F.D.101
2000

THE NATIONAL SAMPLE SURVEY

NUMBER 5

TECHNICAL PAPER ON SOME ASPECTS OF THE DEVELOPMENT OF THE SAMPLE DESIGN

By

D. B. LAHIRI

INDIAN STATISTICAL INSTITUTE, CALCUTTA

NSS 12
5
6/8/91

Issued by

The Department of Economic Affairs
Ministry of Finance : Government of India

March, 1954

For use of the Indian Statistical Institute

THE NATIONAL SAMPLE SURVEY

NUMBER 5

TECHNICAL PAPER ON SOME ASPECTS OF THE DEVELOPMENT OF THE SAMPLE DESIGN

By

D. B. LAHIRI

INDIAN STATISTICAL INSTITUTE, CALCUTTA

NSSD 5
8/8/91

Issued by

The Department of Economic Affairs

Ministry of Finance : Government of India

March 1954

Note of Caution

Being the scanned copy of old NSS report, this document may suffer from following limitations -

i. Poor Quality of the Scanned images.
ii. Page(s) missing in between.
iii. Improper sequencing/arrangement.

THE NATIONAL SAMPLE SURVEY

NUMBER 5

TECHNICAL PAPER ON SOME ASPECTS OF THE DEVELOPMENT OF THE SAMPLE DESIGN

CONTENTS

THE NATIONAL SAMPLE SURVEY

NUMBER 5

TECHNICAL PAPER ON SOME ASPECTS OF THE DEVELOPMENT OF THE SAMPLE DESIGN

The present report is being published in the form in which it was submitted to the Government of India. The views in the notes are not necessarily those of the Government of India.

INTRODUCTION

0.1. The National Sample Survey (NSS) was initiated by the Government of India in 1950 at the instance of the Prime Minister, Jawaharlal Nehru, and has been working continuously since then in the form of one round of survey after another. The first round of field work lasted from October 1950 to March 1951 and covered the rural areas of India. The second round (April–June 1951) also covered only the rural areas. From the third round (August–November 1951) the urban areas (including the four big cities of Bombay, Calcutta, Delhi and Madras) are being covered in every round. The seventh round was about to progress at the time of writing this report. The bulk of the field work is being done under the direct control of the Department of Economic Affairs, Ministry of Finance.

0.2. The statistical work of the NSS (including the preparation of the sample design and schedules, the processing and tabulation of the data, and the preparation of the reports) is being done in the Indian Statistical Institute (ISI). The first report (NSS No. 1) was published in December 1952. Since then three more reports prepared by the Institute have been published[1], and other reports are in progress. In addition, special surveys have been conducted and reports and data in tabular form have been supplied from time to time to different agencies[2].

---

1 a National Sample Survey No. 2: Tables with Notes on the Second Round : April-June 1951 (Government of India, December 1953).

b National Sample Survey No. 3 : Tables with Notes on the Third Round : August-November 1951 (Government of India, March 1954).

c National Sample Survey No. 4 : Survey of Persons in the Live Register of the Delhi Employment Exchange (Government of India, March 1954).

2 a The Fact-finding Committee appointed by the Ministry of Rehabilitation (survey of economic condition of refugees in West Bengal), March 1953.

b The Press Commission (habits of newspaper reading), August 1953.

c The Taxation Enquiry Commission (household consumption by expenditure levels), December 1953.

d The Ministry of Rehabilitation (survey of refugees in Bombay), February 1954.

0.3. The sample design used in the first round was naturally based on the previous experience of the Indian Statistical Institute ; but efforts are being continually made to improve the design in the light of the experience gained from one round of the survey to another. Also, needs are changing and new requirements are arising from time to time. In fact, the sample design of a continuing enterprise like the National Sample Survey must be essentially dynamic and evolutionary in character.

0.4. Shri Debabrata Lahiri, the author of the present report, has been in charge of the sample design of the NSS from the beginning, and he has given in the present report a general account of the development of the sample design (without, however, entering into the mathematical details which will be given later). This report explains broadly the considerations which led to the development of the present design, and also gives some idea of future lines of progress. Problems of estimation of sampling and non-sampling errors have also been briefly discussed. The method of independent interpenetrating samples are being regularly used; and a large volume of material on the margin of uncertainty has accumulated which will be discussed later in a technical report.

0.5. The following background information will be useful in studying the present report. For administrative purposes India is divided into 29 States (some of which are very small), about 300 districts, about 2,500 *tehsils* (or equivalent units), 3,000 towns and about 586,000 villages in round figures. The total area of India is 1.26 million square miles (3.27 million square kilometers) broken up into some hundreds of millions of parcels or "plots" of land ; and the 1951 census population was 360 millions of people (70 millions of households) of whom 60 millions in round figures live in towns and 300 millions in villages. Turning now to the livelihood pattern we find that less than one-third of the population are self-supporting and of these again less than a third derive their income principally from non-agricultural sources.

0.6. As regards the sampling design, it can be stated very briefly that the rural area of the country was divided into about 250 geographical strata from which about a thousand sample villages were selected. In the first three rounds the villages were directly selected within a stratum but in subsequent rounds two *tehsils* were selected in each stratum (with replacement) and two villages were selected within each sample *tehsil* (an administrative unit of about 500 square miles or 1,300 sq. kilometers on an average). Finally, a sample of households was taken up for the household enquiry and a sample of clusters of plots for the land utilisation survey. From the third round the survey was extended to urban areas with stratification of towns by size, and then the selection of a sample of census enumeration blocks (with replacement) with probability proportional to population, and finally a sample of households within each sample block.

0.7. A general idea of the nature of information collected is given below:

(a) *Sample villages :* general economic information, and weekly prices of selected commodities ; rates of daily wages of skilled and unskilled workers etc. ;

(b) *Households:* (1) *General particulars:* age, sex, marital status; economic and employment status; births, deaths etc.; details regarding holdings, use of land under various categories; live-stock, real assets, loans, savings, housing conditions etc.;

(2) *Consumer expenditure:* on a very large number of items ;

(3) *Household enterprises:* agriculture and animal husbandry, acreage and production of different crops; particulars on industry, crafts and trade, including fixed capital, machinery and tools, fuel, power, raw materials, quantity and value of production, source of finance etc.;

(c) *Utilization of land:* survey of a sample of revenue plots;

(d) *Crop survey :* crop acreage and estimates of the yield of crop per acre by direct crop-cutting experiments; and

(e) *Sample survey of manufacturing establishments:* (with 10 operatives or more with power or 20 operatives or more without power) covering practically all groups of industry over the whole of India.

0.8. A general purpose design is used for the household sample (b); and usually, separate designs and some separate staff are used for specific projects under (c) and (d). A different design and an entirely separate staff is used for (e). One of the basic problems of design is to have a closer integration of (b), (c) and (d). A preliminary discussion has been given in this report. Much research is needed on the design of sample surveys, and the present report will be useful in indicating some of the problems which are of basic importance.

27 February 1954 *P. C. Mahalanobis*

1. GENERAL OBSERVATIONS

1.1. The main object of the present paper is to give a broad general account of the development of the sample design of the National Sample Survey (NSS) since the beginning of the first round of the survey in October 1950. It is not intended, however, to enter into a thorough discussion of the reasons which initially led us to the adoption of that part of the present sample design which forms its main structure. Nor shall we discuss, in as much detail as we would have liked, the scope and concrete proposals for future improvements in the sample design. So far we have been primarily concerned with the consolidation, both in the field and in the laboratory, of our position within the broad framework of the main structure already current. This course should cause no surprise to those who have handled large-scale sample surveys, particularly those of a general purpose type. Another reason for paying comparatively less attention to the question of improvement is that with the prospect of obtaining shortly very informative volumes on the 1951 Census of Population, in which basic information on the economic structure of the population will be available for the first time for census-tracts and even to some extent for such small units like villages, the problem of improving the design will undergo substantial changes.

1.2. Keeping the broad structure in view we shall generally examine what improvisations or what modifications have been made or are proposed to be made in the sample plan in the immediate future, including the question of estimation, both of the characteristics under study as well as their sampling error, so as to meet the current needs which a survey of the dimensions of the NSS must face. We shall not try to cover all the points, but shall restrict ourselves to a selected few, and that also without entering into exhaustive details or full mathematical discussions. It must be added, however, that investigations even in a restricted field also have their repercussions on the main structure of the sample design, which will become evident in the following pages.

1.3. We shall make in this and the next few paragraphs some observations on the broad structure of the sample design which has been already briefly described in NSS No. 1 : General Report on the First Round (October 1950—March 1951)[3]. It will be remembered that, generally speaking, in the first three rounds the country was divided into a number of geographical strata within which a number of villages were selected with replacement with varying probability. The number of villages (more than 2,000 on the average) within each stratum was so large compared to the number included in the sample that the chance of repetition of the same village was extremely small. The "with-replacement" scheme was, however, used because of certain great advantages in the estimation of sampling variance (discussed in a later section on sampling with replacement). Within every selected village, a sample

[3] A list of papers and reports prepared by members of the staff of the Indian Statistical Institute will be found at the end of this paper.

of households was taken up for the household enquiry and a sample of a cluster of plots for the land utilization survey.

1.4. In subsequent rounds this structure of the design remained broadly the same for the rural areas excepting that a higher order of multi-stage sampling was introduced mainly for practical convenience as explained in para 4.15. Here *tehsil* was the first stage unit, two *tehsils* being selected from each stratum ; and within each selected *tehsil* , two villages and then an appropriate number of households etc. were chosen as before[4]. It will, however, be found in our subsequent discussions that in a sense this higher order multi-stage scheme was already present in a hidden form in the initial sample-design. Here again the *tehsils* were selected with replacement for reasons stated in the previous paragraph.

1.5. It is intended to go back to the direct selection of villages within a stratum in future rounds. It is felt that in a multi-purpose survey of the type of NSS where information on widely different kinds of items are being collected, (see paragraph 0.7), and where in addition to the estimates of national aggregates etc. the question of providing estimates for at least some of the more important characteristics for each of the six "Population Zones" (1951 census definition)[5] or, smaller sub-regions of the country (like States or State-groups) has to be kept in view, and where it is intended to set up a sampling organization which shou d be able to cater for a special need as and when occasion arises (as for example, the unemployment survey and the news-reading survey actually taken up in 1953) the sample villages should be scattered over the entire country as widely as possible, and therefore the introduction of another stage viz., the *tehsil* with two (or more) sample villages in each selected *tehsil* should be dispensed with at the earliest possible moment, even if it means some additional difficulties in field operations.

1.6. The main effect of the elimination of large first-stage units like *tehsils* would appear to be an increase in travel time. But it is important to point out that the increase is not as much as one would expect on first thought. Whatever may be the condition in countries with a highly developed transport system and with extensive camping facilities, the rural areas of India are so undeveloped in matters of communication that it is not unusual to come across the paradoxical situation that under certain circumstances sample villages separated by long distance are much "nearer in time" and more convenient of approach as regards "physical exhaustion due to travelling" than villages which are apparently quite close to one another.

1.7. In the former case when villages are favourably situated in respect of train or bus service the distance can be covered comparatively easily and within a short period of time. In the latter case one has some time to plod the whole

---

[4] There are on the average 10 *tehsils* per stratum, 225 villages per *tehsil*, and 100 households per village.

[5] *Census of India*, Paper No. 2, 1952 : Population Zones, Natural Regions, Sub-Regions and Divisions. (Government of India, Ministry of Home Affairs.)

distance on foot, or wait for a chance bullock-cart which may or may not give him a lift, and all this possibly in tracts with literally no roads and where movement has to be more or less restricted during the day time. Moreover, the investigator may have to make very large detours to avoid forests, hills, swamps or rivers. It is true that two villages separated by a long distance need not necessarily be fully covered by quick means of transport, but then with the increased scope of availing such means of tansport, by far the larger portion will be usually covered in a short time compared to that taken for the slow-moving portion.

1.8. Thus, unless one contemplates taking within each sample *tehsil*[6] a fair number of villages, say four (or a multiple of four) which is the number an investigator has to cover in a single round there will be no sizeable saving in travel time over that required in the case where all the villages are selected directly within a stratum and therefore flung more widely apart on the average. But even then the saving in travel time (and travel costs[7]) made by bringing down drastically the number of first-stage units is not at all likely to be commensurate with the loss of efficiency particularly when one remembers the requirements set down earlier in para 1.5.

1.9. We have put forward reasons why we have favoured a widely scattered sample of villages. We are now naturally led to the question whether it would not be possible to gain in efficiency by increasing (or decreasing) the number of villages and by making the corresponding change in the number of households to be sampled from each village in order that the total cost of the survey may remain unaltered.

1.10. Before taking up this question it is important to bear in mind one peculiar feature of repetitive surveys of the type of NSS. In a non-repetitive survey one has a good deal of freedom of fixing the size of the investigating team; for example, by employing a larger contingent of investigators one can complete the survey in a shorter period. But in a repetitive survey where the investigators are to be kept continuously in service[8] on a monthly salary this flexibility is lost, for the duration of the survey or rather any round of the survey is more or less fixed, with the consequence that the number of investigators is also more or less fixed thereby, the total cost of the survey of course remaining the same for each round.

1.11. The bearing of this aspect on the question of changing the number of sample villages is that the change, if any, has to be brought about by big jumps. Thus, in the current plan, 240 investigators are engaged on the rural part of the survey

---

6 The average area of a *tehsil* is 500 square miles (1,300 sq. kilometers) while that of a stratum is 5,000 square miles (13,000 sq. kilometers).

7 For convenience the investigators are allowed a fixed travelling allowance (T.A.) which accounts for a comparatively small percentage of the expenditure on field operations. There is, therefore, not much scope for economy. Moreover for administrative reasons it may be difficult to reduce the fixed T.A. even if an investigator's work is restricted to a single *tehsil*.

8 It is obviously desirable not to lose the advantage of the experience gained by the investigators in any round; so a quasi-permanent team of investigators is desirable. Recruitment and disbandment of the investigating team separately for each round is not only difficult but undesirable and, of course, more expensive because of the time lost in fresh recruitment, training, and gaining of experience.

and each investigator covers four villages in each round so that altogether 960 villages are in the sample in a single round. One cannot change the number of investigators and therefore the "smallest" change that one can think of is to assign five or three villages to each investigator (the assignment of a uniform number of villages to an investigator being, of course, a practical necessity) so that we have to go right up to 1,200 or go down to 720 villages from the current size of 960 villages. The implication is that even if considerations of variance and of cost functions lead in a non-repetitive survey to an optimum sample size (for villages) other than 960, even then it cannot be given effect to in a repetitive survey like the NSS.

1.12. Even when the "optimum" size falls in line with the above restrictions there are other considerations of great practical importance. An increase in the number of sample villages is fraught with certain difficulties. Perhaps the greatest drawback is the fact that the investigator is a human being. He cannot be expected to move very frequently from one village to another, which would be necessary if the total number of sample villages (to be covered by him in each round) were increased, especially under the very trying conditions prevailing in the rural tracts of the country. He stays at present for about three weeks in each sample village; and it may be inconvenient to make any substantial decrease of the period of his stay in a village (in order to increase the number of sample villages by reducing the work-load in each village so that the total cost of the survey may remain unaltered).

