably large to round up this elements of personal inaccuracy, with independent replications if bias also is suspected.

3.3. *Try-out experiments for an estimation of yield independent of conventional cuts.* A series of experiments for trying out this method of sampling for estimating the total yield of individual fields were carried out at Giridih in the years 1963 and 1964 on a number of fields with winter paddy.

In 1963 three fields were completely harvested in standard headloads of a specified number of handfuls. From each headload in field 1 two handful units were selected at random, while from field 3 two 'DBL' units were likewise selected in a second stage. In field 2 however, no such selection was made. The selection of 'DBL' units consisted in splitting each headload into approximate halves by eye-estimation and then selecting one of the halves at random. The same procedure was repeated three times in succession by splitting the selected half over again and selecting one of the halves. Precaution was taken to ensure that the random numbers used for selection were consulted only after the splitting was done. The ultimate unit thus represented one-eighth of the original headload, equivalent roughly to two handfuls. In selecting the second 'DBL' unit, splittings already made were made use of, the selection with replacement being however made afresh.

This selection in two stages for field 1 and specially for 'DBL' units in field 3 involved too much handling resulting in some shedding of grains (Table A.6). Besides, an analysis of the stage variances (Table A.7) indicated that a two-stage sampling has no appreciable advantage over a uni-stage one. On the other hand, an increase in the bundle size brings about a small reduction in the coefficient of variation (Table A.8).

In the year 1964, five fields were chosen, four of which were completely harvested in smaller bundles of 8 handfuls. A selection was then made out of these bundles in a number of subsamples, random as well as systematic. The remaining bundles, although harvested, were returned to the cultivators before threshing. In the fifth field, ten systematic subsamples of single handfuls were drawn out of all handfuls arranged in groups of ten on the ground after harvesting. For two of these subsamples, separate weights were recorded for each individual handful unit constituting the sub-samples. But for others, all handfuls in a subsample were pooled up, threshed and weighed as a whole. The results in 1964 were more or less consistent with those in 1963 (Table A.9).

TABLE A.6. LOSS OF GRAINS IN THE SELECTION OF 'DBL' UNITS BY SPLITTING FULL HEADLOADS IN THREE SUCCESSIVE STAGES AND THE STEADINESS ACHIEVED IN SPLITTING, GIRIDIH 1963

team no.	number of headloads	p.c. weight of grains shed	ratio of headload weights to the weight of 'DBL' units	
			mean	c.v.
(1)	(2)	(3)	(4)	(5)
1	39	0.47	8.50	26.3
2	24	.52	7.95	19.8
3	26	.75	7.97	16.8
4	33	.50	8.09	4.2
5	30	.55	8.18	17.5
6	35	.66	7.90	16.3
7	38	.69	8.06	20.3
all	225	0.59	8.11	19.0

TABLE A.7. ANALYSIS OF THE STAGE VARIANCES OF YIELD IN GRAMMES PER SAMPLING UNIT, GIRIDIH, 1963

stage	d.f.	variance		coefficients of true variation
		observed	true	
(1)	(2)	(3)	(4)	(5)
(a) field 1 : 2 handfuls out of each headload of 32 handfuls (mean = 82.5)				
1. between headloads	224	715.79	157.00	15.2
2. within 'h.l.' between 'h.f.'	225	401.68	401.6	24.3
3. unistage between handful units	449		558.60	28.6
(b) field 3 : 2 'DBL' units of 1 in $(2)^2$ out of each headload of 16 handfuls (mean = 226.6)				
1. between headloads	224	4805	1552.36	17.4
2. within 'h.l.' between 'DBL' units	225	1790	1790.14	18.7
3. unistage between 'DBL' units	449		3342.50	25.5

TABLE A.8. MEAN YIELD RATES (gms.) PER 'STANDARD' HEADLOAD OF 32 HANDFULS (H.F.) AND THE COEFFICIENTS OF VARIATION WITHIN INDIVIDUAL FIELDS, AS OBTAINED WITH DIFFERENT SAMPLING UNITS, GIRIDIH, 1963

field no.	area in acres	total number of headloads	yield rate* expressed in grammes of paddy per 'standard' headload: based on h.f. units $\bar{x}_1$, 2 per headload (h.l.)	based on 'DHL' units $\bar{x}_2$, 2 per headload (h.l.)	full headloads**
(1)	(2)	(3)	(4)	(5)	(6)
			(a) mean yield		
1.	.688	225	2619	—	2627
2.	.636	411	—	—	3322
3.	.684	225	—	3625	3676
			(b) coefficients of total variation		
1.	.688	225	28.6	—	20.3
2.	.636	411	—	—	20.6
3.	.684	225	—	25.5	18.4

*green paddy in fields 1 and 2, but dry paddy in field 3.