1.13. Consideration of the human factor is not only important for the investigator but also for the investigated or interviewee. An investigator's stay for a reasonably long period is conducive to the establishment of good relations between the investigator and the villagers. Within limits, the longer he stays in a village the easier it will be for him to secure the necessary cooperation of the villagers and better his chances of collecting more complete and more accurate information by call-backs, if necessary, of the appropriate persons who are in a position to give the correct information.

1.14. As to the prospects of decreasing the number of sample villages to 720 (at most) we have earlier explained how, because of the multi-purpose nature of the survey, it is desirable to scatter the sample villages as widely as possible, and the same reasons also speak in favour of not decreasing the number of sample villages.

1.15. It must, however, be admitted that some portions of these general considerations, explained in the earlier paragraphs, border somewhat on the side of speculation. But such speculations are necessary when quick decisions are to be made, as is often the case in large-scale sample surveys, and therefore these have a distinct place in the evolution of any large-scale sampling enquiry.

1.16. The question of optimum allocation of our resources which, of course, involves the questions of optimum use of all supplementary information in the frame already available (or likely to be available shortly like that provided by the 1951

population census), in the construction of strata, or in the assignment of probability of selection, in determination of shape and size of sampling-unit and so on, is intended to be studied much more thoroughly than has been possible so far. This is a complex problem for a general purpose continuing survey where one has, moreover, to reckon with the boundary conditions already set by the pattern of the initial sample-design (whatever may have been the reasons for choosing it) as it has more or less laid down the pattern of field organization for which some of the multifarious difficulties have been already overcome; to renew these difficulties for the sake of even an otherwise better sample design (calling for considerable changes in this pattern) may be highly undesirable.

1.17. Even leaving apart the field point of view there is the fundamental difficulty of defining what, in some useful practical sense, is meant by an optimum design where widely different (from the design point of view) groups of items, some of which are new or initially unknown but none-the-less quite important, are of interest. Informations on these items are certainly known to be required at the national or Union level but the possibility of supplying figures relating to more important items in respect of different population zones, or different occupational groups, or expenditure classes or enterprises etc., have also to be kept in view. Moreover, we are not only interested in certain population "totals" (e.g., total production of cereals, total consumer expenditure, total cost of fuel and power consumed by household industrial establishments, etc.) but also equally in some population "ratios" (e.g., per capita consumption of cereals, average size of an agricultural holding, cost of cultivation per acre), and in some cases in the distribution of certain variates (such as, distribution of household by size of expenditure, or by size of land owned, or by size of operational holding in agriculture etc.). It must be admitted we have not so far been able to tackle this formidable problem systematically, and have proceeded more or less on semi-intuitive grounds based on previous experience. We now turn to more specific problems where some progress has been made towards their solution.

2. STRATIFICATION AND INTEGRATION OF DIFFERENT ENQUIRIES

2.1. We have referred to the difficulty of integrating somewhat opposing demands made by widely different (from the design point of view) groups of characteristics which are under investigation. We shall here restrict ourselves to only one such problem. This is the problem of drawing up an integrated scheme for the household enquiry and the land-utilization survey. We have not yet solved all the problems raised in this issue of integration; but we shall present here an account of whatever progress that has been made in this direction. We shall limit ourselves

to one of the main difficulties arising from the fact that whereas "population" is naturally the basic consideration for the household enquiry, the "area" is so for the other[9].

2.2. The problem of integration not only arises in the cases referred to in the preceding paragraph but also between different types of household enquiries, for example, between enterprise and consumer expenditure; and even between different types of enterprise where widely different items of information are being collected, as will be evident on going through the facsimile field schedules published in previous NSS Reports Nos. 1, 2 and 3. It is hoped that with the publication of the 1951 census data where for the first time the economic structure of the population by census tracts and units as small as villages will be available, we shall be in a better position to have a closer integration of the different household enquiries than has been possible so far.

2.3. The problems of allocation of sample units to the different strata in the two cases are in a sense not very different. It is true that allocation of units in proportion to the population does not ensure allocation in proportion to area. But if lands which are intimately connected with the activity of the population are given importance over others lying waste or receiving scanty attention, then allocation on the basis of population should give fairly good results. It is, however, not true that information on land not utilized at present is of minor importance ; for example, information on cultivable waste is of great importance in the context of the urgent need of increasing the present supply of food.

2.4. In a multi-purpose survey one must, however, work on some compromise between opposing interests ; and in the first three rounds, the allocation of sample villages was made on the basis of population (with such minor adjustments as appeared to be necessary when the area was disproportionately large in relation to its population, or when a State quota became comparatively small on strict adherence to allocation by the population rule).

2.5. From the fourth-round onwards our approach has been different. Whereas in the earlier rounds it was stratification first and then allocation, from the fourth round it has become, so to say, allocation first and then stratification. It was decided to have a fixed number of sample units in each stratum, and then construct the strata suitably.

2.6. It was thought that in the construction of strata the equalisation of stratum—"sizes", associated with a single sample unit in each, would give optimum

---

9 It should be pointed out that in order to reduce the ascertainment error the information regarding land utilization should preferably be collected by direct physical observation of a sample of plots and not through enquiry from households. Accurate maps showing the boundaries of plots or fields are available. Reference may be made in this connexion to P. C. Mahalanobis : A sample survey of acreage under jute in Bengal, *Sankhya* 4, (1939), 511-30.

results[10]. Also, having an equal number of sample units in each stratum is of great advantage for other reasons which will be evident at the closing stages of the present paper. Finally, in order to supply valid estimates of the sampling error, it was decided to select not one, but two units (*tehsils*) from each stratum.

2.7. The "size" considered most suitable was the total consumer expenditure. But this size was more or less unknown for the *tehsils*, which were the administrative units which would ultimately be combined to form the strata. The only information known about consumer expenditure was that provided by the first round of the survey, and naturally for such small units there was no way of obtaining reliable estimates of the (*tehsil*) "sizes". What was actually done was to allocate the total number of sample units to large zones (State or State-groups) on the basis of the estimated total consumer expenditure as found in the first round and then to reallocate the units to the different "natural divisions"[11], defined by the census department for the 1951 census, in proportion to the population.

2.8. The introduction of "natural divisions" in the process of stratification was an improvement over the earlier procedure in which the country was divided into compact areas by more or less arbitrary combinations of districts, whereas in the formation of the natural divisions the census department took care to ensure homogeneity in respect of geological, climatic and cropping patterns, so that considerable improvements both in regard to land-utilization and household enquiry may be expected.

2.9. The natural divisions were further broken up into (ultimate) strata in such a manner that for any specified natural division the population of any stratum was roughly the same. The number of strata into which a natural division was broken up was exactly equal to half the number (care was taken at the allocation stage to round off to an even number) of first-stage units allocated to a natural division. Further details about the method of formation of the (ultimate) strata (which is of relevance in the assignment of variable probability in the selection of sample units) are explained in the next few paragraphs.

---

[10] This principle of strata of equal size is being used by the Indian Statistical Institute for a long time. For example, in the repeated surveys of land-utilization and crops in Bengal (at first twice, later three times a year) the whole State was divided into cells of equal size (approximately 64 square miles in area) from 1943. This approach has been found advantageous in practice but no entirely satisfactory theoretical discussion is available.

[11] The country as a whole is divided into 5 "Natural Regions", the boundaries of which are fixed solely with reference to physical features. These 5 Natural Regions are sub-divided into 15 "Sub-Regions" on the basis of substantial differences within each Natural Region, in respect primarily of rainfall and climatic conditions and also differences in soil so far as these are broadly identifiable and are reflected in the cropping pattern. The intersection of these Sub-Regions with States is the basis on which Natural Divisions in each State are fixed and related to the All-India scheme of Divisions. The 15 Sub-Regions are sub-divided into 52 "Divisions" each of which is either an entire State or a group of contiguous districts within a State. (*Census of India*, Paper No. 2, 1952)

2.10. The first-stage sampling units, namely, *tehsils* vary widely in size even within the same stratum whether this is judged by the population or by the area; and the selection of units with varying probability is desirable. Apart from certain refinements (which may in practice be somewhat inconvenient) one can say that for the household enquiry the units be better selected with probability proportional to the number (census) of households or population, and for the land utilization survey with probability proportional to area. Other possible methods of dealing with the problem will be first discussed before elaborating upon the method of selection by variable probability which was actually adopted.

2.11. Selection with equal probability, after stratification by size (population and /or area), associated with a simple unbiased estimate of stratum totals based upon the total number of units in the stratum, was one possibility. A modification, namely, stratification by one of the sizes (population or area), and selection of units with probability proportional to the other size (area or population) may give better results, but even this was not suitable for reasons which will be evident in a moment.

2.12. Now a high degree of geographical stratification was considered to be of special importance particularly because of the multi-purpose nature of the survey. But stratification by size in conjunction with geographical stratification would demand an unduly large size of the geographical super-strata if the stratification by size were to control effectively the size within reasonable limits. At the same time any actual (ultimate) stratum would in almost all cases be built up of disjoint *tehsils* distributed over the whole of this geographical area, with the result that the heterogeneity introduced thereby will substantially reduce the gain that would otherwise be introduced.

2.13. Moreover with the approach previously accepted there is really no freedom in the choice of the geographical super-strata, for we have already identified these with the natural divisions; nor do we have any choice in the number of (ultimate) strata for this also is predetermined. With these severe restrictions it is hardly possible to demarcate the (ultimate) strata in such a manner that the variation in size of *tehsils* within any stratum would be really small.

2.14. The construction of small geographical strata with two different sample units selected at random (with equal probability) from each stratum, and estimation of (universe) totals by the ratio method also had apparently great advantages. This would make integration of the two types of enquiry (household and land utilization) comparatively easy; for the former the (census) population and for the latter the area need only be used in the denominators of the ratios.

2.15. There is no question of estimating the (universe) total by using a weighted sum of the estimated ratios of the separate strata as the sample size for each stratum is very small and therefore, there is risk of introducing serious bias in the estimates. We can only think of a combined ratio estimate, i.e. , ratio of simple unbiased estimate of the population total under investigation to the simple unbiased

estimate of area (or population, as the case may be), multiplied by the actual area (population) of the entire country. But here also there would be some uncertainty as to the magnitude of the bias, particularly because stratum to stratum differences in the actual ratios would be very wide for a country of the dimensions of India for several items under investigation[12].

2.16. It should be noted further that although the total number of (first stage) sample units were fairly large for the entire country, India is so large that even for such a sample size the strata, with two units per stratum, had necessarily to be made fairly extensive. The consequence is that even in groups of sufficiently numerous adjoining strata (in order that a necessary condition for the use of the combined ratio method within a group may be satisfied) one is not quite sure that the ratios in the different strata within the same group would be approximately the same, and therefore, even estimates obtainable by pooling together the several combined ratios given by groups of strata would involve an uncertain amount of bias which may in certain cases be not quite negligible. In any case, the complications in computation introduced thereby would be inconvenient in a large-scale tabulation of the data.

2.17. Further complications arise from the multi-stage and other features of the sampling plan. It was decided therefore to adopt a different procedure to be explained shortly, and use an unbiased method of estimating population totals. But as will be found in the course of this paper we had still to use the ratio method when our interest lay not in universe totals but in universe ratios like per capita consumption of cereals etc.

2.18. Other reasons for adopting the preferred method was the simplicity in the estimation of the sampling error, compared particularly with a ratio method, where moreover an "approximate" formula would necessarily have to be used. Developments leading to simplifications in setting confidence limits to the estimates obtained by the accepted procedure will be explained later.

2.19. The method actually adopted will now be described. In dividing any natural division into the appropriate number of strata of approximately equal population, care was taken to see that the *tehsils* constituting any stratum had approximately the same population density (number of persons per unit area), and as far as possible, consistent with geographical compactness[13]. (It may be noted incidentally that as population densities are in general already worked out and published in the census records no work is usually necessary on this account.)

---

[12] Another source of bias of a different nature (in the multiplier "actual" area or population of the entire country) is described in the Section on "Bias due to erroneous stratum sizes". See particularly the foot-note to para 4.7.

[13] On the average in each natural division there are 50 *tehsils* which are to be combined to form roughly 5 strata.

2.20. It may be further noted that the restriction to geographical compactness was not a serious one because, as is to be expected, areas of any specified class of population density occurred more or less in compact blocks.

2.21. Perhaps more interesting and as we shall see later on, important and more difficult to tackle, is the fact that the above feature is true only when the units for which the densities are computed are large, like *tehsils*; and when small units like villages are taken the topographic distribution of population density shows the completely opposite picture of erratic changes in adjoining villages.

2.22. The object of securing homogeneity of the *tehsil*-population-density within a stratum was two-fold. Firstly, there is reason to believe that the economic conditions of an area are related to some extent to the population density, so that stratification by population density would introduce some gains in efficiency.

2.23. Secondly, such a stratification tends to minimise the different between selection with probability proportional to population and that to area. This can be inferred from the obvious fact that if the population densities of the *tehsils* constituting any stratum were exactly identical, then the two modes of assigning the probabilities are also exactly equivalent.

2.24. It is true that the present method did not ensure exact equivalence of the two systems. But slight departures (or even somewhat moderate ones) from any given system will not impair the efficiency appreciably. We can therefore assign a probability which is intermediate between the two systems, that is, use a suitable linear function of the two sizes in the assignment of probability of selection. But the last refinement was not actually employed, and either of the two systems was more or less arbitrarily used. It may be pointed out here that a method of sample selection which avoids the calculation of the linear function for each and every sampling unit is described elsewhere[14].

2.25. The principle involved here is of general applicability in similar situations when a compromise between two modes of assigning variable probability is desired. It should, however, be noted that this method is especially effective when the ratio between the two "sizes" is tolerably stable like *tehsil*-population density. Further, if geographical stratification is of great importance then areas (*tehsil*) of similar "ratio" should form more or less compact blocks.

2.26. It is likely that the information which the 1951 Census would provide for each census tract would be amenable to similar treatment for the purpose of integrating different household enquiries. But this aspect has not so far been examined carefully and it is possible that some modification of the above method will be required to extract full benefit from the variety of information expected in the census reports.

---

[14] The principle of selection is essentially the same as that described in para 3.26 for the selection of units with variable probability. Further details will be found in Lahiri, D. B., A method of sample selection providing unbiased ratio estimates, *Bull. Int. Stat. Inst.*, Vol. XXXIII, pt. II. (International Statistical Conferences, 1951)

3. METHOD OF OVERLAPPING MAPS

3.1. It is more difficult to resolve the problem of the two modes of assigning the probabilities of selection of sample villages within each selected *tehsil*. The principles enunciated above in the construction of strata by grouping together areas of similar population density is not suitable for the following reasons.

3.2. Even if we are prepared to calculate the population density of every village in a sample *tehsil* which itself would amount to a substantial volume of work, the range of variation of population density when worked out on the village basis is extremely large, with the consequence that a large number of density-classes would be required to bring in approximate equivalence of the two modes of probability assignment.

3.3. Moreover the villages falling in any density-class are, unlike *tehsils*, not likely to cluster together to form a more or less compact geographical area; in fact, an examination of the frame shows that the distribution of village population-density is very erratic. It is possible, however, that geographical stratification within a *tehsil* may not be very important.

3.4. In any case a stratified scheme would involve selecting a substantial number of sample villages within each sample *tehsil*, and this we were not prepared to do, for reasons explained earlier. In fact, taking only two sample villages in each sample *tehsil*, as is the current practice from the fourth round onwards, will possibly be replaced by a single village from each sample *tehsil*.

3.5. In the current sampling plan this part of the problem has remained unsolved and the choice between the two alternatives (selection with probability proportional to population or to area) was made more or less arbitrarily (excepting in so far as the availability of the frame at the time of drawing up the revised plan of the fourth round was concerned). However, fair progress has been now made in the collection of both the area and the population figures of all villages and it is time to examine how far the difficulty of integration can be resolved. A brief account is given below of what is being contemplated at present.

3.6. We have seen how attempts to induce equivalence of the two modes of probability assignment at the stage of the selection of the village by the earlier method of stratification proved abortive. The only course appears to be to make separate selections with the respective probabilities for the land utilisation and household enquiries. But this at once doubles the number of sample villages with consequent increase in costs and difficulties in field operations.

3.7. This almost amounts to avoiding the real issue of integration. But this is only so if the selections under the two systems are made independently; and with a dependent scheme of sample selection it is conceivable that actual integration is attainable. Thus, if we could devise a rule of selection which would ensure that corresponding to every village for the household enquiry there is another for land

utilisation survey which is very near, if not completely identical with the former, then we have practically solved our problem, for then from the point of view of field investigator the two villages could be regarded as a single locality so that the number of sample villages would remain effectively almost unaltered.

3.8. The idea of tagging on the land-utilisation village to the household enquiry village is all the more important because the quantum of work for the land utilisation survey is small compared to that for the household enquiry, and a visit to localities separated by long distances with all the attendant trouble just for the land utilisation survey is almost unthinkable. Moreover the scope of dove-tailing the two enquiries which would lead to further economy of investigation time is possible when the two enquiries are restricted to the same locality. In this way it would be possible to carry out the land-utilisation survey at a marginal cost.

3.9. We shall now explain the method that is proposed to be adopted. Suppose that the villages within a *tehsil* are arranged (in the "frame") in a serpentine manner, as is sometimes the practice. Let us now use a convenient graphical representation of the villages. Taking a straight line of any length we can break it up into segments and allocated one segment to each village making the length of the segment proportional to the size of the population of the village. If we arrange the village-segments in the order of their listing, we shall get a linear map of the whole *tehsil* in accordance with the size of the population of the individual villages. In the same way we can construct a linear map of the *tehsil*, in accordance with the size of the area of individual villages. If we now make the length of the two maps the same, and superpose one map on the other, we shall get two different but overlapping mappings of the villages on the same straight line corresponding to the two sizes : area and population. If now a point is thrown at random on this line then a pair of villages (which may or may not be identical) will be selected with probabilities respectively proportional to area and population. If the villages are identical then both enquiries will be done in this village. If two different villages are selected, even then they are likely to be in the same neighbourhood. In fact, the distance between the two villages will depend roughly on the difference between the two serial numbers of the two villages located by the random point. The probability distribution of the "distance" (difference between the serial numbers) between the villages depends on the relative magnitudes of the two sizes and also on the ordering of the villages. But without entering into the mathematics of the problem it may be said from general considerations that the "distance" will be usually small. This has been empirically verified for certain *tehsils* (police stations) in West Bengal.

3.10. As there is a certain amount of freedom in ordering the villages in a serpentine manner, one can utilise this freedom to arrange the villages in such a way that the expected "distance" between the pair of sample villages is minimized. But if the serpentine arrangement is already present in the available frame it may not usually be worth while to rearrange this to give better results. Consider an overlapping map (say, the "population"-map superimposed on the "area"-map) both

covering the entire straight line. Any village will be represented by a population-segment ("p"-segment) as well as an area-segment ("a"-segment). Now, as one proceeds along the serpentine frame from the first village adding on one village after another, a greater and greater area of the *tehsil* will be covered from its one end to the opposite one.

3.11. Now consider how the population density fluctuates for the above "cumulative" area. If the topographic distribution of the village population-density is sufficiently random then the "cumulative" density will also fluctuate rapidly round the *tehsil* population density. And every time the "cumulative" density "crosses" the *tehsil* density either from above or below, we shall reach, as can be easily verified, a village where the a-segment and the p-segment are such that one is contained entirely within the other. If the crossing is from below the a-segment is contained entirely within the p-segment, and *vice versa* if from above.

3.12. Now consider the sections of the map formed by removing the "contained" village segments. Within any given section either the "p"-segment consistently lags behind the corresponding "a"-segment or *vice versa*; and the sections of the two types occur alternately. Although theoretically the lag may be almost equal to the entire length of the section but in actual practice the lag will be much smaller. Thus if the random point falls in any of these sections the pair of villages selected thereby will be fairly close[15] to each other if not completely identical, and the latter will certainly be true if the point falls in any of the "contained" segments.

3.13. Thus when the serpentine arrangement is more or less at our choice we should endeavour to arrange the villages in such a manner that the "cumulative" density fluctuates as frequently as possible about the *tehsil* density.

3.14. The serpentine arrangement of the villages leads to an arrangement which is really transforming a two-dimensional map into a one-dimensional pattern, and therefore really goes only half-way in solving the problem discussed in the preceding paragraphs. We have described this procedure because of the simplicity of the method of selection particularly in the case where the villages are already arranged in a serpentine manner in the available frame. But one would not expect the real geographical distance (as distinguished from the "distance" measured as the difference between the serial numbers considered earlier) to be as small on the average as one would expect if a suitable "two-dimensional" procedure of selection of sample villages could be developed; and we are going to indicate in the next few paragraphs how this can be done. It should be noticed that the main drawback of the serpentine method lies in the fact that although villages separated by a small "distance" (as measured by the difference between the two serial numbers) are really

[15] Another way of looking at the situation is as follows. Every section of the overlapping map (excepting for some minor end-adjustments) may itself be regarded as the overlapping map of some compact sub-region of the *tehsil*. Therefore we at least ensure that both the villages lie in the same sub-region and in actual practice both will not generally lie near the two "extremeties" of the sub-region so that the sample villages are likely to be fairly close to each other.

near, yet villages separated by a large "distance" (large difference between serial numbers) may even be in reality just adjacent to each other.

3.15. When we have to take the real distance into consideration naturally we have to proceed on the basis of a map. Although we are speaking about the real distance it is difficult to give a mathematical definition of the distance between two villages which would precisely measure the effective distance of travel in some sense which really matters from the point of view of investigation. Perhaps the distance between the "centres" (centre of gravity, for example) of the villages may serve as a good measure.

3.16. We shall not just at the moment formulate the problem in a precise mathematical way but shall take the attitude of a practical statistician in devising some method which would in all likelihood approach the "optimum" which may be arrived at on strict mathematical analysis (but even the theoretical "optimum" will be conditioned by the more or less arbitrary definition of the "distance" between two villages).

3.17. The method to be employed is essentially the method of overlapping maps, but unlike the serpentine method the maps are no longer one-dimensional ones but of two dimensions. Our starting point is a map of the region under consideration showing the boundaries of the villages, each village being correctly represented in the map as regards its shape and area. Now keeping the boundary of the region intact, "distort" the map or rather the boundaries of the villages in such a manner that the "new" areas of the villages are proportional to their populations.

3.18. It is now obvious that if a point is thrown at random on the superimposed map (consisting of the "new" map on the original one) then a pair of villages will be selected with the desired probabilities. The success of the method in ensuring a "close pair" will depend upon our ability to accomplish the above transformation with the minimum "amount of distortion". In other words, we want to approach the "optimum distortion" which makes the expected value of the distance between the pair of villages selected by the above process (e.g., mean distance between the "centres" of the pair of villages) a minimum. We do not know the general solution for the optimum distortion. But distortions wihich should give reasonably good results may be found by inspection. It will also be noticed that the "serpentine" method explained earlier is one in which the distortion is restricted to only one direction, namely, the one given by the ordering of the villages.

3.19. Although we have used the concept of map in explaining our methods, it does not mean that arithmetical procedures are not available. The appropriate procedure for the serpentine method is evident. In the case of the general two-dimensional distortion and superposition plan, the basic ideas are the same. Very briefly stated, for the cumulative method of selection we reserve sets of consecutive integers (not for a single village but) for an ordered pair of villages, the system of pairing and fixing of the "composite size" of the sets being appropriately done. A

first-position (or second-position) village may, and generally will, have to occur in more than one such pair.

3.20. As an illustration of the method we are giving the following map and relevant information (Table 1). The villages are numbered in a serpentine fashion.

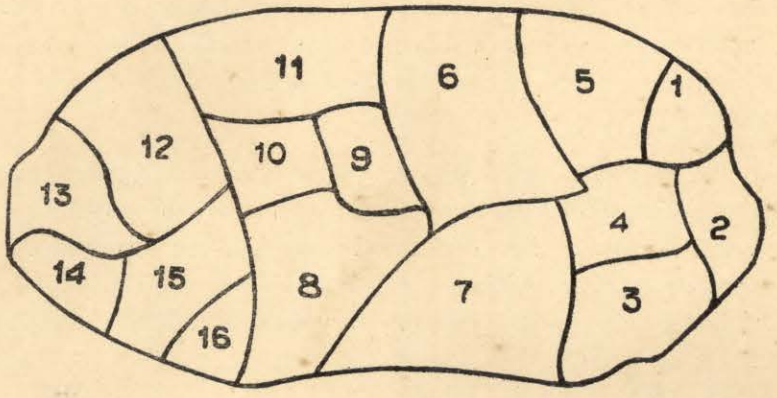

Fig. 1

TABLE 1. AREA AND POPULATION

village no.	area	population	village no.	area	population	village no.	area	population	village no.	area	population
(1.1)	(1.2)	(1.3)	(2.1)	(2.2)	(2.3)	(3.1)	(3.2)	(3.3)	(4.1)	(4.2)	(4.3)
1	3	8	5	11	6	9	5	10	13	4	2
2	4	5	6	15	20	10	5	1	14	4	10
3	6	10	7	15	10	11	10	9	15	6	6
4	5	5	8	11	5	12	8	3	16	4	6

For convenience the unit of area is so chosen that the figures for the area and population of the entire field are identical (116).

3.21. The scheme of pairing of villages together with the assignment of composite sizes is shown below in Table 2. For the two-dimensional case the pairing and fixing of the "composite" size is done by inspection, whereas for the serpentine method once the ordering of the villages is decided upon the rest follows automatically[16]. To choose a pair of villages we choose a number at random from the set 1, 2, ..., 116, say 76. Then for the serpentine method we take up village No. 10 for the land utilisation survey and village No. 9 for the household enquiry. For the other method the land-utilisation village as well as the household enquiry village are both identical viz., No. 10.

[16] For the serpentine method it is basically necessary, as can be easily seen, to arrange both the sizes (area, population) in the *same* serpentine order and then obtain the cumulative totals. The elaboration shown in Table 2 is given only to show the parallelism between the two methods.

TABLE 2. PAIRING OF VILLAGES AND FIXATION OF COMPOSITE SIZE

serpentine method				two-dimensional method			
village no.		composite size	integers reserved	village no.		composite size	integers reserved
l.u.	h.h.			l.u.	h.h.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1	3	1 — 3	1	1	3	1 — 3
2	1	4	4 — 7	2	2	4	4 — 7
3	1	1	8	3	3	6	8 — 13
3	2	5	9 — 13	4	4	4	14 — 17
4	3	5	14 — 18	4	2	1	18
5	3	5	19 — 23	5	5	6	19 — 24
5	4	5	24 — 28	5	1	5	25 — 29
5	5	1	29	6	6	15	30 — 44
6	5	5	30 — 34	7	7	10	45 — 54
6	6	10	35 — 44	7	3	4	55 — 58
7	6	10	45 — 54	7	4	1	59
7	7	5	55 — 59	8	8	5	60 — 64
8	7	5	60 — 64	8	6	4	65 — 68
8	8	5	65 — 69	8	9	2	69 — 70
8	9	1	70	9	9	5	71 — 75
9	9	5	71 — 75	10	10	1	76
10	9	4	76 — 79	10	9	3	77— 79
10	10	1	80	10	12	1	80
11	11	9	81— 89	11	11	9	81— 89
11	12	1	90	11	6	1	90
12	12	2	91 — 92	12	12	2	91 — 92
12	13	2	93 — 94	12	14	6	93 — 98
12	14	4	95 — 98	13	13	2	99—100
13	14	4	99—102	13	14	2	101—102
14	14	2	103—104	14	14	2	103—104
14	15	2	105—106	14	15	2	105—106
15	15	4	107—110	15	15	4	107—110
15	16	2	111—112	15	16	2	111—112
16	16	4	113—116	16	16	4	113—116

l.u. = land utilisation h.h. = household

3.22. To compare the two methods we obtain the following tabular results.

TABLE 3. COMPARISON BETWEEN SERPENTINE AND TWO-DIMENSIONAL METHODS

probability of selecting villages which are	serpentine	two-dimensional
(1)	(2)	(3)
identical	51/116	82/116
adjacent	55/116	28/116
further away	10/116	6/116

Thus the two dimensional method has introduced definite gains.

3.23. It may be thought that the success of the above two methods depends very largely on the existence of a higher positive correlation ($r = +0.5$ nearly in the above illustration) between area and population. It will however be found from the following illustration (Table 4) where there is a negative correlation ($r = -0.4$ nearly) that the existence of high positive correlation is not of overwhelming importance.

TABLE 4. AREA AND POPULATION (MODIFIED)

village no.	area	population	village no.	area	population	village no.	area	population	village no.	area	population
(1.1)	(1.2)	(1.3)	(2.1)	(2.2)	(2.3)	(3.1)	(3.2)	(3.3)	(4.1)	(4.2)	(4.3)
1	3	6	5	11	9	9	5	10	13	4	10
2	4	10	6	15	3	10	5	5	14	4	6
3	6	6	7	15	1	11	10	5	15	6	2
4	5	20	8	11	8	12	8	10	16	4	5

The serpentine method gives the following result.

TABLE 5. EFFECTIVENESS OF SERPENTINE METHOD

probability of selecting villages which are	serpentine
(1)	(2)
identical	34/116
adjacent	66/116
further away	16/116

In the last category "further away" in no case there is more than one village separating the two selected villages.

3.24. Speaking about the cumulative method of selection with variable probability we are naturally drawn to a method developed at the preparatory stages of the survey work which avoids the work on cumulation. Here again the ideas originate from the representation of units in a linear map.

3.25. One of the practical difficulties encountered at the very beginning was the selection of sample villages. It will be remembered that the general plan was to divide the country into a number of geographical strata and then select (with replacement) villages with (varying) probability proportional to "size", namely, either area or population. The use of the known method of cumulative totals for the selection of villages with variable probability involves the determination of cumulative totals for more than half a million villages in India. But in view of the fact that the number of sample villages to be selected from any stratum is very small compared to the total number of villages in that stratum, this part of the work seemed to be somewhat out of proportion to the net result(i.e., selection of the sample villages). This led to the development of the method of selection explained below.