**of 32 handfuls in field 1 and 16 in fields 2 and 3.

TABLE A.9. MEAN YIELD PER BUNDLE WITH THE COEFFICIENTS OF VARIATION BASED ON RANDOM AND SYSTEMATIC SAMPLES USING SINGLE HANDFUL AND BUNDLES OF EIGHT AS SAMPLING UNITS, GIRIDIH, 1964

field no.	sample unit	sub-sample number	number of sample units	random sub-samples: mean in gms. per bundle of 8 h.l.	random sub-samples: coefficient of variation	systematic subsamples: mean in gms. per bundle	systematic subsamples: coefficient of variation
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	bundle of 8 handfuls	1	16	409	42.1	464	37.4
		2	16	397	38.7	398	32.7
		total	32	403	39.9	431	35.9
2	bundle of 8 handfuls	1	25	655	19.2	638	16.9
		2	25	673	15.4	669	19.6
		total	50	664	17.2	654	18.3
3	bundle of 8 handfuls	1	25	736	15.3	688	15.9
		2	25	731	19.7	692	19.0
		total	50	733	17.4	690	17.3
4	bundle of 8 handfuls	1	25	673	19.0	—	—
		2	25	622	18.6	—	—
		3	25	617	20.5	—	—
		4	25	663	16.8	—	—
		total	100	644	18.8	—	—
5	sub-samples of 37 handfuls	$S_1 - S_{10}$	10	—	—	610	4.6
	single handfuls	S_9	37	—	—	610	24.1
		S_{10}	37	—	—	600	27.6
		total		—	—	624	25.8

Note: Weight of dry paddy in field 1, green paddy in fields 2-5.

REFERENCES

HUBBACK, J. A. (1927) : Sampling for rice in Bihar and Orissa, Imperial Agricultural Research Institute, Pusa, Bulletin No. 166, and reprinted in *Sankhyā*, 7(3), 1946, 281-294.

MAHALANOBIS, P. C. (1939) : Statistical Report on crop-cutting experiments on Jute (published by the Indian Central Jute Committee, 1940).

——— (1940) : Statistical Report on crop-cutting experiments on jute in Bengal (published by the Indian Central Jute Committee, 1941).

——— (1944) : On large scale sample survey. *Phil. Trans. Roy. Soc.*, 231, Series B (584), 329-451.

——— (1946) : Sample survey of crop yields in India. *Sankhyā*, 7(3), 269-280.

——— (1946) : Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Stat. Soc.*, 109(4), 326-378.

——— (1946) : Use of small size plots in sample surveys for crop yields. *Nature*, 158 (4022), 798-799.

——— (1955) : Report on Bihar Crop Survey, Rabi season 1943-44 (submitted to the Government of Bihar in November 1944) and published in *Sankhyā*, 7(1), 1945, 29-106.

SUKHATME, P. V. (1944-45) : Report on the second random sampling surveys for estimating the out-turn of wheat in the United Provinces, 1944-45.

——— (1945-46) : Report in the random sample surveys for estimating out-turn of paddy in Madras, 196.

——— (1946) : Size of sampling unit in yield surveys. *Nature*, 158, 345.

——— (1946) : Bias in the use of small size plots in sample surveys for yield. *Current Science*, 15, 119-120.

——— (1947) : Use of small size plots in yield surveys. *Nature*, 160, 642.

——— (1947) : The problem of plot size in large scale yield surveys. *J. Amer. Stat. Assoc.*, 42, 297-310.

SENGUPTA, J. M. and MAHALANOBIS, P. C. (1951) : On the size of sample cuts in crop-cutting experiments in the Indian Statistical Institute, 1939-1950. *Bull. Int. Stat. Inst.*, 33(2), 350-404.

SENGUPTA, J. M. (1964) : On perimeter bias in sample cut of small size. *Sankhyā*, Series B, 26, 53-68.