3.26. In essence of the method is as follows. Let us suppose that the population consists of N sampling units; and let X, say, be the 'size' of the largest unit. Choose a pair of random numbers ν and ξ, the first in the range 1 to N and the second between 0 and X (if the unit of measurement is such that the sizes are in integers then this means selecting an integer at random from the set 1, 2, ..., X). Let X

be the size of the ν-th unit. Then if $\xi > X_\nu$, that is, ξ exceeds the "size" of the ν-th unit reject, otherwise accept the sampling unit. If rejected, repeat the operation until a selection is made. A unit once rejected may however be selected at a subsequent operation. It is possible to replace X by a larger number, but then the number of rejections will be higher. This procedure was explained in the 1951 session in India of the UN Statistical Sub-Commission[17]. Further details, including methods of reducing the number of rejections will be found in a paper presented to the International Statistical Conferences, India, 1951[18].

3.27. We have seen how in the process of integration of the household enquiry and the land utilisation survey the topographic distribution of the ratio between population and area, i.e., population density, played a fundamental ro'e in the process of integration. We have also seen how the character of the topographic distribution changes with the "units" (*tehsils* or villages) chosen for the calculation of this ratio, and in what way it is relevant to the question of integration. We have not so far examined the nature of the topographic distribution, for example, of the proportions under different occupational groups which will be of importance in integrating the different household enterprise enquiries.

3.28. The technique of overlapping maps opens up new avenues of improving the efficiency of the sample design; of course, which particular avenue is to be selected in any specified case depends upon its peculiar conditions. We have already shown its use in selecting a pair of sample villages, one with probability proportional to area and the other proportional to population in such a manner that the two villages are close to one another, if not completely identical, so that from the investigation point of view the pair may be regarded as a single locality. But the possibilities do not end here and we hope to publish in future the ways in which it is proposed to be employed in the NSS.

---

17 United Nations Economic and Social Council; Report to the Statistical Commission on the 5th session of the Sub-Commission on Statistical Sampling, India, December, 1951.

18 The number of rejections is particularly large when there are a few extraordinarily large units. In such cases it is generally desirable to treat them as a separate stratum. However, if it is not proposed to treat them as such a simple method of reducing the number of rejections is available. Mark the large units; this means choose a suitable (round) number X' and mark the units whose size is greater than X'. Let P be the size of a large unit, and let Q be the quotient and R the remainder when P is divided by X'. The large units (already given a serial number along with all the other units) are given additional serial numbers,—thus the above unit of size P is given Q additional serial numbers (starting from $N+1$, if it is the first large unit to be so treated) and so on for the lother large units. Let us suppose that we hereby go upto the serial number N'. Then choose an integer ν' at random in the range 1 to N'; if ν' is in the sub-range $N+1$ to N' then choose the unit bearing the serial number ν'. Otherwise throw a random point ξ' between 0 and X'. If ξ' exceeds the size of the ν'-th unit,—there is one stipulation, namely, the large unit P should be considered to be of size R and similarly for other large units,—reject the unit, otherwise accept. If rejected repeat the entire operation until a selection is made. Lahiri, D. B., *loc. cit.*

4. BIAS DUE TO ERRONEOUS STRATUM-SIZES

4.1. We next turn to a discussion on how certain practical difficulties connected with the sampling frame affected the sample design. One of the problems frequently met in large scale sampling practice is the use of "erroneous"[19] stratum-sizes and to this problem we now turn our attention. Another parallel problem arises in the use of erroneous sizes of first stage units in two-stage (or multi-stage) sampling plans. The possibility of introducing some bias, which may not be entirely negligible, at the estimation stage by the use of such erroneous sizes, was realised in course of the first few rounds of the survey; and it became necessary to develop a procedure which, while taking advantage of the erroneous but otherwise approximate sizes, would nonetheless furnish an unbiased method of estimation.

4.2. Theoretically speaking it is difficult to visualize a situation in which one is in a position to select the sample units rigorously with the desired probability (proportional to size) but is not in a position to determine the actual stratum-sizes. For example, in the first three rounds of the survey the village was the first stage unit and the selection was usually with probability proportional to the 1941 population or to the area of the village, and on first thought it appears that such a selection necessarily implies the availability of population or area figures for all the villages in every stratum and consequently the strata sizes could have been determined correctly if desired. But in actual execution certain difficulties arise which do not usually allow the use of such a procedure.

4.3. We shall employ the following notations in this section. The meaning of the terms used will be found in the text. The villages and *tehsils* all belong to a specified stratum. For the sake of simplicity we shall assume that no sub-sampling is done within the villages.

Frame T is a list showing sizes of all *tehsils* constituting the stratum
Frame V is a list showing sizes of all villages constituting the stratum
Frame V_t is a list showing sizes of all villages constituting the t-th *tehsil*
Frame V is the collection of all V_t's
Σ_{vt} $\equiv$ summation over all villages in the t-th *tehsil*
Σ_{v*} $\equiv$ summation over all villages in the stratum
Σ_{*t} $\equiv$ summation over all tehsils in the stratum
S_{vt} $\equiv$ summation over all sample villages in the t-th *tehsil*
S_{v*} $\equiv$ summation over all sample villages in the stratum
S_{*t} $\equiv$ summation over all sample *tehsil*
$Z_{vt}(V)$ = size of v-th village in t-th *tehsil* as given in frame V
$Z_{v*}(V)$ = size of v-th village (in the stratum) as given in frame V

[19] "erroneous" is used here in a special sense explained in para 4.7. In fact the *real* stratum-size may be "erroneous" in relation to the method of sampling and estimation actually adopted.

$Z_{*t}(T)$ = size of t-th *tehsil* as given in frame T

$Z_{*t}(V)$ = size of t-th *tehsil* as derived from frame V
$= \Sigma_{vt}[E_{vt}(V)]$

$Z_{**}(V)$ = size of the stratum as derived from frame V
$= \Sigma_{*t}[E_{*t}(V)]$

$Z_{**}(T)$ = size of the stratum as derived from frame T
$= \Sigma_{*t}[Z_{*t}(T)]$

$Z_{**}(K)$ = known (or real) stratum size

Y_{vt} = value of some characteristic for the v-th village of the t-th *tehsil*

Y_{v*} = value of the characteristic for the v-th village (in the stratum)

Y_{*t} = value of the characteristic for the t-th *tehsil*
$= \Sigma_{vt}[Y_{vt}]$

Y_{**} = value of the characteristic for the stratum
$= \Sigma_{*t}[Y_{*t}]$

n_{vt} = number of sample villages in the t-th *tehsil*

n_{v*} = number of sample villages in the stratum

n_{*t} = number of sample *tehsils* in the stratum

4.4. Consider the following situation which is one of common occurrence. Any stratum (or first-stage) unit is composed of one or more administrative divisions (*tehsils*), no such division belonging to more than one stratum and the total size, $Z_{*t}(T)$ [population (T) or area (T)] of each such administrative division being given in, say, published records (frame T). It has been noticed, curiously enough, that stratum sizes $Z_{**}(T)$ (area, or population) derived from frame T may not agree with those $Z_{**}(V)$ obtained by the addition of the available figures in frame V of population (V) or area (V) of the villages constituting such stratum (or first-stage unit) ; and either or both of these may differ from the known stratum size $Z_{**}(K)$.

4.5. This discrepancy is explained by a number of causes. The frames T and V as well as the "known" stratum size may have been completed by different agencies[20], or do not refer to exactly the same point of time, or the definitions are not quite identical[21]. Even when the two frames are constructed by the same agency there may still be some discrepancy, for example, the village population figures may be those of a preliminary count while the *tehsil* or district figures refer

[20] For example, two series of figures are "known" for the areas of the districts in Madhya Pradesh (Tables AI and E of *Census of India*, 1951, Vol. VII, Part II-A, General Population Tables and Summary figures for districts). One set of figures is supplied by the Surveyor-General and the other by Deputy Commissioners. The magnitude of the discrepancies for the 22 districts is summarised below (the discrepancies are expressed as percentages of the figures supplied by the Deputy Commissioners) : nearly 30% in 1 case, more than 15% in 2 cases, about 4% in 2 cases, about 3% in 2 cases, about 2% in 3 cases, about 1% in 6 cases and fairly small discrepancy in all the remaining 6 cases. Moreover the Deputy Commissioners' figures are less than the Surveyor-General's figures in all cases of large discrepancy.

[21] Three different figures for the total population of all cities in Uttar Pradesh are obtainable from the Census Reports ; (incidentally it may be pointed out that this is of relevance in the selection of urban areas in the NSS). (1) 3,908,056 from Table V, population of cities (towns with a population of 100,000 and over), *Census of India*, Paper No. 1, 1952, Final Population Totals—1951 Census : (2) 3,370,551 from

to the final count. There may also be genuine clerical and printing mistakes and so on in the available frames.

4.6. In theory there is no difficulty if within a stratum it is desired to select the villages precisely with probability proportional to size Z_{v*} (V) (population or area), necessarily as given in frame V. To calculate

$$\frac{Z_{**}(V)}{n_{v*}} \cdot S_{v*}\left[\frac{Y_{v*}}{Z_{v*}(V)}\right]$$

which is an unbiased estimate of the stratum aggregate Y_{**} of any characteristic (e.g., total unemployed) it would further be necessary to obtain correctly the stratum size $Z_{**}(V)$ from the same frame, thus rejecting frame T for this purpose. But as this involves totalling up a huge number of figures, the use of frame T for obtaining the stratum size $Z_{**}(T)$ may be preferred. Other reason, perhaps more important, arised from the difficulty of collecting the *complete* frame V, so that there is no question of calculating $Z_{**}(V)$.

4.7. But as has been pointed out earlier the analysis of the data collected showed that sometimes the use of such "erroneous" stratum-sizes $Z_{**}(T)$, although employed by competent statisticians, may give rise to the bias

$$\left[\frac{Z_{**}(T)}{Z_{**}(V)} - 1\right].Y_{**}$$

which may not be entirely negligible. From the point of view of bias, it should be further noted, it is irrelevant which frame, if any, gives the real size $Z_{**}(K)$ (population or area), so that when we say that frame T gives erroneous stratum sizes we simply mean that these sizes $Z_{**}(T)$ are different from those, viz., $Z_{**}(V)$ derivable from frame V, which should have been used, to obtain unbiased estimates. Indeed the use of real stratum size $Z_{**}(K)$, if available, would introduce bias if stratum-size $Z_{**}(V)$ derivable from frame V does not agree with the real size $Z_{**}(K)$. The real danger perhaps lies in the fact that one may not even be conscious that he is using a wrong stratum size[22].

4.8. Before discussing how the frame T can be used, but without at the same time introducing any bias it is important to get a clear idea of other inherent

---

Summary Table II, Towns and villages classified by population, *Census of India*, Paper No. 3, 1953, Summary of Demographic and Economic Data, 1951 Census. (3) 3,340,999 from Summary Table III, Towns with population of 20,000 and over, of the above Paper No. 3. Possibly (although not clear from the two above papers) no consistent specification of the territorial limits of a city has been adopted in the three tables. This example is given merely for illustrative purposes ; in cases of cities, because of the smallness of their number, there is no reason to adopt an erroneous stratum size.

[22] For example, this is mostly so when a ratio estimate is used to estimate the stratum total of any characteristic. Thus for such estimates there are two sources of bias, firstly the bias in the formula for estimation of a ratio (which may be negligible when the sample is sufficiently large) and secondly, bias in the multiplier (the stratum-size) arising from sources enumerated in para 4.5.

difficulties in attempting to base the sampling solely on the basis of frame V. In this procedure it is assumed that the complete[23] frame V is already available with the sampler ; and, if not, he has enough time and resource to collect it. Indeed the collection of the frame V may be so expensive that the idea of collection of the *entire* frame may have to be given up. Moreover he must have adequate resources to obtain correctly the strata-sizes $Z_{**}(V)$ (for a country of the size of India). The strata-sizes $Z_{**}(V)$ were required after the survey at the estimation stage for survey designs adopted for the first three rounds, but if a self-weighting (discussed later) system is desired it would be necessary to obtain the strata-sizes before the survey strata and it would then be necessary to examine whether this part of the job could be completed in time, before the scheduled date of the field work.

4.9. The practical difficulties connected with the collection of the frame V was keenly felt at the time of revising the sample design for the fourth round when the unpublished (provisional) 1951 Census figures had been intended to be used. A self-weighting system was also desired. The question therefore initially arose as to how to sample with replacement the villages directly within a stratum with the desired probability and yet overcome the difficulties explained above. (We shall for the moment ignore the fact that finally, on account of difficulties of the collection of the necessary frame, the direct selection of villages within a stratum was given up). The solution very fortunately lay in the "modification" unwittingly introduced in the first three rounds. In these rounds purely with an intention to simplify the operational procedure of selection another stage, as described in the next paragraph, was introduced. But as will be explained later on, this in reality modified, for some unforeseen reasons, the probability of selection that was really intended. However, in subsequent rounds this gave us the real clue to meet the difficulties stated in the preceding paragraphs.

4.10. In the earlier rounds the procedure of selection of sample villages from any stratum (which was generally a single district or a combination of districts) it was found convenient to proceed by two stages—firstly a *tehsil* (or similar smaller administrative unit) with probability proportional to size $Z_{*t}(T)$ (population or area as given in frame T), and then a village with probability proportional to size $Z_{vt}(V)$ (population V or area V) necessarily on the basis of frame V_t of selected *tehsil*. The whole process was repeated until the entire quota of villages were completed, care being taken to replace the selected *tehsil* and village before a fresh selection was made. This two-stage selection procedure was found convenient because popula-

[23] In practice to ensure that there are no omissions or duplications is an extremely difficult task, especially when the different parts of the frame (lists and/or maps) refer to different points of time, as was experienced in the NSS particularly in the initial rounds. The 1951 Census frame refers to the same point of time and is therefore likely to be better to that extent, but even then some scrutiny is necessary, specially for clerical errors of gross omissions and duplications.

tion and area figures were compiled separately for each of the smaller administrative units (*tehsils*)[24].

4.11. It must be admitted here that at the time of sample selection for the first round it was taken for granted that the two frames were mutually consistent, that is, stratum-size as obtainable from frame T is identical with that derivable from frame V. That is, it was assumed $Z_{**}(T) = Z_{**}(V)$. Consequently it was thought that the exact probability of selection of a sample village (required to obtain the necessary estimates) would be easily obtained by dividing the village-size population (V) or area (V), by the stratum-size population (T) or area (T) i.e., by $Z_{vt}(V)/Z_{*t}(T)$. The use of this method for the calculation of the desired probability would not be, however, correct when the two frames are not consistent and may lead to bias (which is called here bias due to the use of erroneous stratum-sizes). But once we realised that the two frames were not mutually consistent there was no reason to adopt a biased method. The exact probability of selection is obtained with a little more trouble, as is easily seen, and therefore an unbiasd procedure is really available.

4.12. There is one interesting fact which may be noted. There are two ways in which the above change of front may be looked upon,—change from what was initially intended to what was actually executed. It may be taken either as prefixing another stage of sampling (*tehsil*) with replacement and then sub-sampling a single unit (village) from each selected first-stage unit; or, as a modification in the probability of selection of a village without the introduction of an additional stage. Thus villages may be supposed to be selected with replacement with probability proportional to adjusted-size directly within a stratum. The adjusted-size is equal to village-size (V) multiplied by the ratio of the size (T) to the size (V) of the *tehsil* in which the village is situated; that is,

$$\frac{Z_{*t}(T)}{Z_{*t}(V)} \cdot Z_{vt}(V).$$

4.13. Coming back to the problems faced at the time of drawing up the sample design of the fourth round we get the indication from the above discussion that in order to obviate the difficulties it is necessary to replace selection with probability proportional to population (V) or area (V) to one proportional to adjusted population or adjusted area. This is equivalent to the introduction of one more stage in the sampling design. Firstly to select (with replacement) a *tehsil* with probability proportional to population (T) or area (T) on the basis of frame T, and then to select a village with probability proportional to population (V) or area (V) but on the basis of frame V.

[24] A similar technique had been used earlier by the ISI in the selection of sample grids within zones or police stations for the Bihar and Bengal Crop Surveys from 1942. Before making the final selection of grids in the village maps (showing the plot boundaries) the quota for each village was fixed by throwing the desired number of points (each corresponding to a sample grid) at random on zone or Police Station maps showing village boundaries (but not plot boundaries). See Report on Bihar Crop Survey : Rabi Season 1943-44, P. C. Mahalanobis, *Sankhyā*, 7(1), 35, 1945.

4.14. It can easily be seen that by the introduction of this procedure the practical difficulties associated with the collection and scrutiny for omissions and duplications[25] of the detailed frame V were much reduced since it was now required to collect detailed information for only a small number of *tehsils*, namely, those included in the sample. Moreover, the huge work of totalling up more than half a million village figures, stratum by stratum, was replaced by a smaller volume involving totals for only the small number of selected *tehsils*. At the same time by the use of appropriate estimation formula, viz.,

$$\frac{Z_{**}(T)}{n_{v*}} \cdot S_{v*} \left[\frac{Y_{vt}}{Z_{vt}(V)} \cdot \frac{Z_{*t}(V)}{Z_{*t}(T)} \right]$$

the bias due to erroneous stratum sizes in the estimation of Y_{**} was avoided. In actual practice sub-sampling (households or plot-clusters) was done within villages and in the above formula for estimation Y_{vt} was replaced by its estimate $\hat{Y}_{vt}$.

4.15. Although the original intention was to sample the villages directly within a stratum by the simple contrivance of first fixing the *tehsil*, the volume of preliminary work was still of unmanageable dimensions under the circumstances prevailing at the time of drawing up the changed design in the fourth round in which the latest but unpublished provisional population (1951) figures were being utilised, that further reduction of the preparatory work became necessary. This was effected by choosing with replacement $n_{*t} = 2$ *tehsils* within every stratum and then $n_{vt} = 2$ (so that $n_{v*} = 4$) villages with replacement in each selected *tehsil*, thus cutting down to half the number of *tehsils* which would be required if a single village was selected from each *tehsil*. This had also one additional advantage over the initial scheme in that it would enable us to obtain "within" *tehsil* variances which may be useful in planning future sample surveys.

4.16. We may also add incidentally that the villages could have been selected without replacement in a selected *tehsil*. This would not introduce any additional trouble of obtaining unbiased estimates of the sampling error of estimates of national "totals", for example. But estimation of "within" *tehsil* variances, if desired, would be much more involved.

4.17. It will be noticed further that this procedure would not quite remove the possibility of the same village occurring twice in the sample as the *tehsils* were selected with replacement, because of the resulting simplicity, discussed later, in the estimation of sampling error. The probability of repetition of the same village was already small with the with-replacement-scheme actually adopted because the number of villages in a *tehsil* was moderately large, and with the without-replacement-scheme discussed in this paragraph, this probability would be smaller still.

---

[25] Scrutiny for the completeness of frame T is of course necessary, but this is a comparatively simple matter.

5. LAND UTILISATION SURVEY[26]

5.1. We next turn to the problem of overcoming the difficulties associated with the collection of a suitable frame[27] for the land utilisation survey. It is not possible to isolate the problem of collection of the frame from the manner in which it can be possibly utilised for sampling and thence for obtaining the necessary estimates. We shall, therefore, have to consider all the aspects simultaneously.

5.2. The land utilisation survey called for detailed maps showing plot boundaries of all the sample villages so that sample plots could be selected for the survey[28]. It has already been stated in the Report on the First Round that, not to speak about such village maps, attempts to collect even *tehsil* maps, showing merely the village boundaries, were unsuccessful in many cases. Thus all hopes of making the sample selection in the Institute Laboratory were shattered. Apparently the only course left open was to wait and with considerable expenditure of time and money collect the necessary maps. This may in certain cases entail the actual presence of the map collector in the village concerned, this being the only known place, where copies of the map may be found.

---

[26] We shall use the following notation in this section. The meaning of the terms used hereunder will be found in the text.

A = area of village

n = number of sample clusters (compact, patterned, or quasi-compact), entry-plots or "spots" in the village

n' = number of (ultimate) plots sampled in each quasi-compact cluster

a = area of plot

p = proportion of the plot under any specified utilisation

ε = 1 or 0 according as the "spot" falls or does not fall on the specified land utilisation type

Σ = summation over all plots constituting a "compact" or "patterned" cluster

Σ' = summation over all plots constituting a "quasi-compact" cluster

S = summation over all sample clusters ("compact" or "patterned") in the village

S' = summation over all sample quasi-compact clusters in the village

S_e = summation over all sample entry-plots in the village

S_s = summation over all sample "spots" in the village

[27] See footnote to para 2.1.

[28] The ISI method developed earlier in connection with the Bengal Crop Surveys was a stratified scheme in which the sampling units were square grids of a few acres. The sample grids were allocated to a stratum and then at random to individual villages, each village either receiving no grid at all in which case it was excluded from the sample or receiving one or more grids. The appropriate number of grids (one or more) were stamped at random on village maps showing plot boundaries. The investigators then went to the sample-villages; identified, with the help of the village maps, the plots which fell entirely or partly within each grid; and, by direct physical observation of the crops on the ground, recorded the proportion of each plot sown with each crop. The acreage under any specified crop within any sample-grid was determined by adding together the products of plot area (measured on the map) and the proportion of plot under the specified crop (observed by the investigator) for all the plots included in the grid. If a plot was only partially included within the grid, its contribution to the crop is recorded as the product of the actual area of the plot included within the grid and the proportion of the whole plot under the specified crop. The estimation of the crop acreage for the stratum is thereafter quite straightforward. The optimum size-density distribution of grids was determined by very detailed studies of variance and cost functions (Mahalanobis, P. C., *Phil. Trans. Roy. Soc.*, London, Series B, No. 584, 231, 329-451, 1944).

5.3. The question of postponing the general survey only for this lacuna was obviously impracticable. The land utilisation survey itself could be postponed for the first one or two rounds; and in the meantime attempt could of course have been made first to select a number of sample villages, and then to collect maps of these villages with a view to utilising these for some subsequent round of the NSS. This would have however involved either using the same set of villages as in the first round, or fixing the set of sample villages in advance for the subsequent surveys, once for all, at least for some considerable time.

5.4. Using the same set of sample villages round after round is however not desirable; and in fact, is not workable as was realised during the course of the survey, when it was found that the villagers as a group resented being interviewed repeatedly round after round. Moreover, using a fixed set of sample villages would necessarily entail loss of that flexibility in the sample design which would make possible the utilisation of the information collected during the course of the different rounds of the survey or the utilisation of information collected by other agencies like the 1951 Population Census for the improvement of the sample design for subsequent surveys.

5.5. Although excellent "frames" of more or less permanent nature, namely, cadastral maps showing plot boundaries, existed for the greater part of the more important agricultural regions, it was a stupendous and therefore practically impossible task to collect all these millions of maps, in order that full flexibility could be maintained in the sample design. It was necessary, therefore, to improvise some reasonably simple but statistically sound method of selecting the plots locally by the investigators themselves.

5.6. Moreover, in preparing the rule of selection of the sample plots one has to bear in mind the well known fact that the time taken for the identification of a plot is comparatively large compared to the time taken for noting down the actual land-utilisation of such plots, so that it is usually worth while to survey a compact block of plots instead of a single plot. This was realised quite early in the Bengal Crop Surveys when the use of individual plots as the sampling unit was rejected and a grid of plots was favoured.

5.7. The rule of division of a sample village into blocks or cluster of plots and the method of selection should of course be such that strictly valid estimates may be formed. The formula for estimation should be sufficiently simple, an unbiased formula should generally be preferred, unless of course it can be shown that an alternative formula, although biased, is of higher precision, the bias being of negligible magnitude. And from the operational point of view, the rule for the determination and selection of the cluster of plots should be sufficiently simple for the investigator to act upon.

5.8. It appeared that if a sufficiently simple procedure of selecting (with replacement) clusters of plots with probability proportional to area could be devised, then an unbiased estimate ($\hat{Y}_{vt}$) of total acreage under different types of utilisation

could be obtained easily for the sample village, the village area being known, and thence for the corresponding stratum by using the estimator given in para 4.14. Before describing the procedure actually adopted for this purpose (para 5.10 etc.) we shall in the next paragraph briefly discuss the possibility of using other methods.

5.9. One method could be the selection of clusters with equal probability, which is not a difficult task, and then making an unbiased estimate based on the total number of plot-clusters in the village. But this would give results of lower precision. The ratio method of estimation using areas of plot-clusters as supplementary information could have been adopted to increase the precision but even leaving apart the computational difficulties that would necessarily be involved there was no point in adopting a biased method when an unbiased method described below (which was likely to produce results of comparable precision) was available. The proposed method had other advantages which will be noticed in course of the next few paragraphs. The advantages were so important that these led to the rejection of the other alternatives discussed in this paragraph[29]. We now revert back to the question of selection with probability proportional to area.

5.10. The first step depended upon a very simple improvisation which worked extremely well. This was merely a rectangular sheet of paper with circular holes punched at random on the basis of a pair of random numbers. The perforations were numbered in the order of random selection (with replacement). This sheet of paper when superimposed on the village map would enable the investigator to select the requisite number of plots with replacement with probability proportional to the area of the plots, the plots being located by the centres of the circular holes. Several such perforated sheets arrived at by independent process of randomisation were used.

5.11. The above procedure would have sufficed if the object were to select single plots, but the aim was to select clusters of plots, with probability proportional to the area of the cluster. For this purpose full advantage of the serpentine method of assigning the survey numbers to the plots was taken in defining the clusters in a very simple manner, so that the "entry-plot" (that is, the plot on which the centre of a random hole in the perforated sheet fell) would automatically determine the cluster. Moreover, the selection of the cluster would also be proportional to its area.

5.12. There are a variety of ways in which the rule for the determination and selection of the cluster may be given. In every case the definition and selection of the cluster is through the entry-plot. Thus for a "compact" cluster of five plots we may have the rule: if the digit in the units place of the survey number of the entry-plot is any one among 0,1, 2, 3, 4 then choose the cluster formed by the plots whose survey numbers are obtained by substituting all these numbers one by one in the units place; treat similarly the case for 5, 6, 7, 8, 9.

---

[29] In areas where a list of plots (or holdings) were available but not the village map showing plot boundaries the plots (or holdings) were selected at random with equal probability. But in such cases the investigator had necessarily to depend upon a local man for the identification of the plots (or holdings).

5.13. If a greater spread of the plots is desired (without, however, running counter to the saving in the identification time) then we may have the rule for a "patterned" cluster: substitute the odd digits 1, 3, 5, 7, 9 if the digit in the units place of the entry-plot survey number is odd, and similarly for the even digits. And for still further spread, we may proceed by changing the digit in the ten's place. If any of the above rules gives a non-existent survey number (as sometimes happen) then just ignore that fact and treat the cluster formed by the remaining existent plots as the desired cluster. Thus the cluster size may occasionally be less than five, but the probability of selection will still be proportional to the area of this "numerically depleted" (and hence the "physically correct") cluster.

5.14. It will be noticed that the above method avoids the actual demarcation of the clusters on the village map, but what is more it eliminates the possibility of introducing any bias which might have arisen if the investigators were left to themselves to form the clusters with the mere restriction that each should be of a specified size (5 or 10 plots) instead of the present rigid rules where the investigator has no choice. The above procedure, however, still left something to be desired and it was necessary to determine how this could be achieved. There were still two difficulties, one pertaining to the frame of plots and the other to the arithmetical complexities at the estimation stage.

5.15. Firstly, it is necessary to know the areas of the plots constituting the clusters. The investigators were expected to collect these by local enquiry—either copying the area figures from the village record (if any, and easily available), or ascertaining it from the local people. This arrangement was not quite satisfactory because our earlier experience in the Bengal Crop Survey indicated that the reliability of such collected figures may be comparatively poor[30]. In fact, in the later crop surveys in Bengal the collection of area figures by local enquiry was replaced by the measurement of plot areas (called "area extraction") in the laboratory directly from the maps, and this was definitely more satisfactory. Unfortunately recourse to this procedure was not possible in the NSS as the village maps were not available in the Laboratory.

5.16. The second difficulty lay in the calculation of the proportions of the total area of a cluster under different types of utilisation—a necessary step in the estimation of the area under different utilisation for the whole country. The "cluster-proportions" had to be obtained by weighting the "plot-proportions" by the corresponding areas. (This is known as "anna-conversion" in our laboratory parlance). It may be pointed out here that what the investigators noted against each plot in the selected clusters were the proportions (in annas i.e., units of 1/16, and hence called "anna proportions") of the area of the plot under different utilisation types,

[30] This is one of the reasons why selection of plots with equal probability is not adopted in the NSS. The investigators directly note the proportion of the plot included under any specified land utilisation, and when plots are selected singly or as we shall see later on in quasi-compact clusters with probability proportional to area the collection of plot-area figures become superfluous.

and not the actual corresponding area in acres. This practice was followed because we were convinced from our experience in the Bengal Crop Survey that investigators were in a better position to assess correctly the proportions, rather than the areas. It should be particularly noted that in view of the fact that in an overwhelmingly large number of cases the *entire* plot belonged to the same utilization type the estimation of the proportions presented no very great volume of work.

5.17. The genesis of the solution to the above two problems was really to be found in the current sampling plan described earlier. It will be remembered that a cluster was determined by means of an entry-plot, which in its turn was fixed by the centre of a circular hole on the perforated sheet and as a consequence both the cluster and the entry-plot were selected (with replacement and) with probability proportional to area.