GOVERNMENT OF INDIA : Some results of the Land Utilisation Survey and Crop-Cutting Experiments (NSS Thirteenth Round : September 1957-May 1958).

Paper received : August, 1965.

ON OPTIMUM PAIRING OF UNITS

By V. K. SETHI

Agra University, India

SUMMARY. The paper gives the best method of selecting two units from a population for estimating the population total (or mean) under the restriction of given expected number of occurences of different units in the sample. This involves only the arrangement of units in increasing order of the value of the variable divided by the given expected number of occurrences. Some modifications of the method and some extension for larger clusters are also suggested.

1. Introduction

Cluster sampling is one of the well-known and widely used techniques of sampling. The population is divided into a number of clusters of units and one or more of these clusters are drawn at random. Generally by cluster we understand a set of units such that for any proper sub-set there is at least one more unit in the set which is adjacent to it in location. In this paper, however, the word 'cluster' will be used for any set of units whether it satisfies the above mentioned property or not. The problem considered here is to form clusters so that, given the method of estimation and the loss function, the risk is minimized. An exact solution is obtained for clusters of two units. This solution finds an immediate application for selection of a pair of units with varying probabilities.

The method suggests a different approach to the problem of stratification when a pair of units are to be drawn from each stratum in an optimum manner. This consists in dividing the population into symmetric sub-populations.

Repeated use of the method for different characters, or its combination with stratified sampling with probabilities proportional to sizes may provide a satisfactory solution for the problem of multi-character surveys.

No simple solution exists for optimum formation of clusters of more than two units. Some methods are, however, suggested which are nearly optimum.

2. Optimum pairing for a population of four units

We start by considering a population with only four units (U_i), $i = 1, 2, 3, 4$. By X_i we denote the value of the character x for U_i. The subscripts are so assinged that

$$X_1 \leqslant X_2 \leqslant X_3 \leqslant X_4 \qquad \ldots \quad (2.1)$$

and let

$$\sum_{i=1}^{4} X_i = X. \qquad \ldots \quad (2.2)$$

There are only three different ways of dividing this population into two pairs of units

(1) (U_1, U_2) and (U_3, U_4) : pairing P_1

(2) (U_1, U_3) and (U_2, U_4) : pairing P_2

(3) (U_1, U_4) and (U_2, U_3) : pairing P_3.

If, after cluster formation, one of the pairs is to be selected with equal probabilities and the population total X estimated by the usual unbiased estimator, then the absolute error of estimate for the pairing P_i is given by

$$e_i = |X^{(i)} - X|$$

where $X^{(i)}$ is an estimate of X based on the i-th pairing. Clearly

$$\begin{aligned} e_1 &= |(X_3+X_4)-(X_1+X_2)| \\ e_2 &= |(X_2+X_4)-(X_1+X_3)| \\ e_3 &= |(X_1+X_4)-(X_2+X_3)|. \end{aligned} \qquad \ldots \quad (2.3)$$

It may be mentioned that these absolute errors are independent of the particular pair selected. It can be easily verified that

$$e_3 \leqslant e_2 \leqslant e_1. \qquad \ldots \quad (2.4)$$

Let $L(e)$ denote the loss corresponding to an absolute error e. If it is a monotonically increasing function of e then obviously

$$L(e_3) \leqslant L(e_2) \leqslant L(e_1). \qquad \ldots \quad (2.5)$$

If we denote the expected loss over the pairs of P_i by $R(P_i)$, then for a population of four units

$$R(P_i) = L(e_i) \qquad \ldots \quad (2.6)$$

and the risk is the minimum for pairing P_3. In particular if the risk function is proportional to the variance of the estimator of X, $L(e) \propto e^2$ is a strictly increasing function of e. Thus P_3 leads to the minimum variance.

3. Optimum pairing for populations with an even number of units

Let us consider a population of $2N$ units. This is to be divided into N clusters of 2 units each, one of the clusters selected at random with equal probabilities and the population total X estimated by

$$\hat{X} = N(x_1+x_2) \qquad \ldots \quad (3.1)$$

where x_1 and x_2 are the values of the character x for the two units of the selected cluster. The problem is to find the optimum pairing in the sense of minimizing the risk corresponding to a monotonically increasing convex downward or linear loss function of the absolute error.