5.18. Thus two unbiased village estimates ($\hat{Y}_{vt}$) were available, the first, viz.,

$$\frac{A}{n} S\left[\frac{\Sigma(ap)}{\Sigma(a)}\right]$$

deduced from the cluster-data and the other, $\frac{A}{n} S_e(p)$ from the entry-plot information. The latter would naturally be subject to higher sampling fluctuations, but would not be affected by bias arising from possibly faulty plot-area figures, for in the second estimate the areas of entry plots were not at all required. It can easily be seen that not only the "area extraction" part of the job was thus eliminated by using the entry plot estimate but the question of "anna-conversion" moreover did not arise at all.

5.19. This gave the clue to the next step for meeting the requirement that plots be selected in clusters so as to properly reduce the "identification" time. The important step was to recognise that the main hurdle lay in the tacit assumption that the clusters should strictly form a geographically compact block. It can easily be realised, however, that from the point of view of identification time a semi-compact cluster was equally good.

5.20. Once this real need was clearly recognised the final step lay in introducing another stage of sampling. In the first stage a cluster had to be selected with a probability proportional to area, and in the second stage a fixed but adequate number of plots, n', had also to be individually selected with replacement with probability proportional to area, care being taken to see that a quasi-compact cluster was really formed. Both "area-extraction" and "anna-conversion" were no longer required, as can be easily seen from the formula for unbiased village-estimate, viz., $\frac{A}{nn'} S'\Sigma'(p)$. This method was partially tried with success in a special crop survey in West Bengal during the Jute-*Aus* Season in 1952. Further examination as to how far the investigating team can be depended upon in the selection of the ultimate plots is necessary before adopting it for the all-India survey. It is, however, possible

that after detailed study of this peculiar practical difficulty selection of plots singly with probability proportional to area may be preferable.

5.21. Quasi-compact clusters are likely to be more efficient than absolutely compact clusters of the same number of ultimate units (plots) actually surveyed. The optimum size and degree of compactness of a cluster for fixed cost has not so far been determined.

5.22. There is another aspect of the survey design to which attention may perhaps be drawn. In the current plan three different, but not independent, estimates of the area under different types of utilisation may be obtained. We have already referred to the first two (para 5.18). A third, viz., $\frac{A}{n} S_s(\epsilon)$, is obtained very simply on the basis of mere counting, from the data showing the utilisation of the "spot" determined by the random point fixed by the centre of a random perforation in the selection-sheet (col. 12, schedules 4, General Report No. 1).

5.23. In the present method it is possible to get some indication of the quality of the field work of an investigator by testing whether the three estimates are mutually consistent; but this has not yet been explored. The appraisal of his performance may perhaps be better made by studying the magnitudes of the correlations between the three sets of "observations" recorded at every random point, and comparing them with certain norms set up on the results of several rounds of the survey. Unfortunately, we have not had leisure so far to pay sufficient attention to the development of the necessary methodology.

6. HOUSEHOLD ENQUIRY : STRATIFICATION AND SELF-WEIGHTING SYSTEM

6.1. In the last few paragraphs we took up the question of sampling within a sample village as far as the land utilisation survey is concerned. We next turn to the other facet of sampling within villages; here we are concerned with sampling the households. Before proceeding further it is necessary to appreciate one peculiar feature of the household enquiry so that some of the steps which we took in sampling households (in the post-primary stage) may not appear to be refinements of mere academic interest, that is, devoid of much practical importance.

6.2. We are here referring to the problems arising from the smallness of the (household) sample size within a sample village. (For a given cost there is some reason to believe that generally in a multipurpose survey "optimum" results are approached by having a "very small" number of households for any *single* purpose in each of a 'large" sample of villages.) Of course, there are other reasons for adopting the steps referred to in the previous paragraph but this is the one which we have to face continually. The number of households taken up for the consumer-expenditure part of the enquiry is round about only five in a sample village for a single round; and a different set (but of the same size) of households is taken for the enterprise part. In the first round the sample size was still smaller.

6.3. The wide variation from household to household, even within the same village in respect of economic enterprise and consumer expenditure (which form the major part of the enquiry) called for stratification (or some similar device) of the households for purpose of sample selection. It is pertinent to point out that such stratification would enable us to provide separate figures for the different classes (enterprise classes and/or expenditure levels if these form the basis of stratification) of the population with a margin of error smaller than what would otherwise be possible.

6.4. In the first round, in the selection of a sample of fixed number of households from each selected village, the stratification used was the two important groups "agricultural" and "non-agricultural"; and greater representation was given to the "non-agricultural" group because of the greater variety within its fold, so that tolerably good estimates of certain selected characteristics for each group may be obtainable if desired. In the tabulation stage, however, this necessarily introduced some labour for obtaining unbiased estimate $\hat{Y}_{vt}$ of the various characteristics Y_{vt} for the general household population as the two groups had to be differently weighted. Therefore this procedure was omitted in the second round of the survey and an unstratified random sample of households was taken from each selected village. While this simplified the tabulation work, it lost the advantage of stratification, and the question naturally arose whether, without introducing any additional complications in the tabulation stage, the sampling plan could be so modified that some advantage of stratification was still retained.

6.5. A solution which at first sight appeared to be natural was to make proportional allocation to the different strata. This would have been quite feasible if the size of the sample for each selected village were adequate. But in the NSS, this sample size within the village is so small that the necessary approximations would really make the allocation non-proportional, (and more so, if the number of strata were increased for further gain in efficiency), with the result that the tabulation work will not really be simplified.

6.6. A compromise solution, adopted in subsequent rounds, was to arrange (or rather to assign a continuous serial number which gave effect to the desired arrangement) the households in such a manner that all the agricultural households occurred first and then the non-agricultural households. A systematic sample of the desired size (see however paragraph 6.8) was then taken with a random start. It will be noticed that the effect of this procedure was to make *almost* proportional allocation, simplify the work at tabulation stage, and from the field point of view giving perhaps a somewhat simpler method of selecting sample households.

6.7. It will be further noticed that stratification to a greater extent may be easily availed of without practically any additional work. Care should, however, be taken to ensure that the number of strata is compatible with the size of the sample. Where stratification is on the basis of a quantitative character, the class-intervals defining the strata should be so fixed that approximately equal number of households fall in each class.

6.8. As the total number of households in a selected village was rarely a multiple of the desired sample size, some of the possible systematic samples (in the usual sense) from a village varied by one unit from the desired size. The disturbance arising from this source may be particularly serious in view of the smallness of the sample size of any selected village. It was, therefore, decided to assume further, for purposes of selection, that the households were arranged in a circular order, and then there was no difficulty in selecting exactly the desired number of households with a random start, without of course, complicating the process of estimation.

6.9. From the point of view of tabulation although the above method was an improvement, it was not quite what the man in charge of tabulation would desire. To obtain unbiased estimates of "national totals", for example, every sample village would still have a different inflation factor (in place of original two) and would involve a very substantial volume of work. What appeared to be necessary was a self-weighting sample design. While in principle this was quite feasible being merely the proper fixation (in the Laboratory before the field-work commenced) of the proportion of households (or plots) to be taken up for the actual survey for each of the sample-villages, nonetheless a number of problems was thereby raised.

6.10. The first important departure from the original design, due to the introduction of a self-weighting system, was to remove the very desirable restriction of having the same work-load in every village. This change made the field operations somewhat more difficult. Thus the adoption of the self-weighting system shifted, but with justification, the emphasis to simplicity of tabulation from that of field operation. It must, however, be admitted that this did not quite meet the real problem which was to develop a method which while retaining the advantages of a fixed work-load per village would at the same time simplify the tabulation programme, of course, the whole thing being judged on the basis of the accuracy attained in the final results.

6.11. Another problem requiring solution in the actual attainment of a self-weighting system was, what may be called, the problem of "rounding-off". To secure an exact self-weighting system for the household enquiry it would be theoretically necessary to choose in almost all cases a fractional number of households; and this was quite meaningless in practice. The natural step would be to round off to the nearest integer, and ignore any bias that may possibly be introduced thereby. Such a step would no doubt be fully justified from a practical point of view provided the sample size was large compared to the rounding-off errors.

6.12. But in the NSS such a happy situation did not exist; and the difficulty would appear to be still greater when it was remembered that the small number of households sampled in any selected village were not all taken up for every one of the schedules. The stipulation, namely, that no household taken up for consumption enquiry should be taken up for enterprise enquiry augmented the rounding-off error, it being now necessary to round off to the nearest even number, equal number of households being allotted to the two enquries. And when in one of the rounds

it was desired to collect certain items of information on the consumption schedule, once on a monthly basis from one set of households, and then on a weekly basis from a different set of households, the rounding-off had to be done to the nearest multiple of three.

6.13. The problem, therefore, was to adopt some artifice which would circumvent this rounding-off error and retain the advantages of a strictly self-weighting system. The solution is very simple; the rounding-off is done either to the nearest smaller or larger integer[31] (or multiple of 2 or 3 as the case may be) by a random process, which assigns different probabilities of selection to the two alternatives in such a manner that the expected number of households is precisely equal to that desired to secure exact self-weighting.

6.14. There were a number of other issues raised by this problem of self-weighting. Only one will be discussed here. The difficulty experienced in some cases was that self-weighting demanded a larger number of households than what the sample village was actually inhabited by. Obviously one cannot choose more households than actually existing, but then it is tacitly assumed that the same household is not to be taken more than once. Once this restriction is removed there is no difficulty—that is, sampling is to be done with replacement.

6.15. But for a fixed sample size, random sampling without replacement is generally preferred to that with replacement because of the resulting loss of efficiency due to repetition. It appears, therefore, that repetition should be avoided whenever possible, and evidently this would not be secured if the investigators are asked to sample the households with replacement in every selected village. It will be noticed, however, that systematic sampling taken in a circular order described earlier, which was adopted for other reasons, would give rise to repetition only when the sample size exceeds the total number of households in the village.

6.16. Reverting back to the original design where we gave a greater weightage to the non-agricultural population it may be pointed out that one aspect of the problem has not been discussed as yet. The sampling plan developed so far ensured almost proportional representation to the different strata. But it is conceivable that it may be necessary to attach different weights to different sections of the population, either to improve the estimates for the general population or to secure adequate representation of the different sectors so that separate estimates of reasonably good precision may be provided for each of these sectors. Thus to secure better estimates of the total production of cereals, it may be desirable to give greater representation to the bigger farmers ; or to secure reasonably good estimates for individual enterprises, it would be necessary to give greater weightage to the numerically smaller groups.

---

31 Suppose that to secure exact self-weighting it is theoretically required to select $2\frac{9}{17}$ households. Then choose a random number in the range 1 to 17; if the random number falls in the sub-range 1 to 9 round off to 3 households, otherwise to 2 households only.

6.17. It has, however, to be kept in mind that the above requirement is to be secured with the minimum increase in the volume of not only the field work but also of the work at the tabulation stage and this is the crux of the problem. Really no new principles are involved. All that is necessary is to treat the different sectors separately and give each one its proper representation, but ensure that the sample for each sector is separately self-weighting. It will be noted however that computation of estimates, for the general population, of items common to more than one sector will be slightly more complicated than that for the "almost" proportional allocation.

6.18. Speaking about the problem of self-weighting sampling system, we had tacitly assumed that it had reference only to the question of estimating the population totals of items of interest e.g., total expenditure on food items, total production of cereals, total number of unemployed persons, total acreage of cultivable waste, etc. There is, however, another class of estimates, which is of considerable importance ; these are the estimates of certain population ratios; for example, per capita consumption of cereals, average size of an agricultural household, cost of cultivation per acre, etc. Naturally the sampling plan must have a sufficiently simple procedure for obtaining estimates of such ratios.

6.19. The method actually adopted is the combined ratio method of estimation, which merely involves the calculation of the ratio between the estimated totals of the numerator and the denominator of the ratio which is being estimated. The introduction of self-weighting system for estimation of population totals naturally simplified the estimation by the combined ratio method also.

6.20. The adoption of the combined ratio method of estimation raises the question of the bias of the estimate. The use of ratio methods has recently found wide acceptance in sampling practice and it is generally believed that the bias is negligible when the sample is large. This belief is based partly upon experience and partly on theoretical discussions, but unfortunately at none too rigorous level. Whatever theoretical discussion is available in published form does not fit in directly with the somewhat complex sampling design adopted in the NSS. But it can be shown that the NSS design is amenable to similar treatment.

7. SAMPLING WITH REPLACEMENT

7.1. In developing the sampling system there was another objective which had to be kept in view. There must be adequate provision for estimating the sampling errors, and that with the minimum amount of computation.

7.2. In the first round itself the sample design was fairly complex: a multi-stage scheme with stratification at more than one stage. In subsequent designs the complications, in a sense, were increased; these were increase in number of stages,

higher degree of stratification of households, fixation of sample size (household) by a random process, introduction of systematic sampling in the selection of households, etc.

7.3. At first sight it may appear that with such complex multi-stage designs method of obtaining unbiased estimates of the sampling variance would be difficult, and in any case very heavy computational work will be involved. But there is a safety valve, namely, the selection of first-stage units with replacement, and this has very wide implications.

7.4. This single precaution reduced, in effect, the whole problem of estimation of sampling variance of an estimated total to one of unistage sampling[32]. The whole basis of obtaining an unbiased estimate of the sampling variance of estimated stratum total is extremely simple. Corresponding to every first-stage unit selected within a stratum it is possible to obtain, on the basis of the characteristics of the ultimate units actually surveyed within that first-stage unit, an unbiased estimate of the stratum total. And as the n, say first-stage units are selected with replacement, these estimates ar all independent of one another, it being assumed, as is the normal practice, that the sub-sampling within any first-stage unit is independent of that within another. We are thus provided with n independent values from the population of such estimates and therefore the applicability is ensured of the well-known formula for obtaining an unbiased estimate, of the sampling variance of the mean (of the above n independent values), which in its turn is an unbiased estimate of the stratum total.

7.5. It will also be noticed from the argument given above that if each first-stage unit provides a biased but independent estimate of the stratum total, then also an unbiased estimate of the sampling variance is obtainable in precisely the same manner. Mean-square-error which is equal to the sum of the sampling variance and the square of the bias will not, however, be estimated unbiasedly thereby.

7.6. The second advantage accruing from sampling the first-stage units with replacement will be now explained. It is well known that a strictly unbiased estimate of the sampling variance of an estimated total is unobtainable when a single element is selected from each stratum. Although the above statement is universally true when the stratification is of the first-stage units but this is not so when it is at the second stage, that is, within each first-stage unit (like the stratification of households within a sample village adopted in the NSS). The statement is, no doubt, true when the sampling is without replacement at the first stage, but certainly not true when the first-stage sampling is with replacement. The same is also true if a single systematic sample with a random start or any other single cluster or even a single unit is chosen at random in the post-primary stage, even though there is no stratification at that stage.

---

[32] This important aspect of sampling the first-stage units with replacement was referred to in our note on the NSS forwarded in November 1952 to the United Nations Statistical Office in New York for incorporation in their Report on Sample Surveys of Current Interest.

7.7. Selection of single second-stage unit within every selected first-stage unit, for example, selection of a single household from every sample village, is not really so absurd in practice as it many appear on first thought. In a multipurpose survey we may select from each sample village several households, but each for a different purpose. For example to study the average characteristics of a household in each expenditure group, or each enterprise class, or each agricultural holding-size it may be desirable to select one household from each of these groups in a sample village. Thus for any one purpose a single unit (household) is selected in the post-primary stage.

7.8. From the argument given earlier it is evident that a strictly unbiased estimate of the sampling variance is obtainable in the cases where the first-stage units are selected with replacement as every first-stage unit provides an independent estimate of the population characteristic under consideration. This introduces great flexibility in the post-primary stages which can be utilised to increase the precision of the estimates or provide operational facilities like the use of systematic samples instead of random ones.

7.9. It may incidentally be pointed out that, under certain circumstances, sampling the first-stage units without replacement may give a lower standard error of the estimates than the corresponding sampling with identical sub-sampling system but with a replacement scheme for the first-stage units. But even in such cases, if provision for unbiased estimate of the sampling variance is desired, and if absolute identity of the sub-sampling system in the two alternatives is not insisted upon, sampling with replacement at the first stage may yield more precise estimates than that provided by the without-replacement scheme, by virtue of the gain in precision, as explained in the preceding paragraph, resulting from the use of one unit per stratum or like device in the post-primary stage, which is admissible only in the with-replacement scheme but not in the without-replacement one where at least two units are to be selected from each stratum.

7.10. The possibility of selecting, for example, from each sample village a single household for a specified purpose in a multipurpose scheme but at the same time providing for unbiased estimation of sampling variance when the sampling of first-stage units is with replacement, opens up the possibility of increasing for the same cost the number of first-stage sample units beyond the number which the without-replacement-scheme cannot necessarily exceed because of the restriction of having at least two second-stage units within each first-stage sample unit. Thus under certain circumstances the with-replacement scheme may yield more precise results than the without-replacement one.

7.11. While sampling with replacement introduced a good deal of advantage for reasons explained above, it was not sufficiently so for the estimation of error for a survey of this magnitude where hundreds and hundreds of different items were involved. There still remained two main drawbacks. Firstly, there is the necessary step of squaring which still remained of such dimensions that it is possible to do

justice to only a very few selected items, with the time and resources at our disposal. Secondly, the advantages of sampling with replacement are separately non-existent when the combined ratio method of estimation is used for estimation of per capita consumption for cereals and similar ratios; because this advantage hinges upon the possibility of expressing the estimate under question as the mean of independent estimates of the same character provided by each of the first-stage units, and obviously such possibility is non-existent in the present case.

7.12. Leaving apart the question of bias, which may possibly be negligible in large samples, no unbiased method of estimation of the sampling variance is available in the literature on sampling theory for the combined ratio estimate. There are, however, certain methods which furnish "approximate" (in some peculiar sense) estimates of the sampling variance, and these methods are currently being used everywhere for the estimation of sampling error. Once we concede the use of such methods the second drawback ceases to exist, for it is now possible to derive the estimate of "approximate" sampling error on the basis of the "totals" for each first-stage unit included in the sample. This unfortunately is not the end of our troubles, for now we are left with not only the problem of determining a large number of sums of squares, but also with the labour involved in obtaining certain sums of products.

8. INTERPENETRATING SUB-SAMPLES

8.1. In view of the practical difficulties pointed out at the closing stages of the last section it is important to use such methods of estimation of sampling error which will demand a comparatively small volume of computational work, even at the risk, if necessary, of a little inefficiency compared to the estimate derivable by more laborious processes.

8.2. Some progress has been made towards the solution of the above problem but a good deal more work is necessary. The first step found necessary is to choose a sample design such that an equal but adequate number, say n, of first-stage units are sampled *with replacement* from each (primary) stratum.

8.3. This step incidentally brings in the question of defining the strata in such a manner that an equal number of first-stage units will, to some extent, improve the precision of the estimates. It is likely that the principle of equalisation of the "size" of each (to which we have referred earlier) stratum will serve the purpose. We have already mentioned that with this purpose in view attempts were made to equalise the size of strata on the basis of total consumer expenditure in later rounds of the survey[33].

[33] Although we have not so far come across such a situation in the NSS, it is well to point out that when the basis of stratification is such that equalisation of "size" is not possible so that the selection of an unequal number of first-stage units seems more desirable, the sampling plan requires some modification. This is discussed later in para 9.13.

8.4. Consider the case in which a single first-stage unit is selected with the appropriate probability from every one of the strata. Such a set may be called a "distributive" sample. It should then be possible to obtain estimates of a population total (or ratio) from this set of first-stage units (or, rather from strata estimates based on the data provided by the ultimate units actually taken up for investigation from each of these selected first-stage units). Now, when n units are selected with replacement from every stratum, it is possible to form n independent distributive samples, that is, sets of the above type. In other words, the entire sample is broken up into n independent interpenetrating sub-samples. And each such sub-sample provides us with an estimate of the characteristic under consideration.

8.5. Thus n independent estimates are available, it being supposed, as is the normal practice that selection of second or higher-stage units within any first-stage unit is independent of that within another. The mean of these estimates is then itself an estimate of the population total (or ratio) under consideration, and not only this, it is also obviously possible to obtain an unbiased estimate of the sampling variance of this mean from these n estimates.

8.6. It may also be pointed out here that the n replications may be arranged as an interpenetrating net-work of sub-samples with different field, processing, and tabulating teams for the different sub-samples, and it cannot be over-emphasised that this, as is well known, helps to control various kinds of errors[34] introduced in the above phases of the work. Interpenetrating sub-samples are being invariably used from the second round and the results are being separately tabulated. A great deal of information on the margin of uncertainty has already accumulated, and it is hoped that some of the results would be published at an early date. This method is used, for example, in the fourth round to compare the effect of two different periods of reference (week and month), in respect of the consumption of a number of articles.

8.7. Before entering into a general discussion into the merits and demerits of this procedure it is important to examine the implications of this method in relation to a combined ratio estimate. The first change brought about is in the method of estimation of the ratio itself; as pointed out earlier (para 8.5) the estimate itself is no longer a single ratio but the mean of a number of ratios. The effect of this procedure on the magnitude of the bias cannot be precisely assessed in the present state of our knowledge. There appears to be some possibility of the bias increasing, but if the number of strata is large, as happens to be the case in the NSS. the bias may not be serious.

---

[34] The technique of interpenetrating sub-samples was introduced by Mahalanobis in 1939 in The Bengal Crop Surveys mainly for statistical control. Usually two interpenetrating sub-samples were taken which supplied two estimates and the probability that these two estimates (when independent) "enclose" the "true" value when the size of the sample is large was 50%. Mahalanobis, P. C., Recent experiments in statistical sampling in the Indian Statistical Institute, *J. R. S. S.* 109(4), 1946.

8.8. Leaving apart the question of bias, the procedure introduces a very striking change in relation to the problem of estimation of sampling variance of an estimated ratio. As has already been explained, this can now be obtained in an unbiased manner, and no "approximate" formula of doubtful validity need be employed.

8.9. We can now compare the advantages and disadvantages of the above suggested procedure with the more usual one of pooling the estimated variances within different strata. We shall first see its effect in regard to the volume of computational work. It is no longer required to obtain separate "totals" for each first-stage unit included in the sample but merely the "totals" for each of the n interpenetrating sub-samples, and this is operationally more convenient. The volume of work on squaring is reduced to a mere fraction ($1/h$, where h is the number of strata) of what it would have been otherwise. Moreover, the question of sum of products (for ratio estimates) does not arise.

8.10. While the method of interpenetrating sub-samples, no doubt, simplified the computational work, there seems to be considerable apprehension, as will be found in current literature on sampling theory, that the efficiency of the estimation of standard errors is considerably impaired due to the very drastic reduction in the degrees of freedom on which such estimates are based. This point requires careful consideration.

8.11. The use of the "standard error" in measuring the precision of an estimate may be more or less appropriate in sample surveys because of the largeness of the sample. It must however be borne in mind that in interpreting the significance of the estimated standard error two assumptions are involved. Firstly, the sampling distribution of the estimate is normal, which for the type of populations commonly met in practice is approximately true for the large sample sizes normally used. Secondly, the estimated standard error should be sufficiently precise. Failure to ensure the second condition may under certain circumstances, not quite uncommon, give an erroneous impression of the precision actually attained in any specific case.

8.12. In raising the above question our immediate object is not to dwell upon the role played here by the kurtosis of the original distribution, but merely to point out that the large sample interpretation of the estimated standard error is no longer applicable when this estimate is based, in the manner described earlier, on a small number (as it should be at least for practical convenience) of interpenetrating sub-samples, because then for the resulting smallness of the number of degrees of freedom the standard error is ill-determined. We shall, therefore, make a direct approach to the problem of setting confidence limits to the population total or mean. And, as we shall employ the normal theory distribution of Student's ratio for this purpose, we shall naturally have to take account of the departures (in skewness, kurtosis etc.), of our populations from normality.

9. CONFIDENCE LIMITS AND NON-NORMALITY

9.1. In any extensive survey like the NSS it is almost always necessary to use stratified sampling. To start with, however, we shall find it instructive to see how parallel ideas work in the case of unstratified unistage random sampling. To simplify matters we assume that a random sample of v observations are drawn from an infinite non-normal population. It is proposed to set confidence limits to the mean by making use of the distribution of Student's ratio for the normal case.

9.2. In addition to the usual approach to this problem there is an alternative approach, namely, the approach of grouping of observations. Here the observationss ($v = nk$) are randomly grouped into n set of k observations each. We now work with the mean (or total) values of the sets. Thus our basic distribution is no longer the original distribution, P_1, but the distribution, P_k, of the mean of random sample of k observations from the original population. Now the second distribution, P_k, as is well known, deviates from normality to a smaller extent than the original distribution. In consequence, the normal theory of sampling distribution of Student's ratio is likely to be a better approximation to the actual sampling distribution of the ratio for P_k than that for P_1. And as such, confidence limits based on the grouped data (P_k) are likely to reveal the position more correctly at the confidence level given by the normal theory.

9.3. From the point of view of minimising the effects of non-normality it appears from the above discussions that the larger the size of the groups the better is the situation. But there are other considerations as well. The lowering of the degrees of freedom on which the estimated variance is based will also lower the efficiency of interval estimation. But one should not be unduly perturbed over this lowering of degrees of freedom when the object is setting confidence limits, as we shall presently indicate. Even when the original distribution is perfectly normal the lowering of the degrees of freedom by using the grouped method does not entail any great loss of efficiency for all practical purposes even when a small number of groups are taken.

9.4. It is important at this stage to give some idea of the magnitude of the loss of efficiency as judged by the increase in the average length of the confidence interval, of course at a specified confidence level when the distribution is normal. Thus, working with a 95% confidence coefficient it can be shown that if a (large) sample is broken up into 4, 5, 6, 10, 20 or 30 groups then the corresponding increase of the average length of the central confidence interval is respectively about one-half, one-third, one-fourth, one-seventh, one-twentieth and one-thirtieth of what it would have been for the usual ungrouped method. For 2 and 3 groups there is considerable loss; for 2 groups the average length jumps up to more than five-fold, and for 3 groups to about two-fold of that for the ungrouped method. The loss is slightly less than what has been stated above if the original sample is comparatively small and consequently the size of a group is also so.

9.5. Turning now to non-normal distributions, which are more likely to occur in sample surveys, the "gain", in relation to the application of the normal theory distribution of Student's ratio, due to approximation to normality induced by the method of grouping, may more than offset the "loss" due to the lowering of the number of degrees of freedom. The optimum extent of grouping will depend upon the nature and degree of non-normality.

9.6. In other words the effectiveness brought into the confidence statements by grouping the data and hence making them conform more nearly to the basic assumption of normality (involved in the application of the normal theory t-distribution) may be greater than the effectiveness gained by using the ungrouped data with a larger number of degrees of freedom but affected by a greater deviation from normality. With ungrouped data the average length of such confidence interval at the *supposed* level may be slightly shorter (as far as can be guessed from the behaviour for the normal case discussed earlier in paragraph 9.4) than that for the grouped method at the same *supposed* level, but compared to the former in the latter case the *supposed* level may be much nearer to the *real* level because of its closer approximation to normality. Much deeper probing is needed into this question of average length than what we have been able to give it so far; indeed, we are not even quite sure that the average length of confidence intervals for ungrouped data will necessarily be shorter, even at the same supposed confidence level, than the corresponding interval given by the grouped method in all non-normal cases, the confidence limits being however set (by the usual method) as if the distributions were normal.

9.7. We now turn to the case of stratified sampling which is of special relevance to any extensive sample survey like NSS. We shall for the sake of simplicity restrict ourselves to stratified unistage random sampling with replacement. In the first instance, in the usual method, the estimated sampling variance is a linear function of the estimated variances for the different strata. Even in the case where the "observations" within different strata are supposed to be normally distributed, with possibly unequal means and variances (which is very much likely to occur to a fairly large extent in a country of the dimensions of India) the distribution of this complex estimated variance will also be necessarily very complex and the use of the (normal theory) t-distribution may not fully be justified. On the other hand, it will be noticed that in the case of the proposed procedure (see particularly paras 8.4 and 8.5) the t-distribution is fully justified under the assumption of normality for each stratum.

9.8. It is true that the distribution within any stratum may be and usually is moderately and even extremely skew for a large number of economic and social variables. It is also equally true that the definitely leptokurtic and to some extent platykurtic distributions are much more common than approximately mesokurtic ones. Thus the distributions within the strata deviate both in nature and in degree from normality ; and the question arises how far the proposed method overcomes this difficulty.

9.9. Of course, the proposed method does not attempt to transform the non-normal strata distributions to normal ones. But fortunately in the application of the normal theory distribution of Student's ratio in the suggested method, we are not directly concerned with the nature of the distributions within different strata. We are basically concerned with the population, π_h, of estimates of the mean derivable from a "distributive" sample of h units, the units being distributed over the h strata — one at random from every stratum. Each of the n independent interpenetrating sub-samples obtained by the suggested method is a "distributive" sample in the sense explained in this paragraph and therefore provides n independent "observations" from the population π_h, each "observation" being an estimate of the mean in question.

9.10. Now our "observations" are a linear combination of independent contributions from each stratum (it is assumed that the samples are drawn independently in different strata), and as such in spite of the non-normality within strata its distribution will be approximately normal if, h, the number of strata is moderately large. In other words, the population, π_h, is approximately normal, and as such the normal theory confidence limits to the mean based upon n random "observations", provided by the n independent interpenetrating sub-samples are likely to reflect very nearly the *real* situation.

9.11. For stratified sampling, it is seen from earlier discussions that, to set confidence limits, the real issue is not whether to estimate the standard error by the classical method or by the suggested method. The latter method may be usefully employed, for it has both theoretical and practical advantages.

9.12. The real issue is how many replications, that is, how many interpenetrating sub-samples should we use; or, what is the same thing, to what degree of stratification should we proceed ? On the one hand, an increase in the number of replications, and, therefore, in the degrees of freedom, increases the efficiency of interval estimation; but, on the other hand, the consequent decrease in the number of strata, which in the extreme case may even be one (i.e., no stratification) will lessen the precision of the estimates. Moreover, an increase in the number of sub-samples will also mean some increase in the volume of work. Considering these factors and remembering that with the method of grouping, even for a normal population, a small number of degrees of freedom is not a serious handicap (para 9.4); and that for non-normal parent populations, generally speaking, the applicability of the t-distribution is more and more justified as h increases, a small number of sub-samples would seem to be quite adequate.