Let us denote by X_i the value of the character x for the i-th unit U_i, $i = 1, 2, \ldots, 2N$. The subscripts are given in such a way that

$$X_1 \leqslant X_2 \leqslant X_3 \leqslant \cdots \leqslant X_{2N}. \qquad \ldots \quad (3.2)$$

Let us consider a pairing P where two of the pairs are (U_1, U_i) and (U_j, U_{2N}). Corresponding to P we can find a pairing P' which has pairs (U_1, U_{2N}), (U_i, U_j) and the rest of the pairs the same as in P. Then for comparing the risks for P and P', we have to consider the losses corresponding to the pair formed by U_1, U_i, U_j, and U_{2N} alone. As we are not estimating the total of these four units we cannot apply the result of Section 2 directly.

Let e_{ik} denote the absolute error of the estimate based on the pair (U_i, U_k). Then

$$\begin{aligned} e_{ik} &= |N(X_1+X_i)-X| \\ &= |\{(N/2)(X_1+X_i+X_j+X_{2N})-X\}-(N/2)\{(X_j+X_{2N})-(X_1+X_i)\} \\ &= |A-e_1| \end{aligned} \quad \ldots (3.3)$$

where $$A = (N/2)(X_1+X_i+X_j+X_{2N})-X$$

and $$e_1 = (N/2)\{(X_j+X_{2N})-(X_1+X_i)\}.$$

Similarly,

$$e_{j2N} = |A+e_1| \quad \ldots (3.4)$$

$$e_{12N} = |A-e_2| \quad \ldots (3.5)$$

$$e_{ij} = |A+e_2| \quad \ldots (3.6)$$

where $$e_2 = (N/2)\{(X_1+X_{2N})-(X_i+X_j)\}.$$

It follows from Section 2 that $|e_2| \leqslant |e_1|$. Thus if the loss function L be such that for all values and for $|x| \leqslant |x'|$

$$L(|A+x'|)+L(|A-x'|) \geqslant L(|A+x|)+L(|A-x|) \quad \ldots (3.7)$$

then clearly the risk corresponding to P' is not greater than the risk corresponding to P. Thus for such a loss function we can stipulate that the optimum pairing should have (U_1, U_{2N}) as one of the pairs. By following the same arguments repeatedly it is easy to see that it is not possible to reduce the risk to less than the risk corresponding to the pairing

$$P_0 : (U_i, U_{2N-i}), \quad i = 1, 2, \ldots, N. \quad \ldots (3.8)$$

P_0 can thus be termed as the optimum pairing for all loss functions satisfying (3.7). It can be verified that (3.7) is satisfied for all non-decreasing convex downward loss functions. In particular $L(e) = ke^2$ is such a loss function and the risk corresponding to it is proportional to the variance of the estimator of population total. Thus of all the pairings P_0 leads to minimum variance.

The optimum method of forming pairs can be described as follows :

Arrange the units in increasing order of magnitude of the character under study. Form pairs of units which are equi-distant from the two ends in this arrangement.

In practice one would only approximate the optimum by arranging the units in increasing order of some character for which values of all the units in the population are known. This is not a drawback for this problem in particular but arises in all problems of optimization like optimum allocation and optimum stratification. The closer the relationship between the variable used for pair formation and the variable under study the better is the approximation to the optimum. It is not necessary for this relationship to be linear regression. It would be enough if the regression of the variable under study on the pair-formation variable is a monotonic function of the latter.

It should be remarked that if the population were symmetrical the risk would reduce to zero. If one has decided to use optimum pairing for selection of units within strata it would help to form strata by dividing the population into nearly symmetrical populations.

4. Selection of a pair of units with varying probabilities

Let there be N units in the population (U_i), $i = 1, 2, ..., N$. One has to select a pair of units with the restriction that the expected number of times the i-th unit occurs in the sample is the prescribed value w_i, $i = 1, 2, ..., N$. There is no restriction that the unit should not be repeated. The problem is to find the optimum set of values w_{ij} where w_{ij} represents the probability of selecting the pair (U_i, U_j).

Clearly,

$$\sum_{j \neq i} w_{ij} + 2w_{ii} = w_i \qquad i = 1, 2, ..., N. \qquad \ldots (4.1)$$

We shall consider that set w_{ij} as optimum which leads to the minimum risk for an estimator of the type

$$\hat{X} = \frac{X_i}{w_i} + \frac{X_j}{w_j}. \qquad \ldots (4.2)$$

It will be assumed that w_{ij}'s are all rational numbers. This restriction is necessary because no known method can assure some given irrational probabilities of selection. Let

$$w_i = \frac{n_i}{d_i}, \qquad i = 1, 2, ..., N \qquad \ldots (4.3)$$

be the given set of w_i's, and L be a number divisible by all d_i's. Let us associate Lw_i sub-units with U_i for all i. If X_i be the value of the character x for U_i, impute a value X_i/Lw_i to each of the sub-units corresponding to U_i. Thus the total of the imputed values is also $X = \sum_{i=1}^{N} X_i$. The total number of sub-units is $2L$. The total X can thus be estimated from a pair of sub-units selected with equal probabilities. The estimate would be of the type (4.2), and expected number of sub-units corresponding to U_i in a pair selected at random is w_i.

If the loss function be non-decreasing convex-downwards, the best method is to arrange the sub-units in increasing or decreasing order of their assigned values and take the optimum pairing as described in the previous section. The whole method can be summarised as below :

(1) order the units in increasing magnitude of X_i/w_i ;

(2) find the cumulative totals $S_i = L \sum_{j=1}^{i} w_j$, where Lw_i is an integer for all values of i ; put $S_0 = 0$;

(3) select a number r at random from 1 to L with equal probabilities ;

(4) if $S_{i-1} < r \leqslant S_i$, select U_i ;

(5) if $S_{j-1} < 2L+1-r \leqslant S_j$, select U_j.

The probabilities resulting from this procedure will form an optimum set. The remarks at the end of Section 3 apply to this method also. In practice we can only approximate the optimum set of probabilities.

In considering the problem in this section we have allowed the possibility of having either one or two units in a cluster. With this relaxation of conditions, it may be possible to improve upon the optimum clustering derived in Section 3. This improvement consists in making the population of sub-units more symmetric by (i) addition of dummy sub-units, (ii) arbitrary increment in the value of some of the units, or (iii) unequal assignment of values to the sub-units.

5. Use of optimum pairing in multipurpose surveys

No completely satisfactory sampling design exists for multipurpose surveys. The design most suitable for one character may not be very good for another. Several methods for controlling errors exist and they may be used in an optimum manner for different variables. One may use the best method of stratification for one variable, the most suitable size for deciding the probabilities of selection for a second character and optimum pairing for the third.

In deep stratification, the strata homogeneous with respect to one character are further sub-divided to make the sub-strata homogeneous with respect to another character. A similar technique may be adopted in pair formation. Initially optimum pairs may be formed with respect to one variable; then optimum pairs of these pairs may be formed for another variable, and so on.

6. Formation of larger clusters

From the following example it would appear that the knowledge of order is not sufficient for optimum formation of clusters of size more than two.

Two populations each of six units are to be divided into two clusters of three units each and one of the clusters selected with equal probabilities for estimating the population total. The units are arranged in increasing order of values of both the characters under study.

	X_1	X_2	X_3	X_4	X_5	X_6
population I	1	2	4	8	16	32
population II	1	2	3	4	5	6

The optimum clustering for population I is (U_1, U_2, U_3) and (U_4, U_5, U_6). This clustering is not the optimum for population II. Thus considerations other than order must enter optimum formation of clusters of size larger than two.

Suppose there are Nn units in the population which is divided into N clusters of n units each. Let two of the clusters have the following values of the character x.

$$\text{Cluster } C_i \qquad X_{i1} \leqslant X_{i2} \leqslant \cdots \leqslant X_{in}$$

$$\text{Cluster } C_{i'} \qquad X_{i'1} \leqslant X_{i'2} \leqslant \cdots \leqslant X_{i'n}. \qquad \cdots (6.1)$$

If

$$X_{ij} < X_{i'j}, \quad j = 1, 2, \ldots, n \qquad \cdots (6.2)$$

then by exchange of any one U_{ij} with $U_{i'j}$ in the two clusters would obviously lead to a reduction in risk. A clustering which does not satisfy (6.2) for any pair of clusters will be called a *nearly optimum clustering*.

A simple method of getting a nearly optimum clustering is the following.

(1) Arrange the units in increasing order of magnitude of the variable under study.

(2) Divide them into n homogeneous strata of N units each. Thus the first N units in the arrangement form the first stratum, next N units the second stratum and so on.