9.13. In stratified sampling when the number of strata is small, a procedure (which may perhaps be described as a combination of the method for the unstratified case and that for the stratified case given earlier) may be desirable to justify to a greater extent the applicability of the normal theory distribution of Student's ratio. A similar procedure is also desirable when the number of units to be selected from different strata are unequal, even when the number of strata is large. Thus if

$nk_1, nk_2, \ldots, nk_h$ units are to be selected from h strata then the entire sample is to be broken up into n independent sub-samples, each sub-sample being made up of $k_1, k_2, \ldots, k_h$ units selected respectively from the different strata. It will be noticed that it is only necessary to ensure that the sets, each consisting of $k_1, k_2, \ldots, k_h$ units from the respective strata, are to be selected with replacement, but it does not mean that the k_i units to be selected from the i-th stratum, for any specified set, are also to be selected with replacement.

9.14. Indeed, subject to the restriction that each set provides us with a valid and independent estimate of the population characteristic under consideration, there is complete freedom to define the "cluster" of k_i units in any manner one pleases, and naturally so as to improve the efficiency of the design. It may sometimes be worthwhile to group together the strata (including the case where all the k's are equal) so as to improve the efficiency of the design. We have previously suggested that selections within any stratum should be independent of that within another. We can relax this condition to the extent that selection within any group is independent of that within another, and thus give some freedom of selection within a group to improve the efficiency.

9.15. The units (or ultimate units, if a multi-stage sampling is considered) within a group for a given set (of an interpenetrating sub-sample) may be regarded as a single "complex" unit. Thus if the number of groups is large our interpenetrating sub-samples will each consist of a large number of complex units independently selected and therefore the population, π_h, (of para 9.9) may still be considered to be approximately normal.

9.16. The considerations stated in the last few paragraph favour the use of independent interpenetrating sub-samples in setting confidence limits to an estimated total or mean. We have invariably employed, in the manner described earlier, the normal theory distribution of Student's ratio in setting these limits. It is conceivable that, for any specified characteristic, if the nature of its distribution is more accurately known then a better method of calculating the limits may be used by a suitable transformation or some other device. But in a sample survey where several items are to be estimated it is hardly practicable to pay individual attention to particular items, and a general method for all items is desirable.

9.17. The suggested procedure no doubt simplifies the calculations for setting the confidence limits, but with hundreds of different items under study, round after round, it is desirable to simplify the computational work still further, with as small a loss of efficiency as possible. For this purpose the characteristics of the population which are being estimated may conveniently be classified into three categories according to their importance.

9.18. Firstly, for the most important group of items the standard error may be estimated by the usual process of squaring the estimates provided by each of the independent interpenetrating sub-samples, and the confidence limits may

be set on the basis of the t-distribution. Only extremely important items should be placed in this group.

9.19. Secondly, for moderately important characteristics the confidence limits can be set on the basis of the u-distribution (distribution of the ratio of a normal deviate to estimate of standard error based on the range). It will be noticed that hereby we dispense with the business of squaring. The loss in efficiency is practically of no importance if the "observations" (the estimates provided by the independent sub-samples) are small in number as they would be (usually in the proposed estimation plan). Alternatively, when the number of observations is large, the method of mean range may be used, but this is not likely to occur in practice.

9.20. For the third and last group, even the use of the u-distribution is dispensed with, and here the range between the extreme values (in general between a pair of order statistics) of the estimates provided by the replicated interpenetrating sub-samples would directly give the confidence interval of the median at an appropriate level of significance depending on the number of replications. For example, five sub-sample estimates would provide the 93.75[35] per cent confidence interval of the median (which may be taken to be practically the same as the mean). Usually the distribution of the sub-sample estimates may be supposed to be normal, and this provides a sufficient, although not necessary, condition for the identity of mean and median.

9.21. It may be further noted that in the publication of the results great simplification is hereby possible. It appears to be a good plan to publish all the four or five estimates (it is tentatively decided to have in future four or five independent interpenetrating sub-samples in the NSS instead of two such sub-samples as at present). For the first group, in addition to the mean of the four or five estimates its standard error may also be published. For the second group only the mean; and for the third group not even the mean need be published. Any one specially interested in any particular item may himself calculate the mean, which is obtained quite easily as only four or five different figures are involved, or to use the slightly less efficient estimate, namely, the median by inspection (for which an odd number of sub-samples will be slightly more convenient).

9.22. It will also be noticed that when all the sub-sample estimates are published it would be possible, with increasing labour of computation, to obtain more and more efficient modes of calculating the confidence interval, by using the range, the u-distribution and the t-distribution respectively. Ordinarily the range should serve the purpose if only the corrresponding confidence co-efficient (93.75

[35] The probability that the five estimates are all below the median is $(\frac{1}{2})^5$; also the probability that the five estimates are all above the median is $(\frac{1}{2})^5$; therefore, the probability of the five estimates being either all below or all above the median is $(\frac{1}{2})^4$ (it being assumed that almost all estimates provided by all possible distributive samples are different from the median). Therefore, the probability that the range will cover the median is 0.9375. In general, for n sub-samples, the range would give the confidence interval at the probability level of $1-(\frac{1}{2})^{n-1}$. This method is being used in the Indian Statistical Institute for a long time.

per cent for five sub-sample estimates, or 87.5 per cent for four such estimates) is considered quite suitable.

9.23. Another reason in favour of the publication of all the independent sub-sample estimates will now be explained. Sometimes the user of the published results may be interested in setting confidence limits to certain estimates which he can derive from the published results; the scope for the use of such "derived" estimates is particularly high in a multipurpose survey. Thus one may be interested in certain linear functions of the estimates of several different items. For example, the published results may show estimates of the expenditure on several differet items and one may be interested in the total expenditure on a group of such items; or, the published results show the distribution of agricultural holdings in certain specified size-groups, but one is interested in the number of holdings obtained by pooling together several of the large size-groups ; or the published results show the value of raw materials used and the value of the finished products in a certain class of manufacturing establishments but one is interested in the "value added". In such situations even if standard errors of each of the estimated items are published, it is not usually possible to estimate merely from such results the standard error of the "derived estimates". But if the separate sub-sample estimates are available then it is easy to set the confidence limits of such "derived" estimates by precisely the same method as outlined earlier. Even for non-linear "derived" estimates it may under circumstances be not quite improper to use similar method for setting confidence limits. It will also be noted that in a continuing survey like the NSS it is of considerable importance to study the change from one round to another. The confidence limits to such changes are easily set if the same number of sub-sample estimates are published for each round even in the case where the sample remains wholly (or partially) unaltered as it should be (if practicable), if the study of change were the sole objective.

10. CONCLUDING REMARKS

10.1. It should be pointed out here that the present exposition has been intentionally restricted to what may be called the general design of the household and land-utilisation enquiry of the National Sample Survey. From the first round work is also being done in the NSS on various other problems. Thus we have not discussed the work of the Crop Survey wing of the NSS,—acreage and yield surveys, preharvest survey for crop acreages, assessment of the grow more food campaign etc. Other activities of the NSS, like the sample census of manufacturing industries, unemployment survey etc., have also not found place here. The question of control of non-sampling errors has also not been taken up.

10.2. No attempt has been made to give a thorough mathematically demonstration of all the statements made in this note ; moreover, some matters directly

connected with the problems discussed or with the solutions offered have not been mentioned. Important elaborations of the ideas presented here are possible, but these also have been left out to save space, as the note is already long. Work is proceeding on other aspects, and much further research on the sample design will have to be done to meet new contingencies or to solve old but difficult problems.

10.3. The general approach in planning the National Sample Survey was naturally based on the extensive experience gathered by the workers of the Indian Statistical Institute in the field of sample survey since about 1935. The sample design (in its more restricted sense of the allocation and method of selection of the sample units etc.), of the first round of the NSS was broadly based on what may be called the traditions of the Institute[36]. As the work progresses from one round to another, continuing efforts are being made to improve the sample design in the light of past experience.

10.4. Thus the development of the NSS design evidently is not entirely due to the efforts of any single individual. Professor Mahalanobis has been the most inspiring figure with his valuable suggestions. Among others who deserve special mention are Shri S. B. Sen, Shri J. M. Sengupta, Shri N. C. Ghosh, Shri S. Raja Rao and Shri A. Ganguly who by their valuable discussions and other help made substantial contribution to the development of the design.

[36] In this connection a list of important papers by the workers of the Institute is given in the bibliography at the end.

LIST OF SELECTED PAPERS AND REPORTS PREPARED BY MEMBERS OF THE STAFF OF THE INDIAN STATISTICAL INSTITUTE—CALCUTTA

1. Study of consumer preference in Calcutta By P. C. Mahalanobis (Report submitted in 1935).
2. Preliminary report of the survey of handloom weaving industry in bengal By N. C. Chakravarti (Government of Bengal, 1937).
3. Statistical report on the experimental crop census, 1937 By P. C. Mahalanobis (Indian Central Jute Committee, 1938).
4. A note on grid sampling By P. C. Mahalanobis *Science & Culture*, iv(5), November, 1938, 300.
5. First report on the crop census of 1938 By P. C. Mahalanobis (Indian Central Jute Committee, 1939).
6. General report on the enquiry into the prevalence of drinking tea among middle class Indian families in Calcutta, 1939 By P. C. Mahalanobis (Tea Market Expansion Board).
7. Progress report on the jute census scheme for 1939 By P. C. Mahalanobis (Indian Central Jute Committee).
8. Report on the sample census of jute in 1938 By P. C. Mahalanobis (Indian Central Jute Committee).
9. A sample survey of the acreage under jute in Bengal By P. C. Mahalanobis *Sankhyā*, 4(4), 1940, 511-530.
10. Statistical note on crop-cutting experiments on paddy in Mymensingh By P. C. Mahalanobis (Submitted to the Government of Bengal, May 1940).
11. Statistical report on crop-cutting experiments on jute 1940 By P. C. Mahalanobis (Indian Central Jute Committee).
12. Preliminary reports on the sample census of the area under jute By P. C. Mahalanobis (Indian Central Jute Committee, August 1940).
13. A note on the expenditure on tea among working-class families in Howrah and Kankinara By P. C. Mahalanobis (Submitted to the Tea Market Expansion Board, October 1940).
14. Some problems of field operations in labour enquiries By A. N. Bose *Sankhyā*, 5(2), 1941, 229-230.
15. Report on the enquiry into prevalence of drinking tea in Nagpur By P. C. Mahalanobis (Submitted to the Tea Market Expansion Board, March, 1941).
16. Statistical survey of 'Public Opinion' By P. C. Mahalanobis *Modern Review*. April 1941. 393-397.
17. Preliminary statistical report on the regional survey of Borer Pests of sugarcane By P. C. Mahalanobis (Submitted to the Imperial Council of Agricultural Research, June 1941).
18. A statistical report on the rupee census By P. C. Mahalanobis (Published in the *Report on Currency and Finance*, 1940-41, By Reserve Bank of India, June 1941, 49-55).
19. Statistical report on crop-cutting experiments on jute in Bengal, 1940 By P. C. Mahalanobis (Indian Central Jute Committee, June 1941).
20. A method of estimating variance of sample grand mean and zone variances in unequal nested sampling By M. Ganguly. *Science & Culture*, 6(2), 1941, 724.
21. A note on random fields By P. C. Mahalanobis. *Science and Culture*, 7(1), 1941, 54.
22. Reports on programme preference and broadcast reaction, Calcutta By P. C. Mahalanobis (Submitted to the Government of India, July 1941).
23. Report on the sampling technique for forecasting the block yield of cinchona plants : experiments series B By P. C. Mahalanobis (Submitted to the Government of Bengal, October 1941).
24. A statistical note on nutritional investigations in College hostels in Calcutta By P. C. Mahalanobis. *Sankhyā*, 5(4), 1941, 439-448.

25. General report on the sample census of the area under jute in Bengal 1941 By P. C. Mahalanobis (Indian Central Jute Committee, 1941).

26. A note on nested sampling By M. Ganguly, *Sankhyā*, 5(4), 1941, 449-452.

27. Family budget enquiries of labourers, Jagaddal By P. C. Mahalanobis (Board of Economic Enquiry, Bengal, January 1942).

28. Sample surveys (Presidential address, Section of Mathematics and Statistics, Indian Science Congress 1942) By P. C. Mahalanobis.

29. Report on the precision of family budget enquiry at Jagaddal, August 1942 (Bengal Board of Economic Enquiry) By P. C. Mahalanobis.

30. Preliminary report on Burdwan-Hooghly-Howrah crop-cutting survey By P. C. Mahalanobis (Submitted to the Government of Bengal).

31. Preliminary report on the sample census of the area under jute in Bengal, 1942 By P. C. Mahalanobis (Submitted to the Government of Bengal).

32. Preliminary report on the sample census of the area under aus paddy in Bengal, 1942 By P. C. Mahalanobis (Submitted to the Government of Bengal).

33. Preliminary report on the crop-cutting experiment on jute 1941 By P. C. Mahalanobis (Submitted to the Government of Bengal).

34. A note on the sampling error in the method of double sampling By Chameli Bose, *Sankhyā*, 6(3), 1943, 329-330.

35. An enquiry into prevalence of drinking tea among middle class Indian families in Calcutta By P. C. Mahalanobis, *Sankhyā*, 6(3), 1943, 283-312.

36. Report on the survey of Borer Pests on sugarcane, Parts 1–9 By P. C. Mahalanobis (Submitted to Imperial Council of Agricultural Research).

37. Final report on the sample census of the area under jute and aus paddy in Bengal, 1943 By P. C. Mahalanobis (Submitted to the Government of Bengal).

38. Preliminary report on the sample census of the area under aman paddy in Bengal, 1943 By P. C. Mahalanobis (Submitted to the Government of Bengal).

39. On large scale sample surveys By P. C. Mahalanobis (*Phil. Trans. Roy. Soc.*, London, Series B No. 584, 231, 329-451), 1944.

40. Report on the Bihar crop survey : Rabi season 1943-44 By P. C. Mahalanobis, *Sankhyā*, 7(1), 1945, 29–106.

41. Sample surveys of crop yields in India By P. C. Mahalanobis, *Sankhyā*, 7(3), 1946, 269-280.

42. A sample survey of after-effects of the Bengal famine of 1943 By P. C. Mahalanobis, R. K. Mukherjea and A. Ghosh, *Sankhyā*, 7(4), 1946, 337-400.

43. Recent experiments in statistical sampling in the Indian Statistical Institute By P. C. Mahalanobis. *J.R.S.S.*, 109(4), 1946, 325-378.

44. Traffic census on new Howrah bridge By P. C. Mahalanobis (Report submitted to the Government of India, November 1946).

45. Crop-cutting experiment on sugarcane in a farm cultivation By J. M. Sengupta, *Sankhyā*, 9(1), 1949, 47–50.

46. Two dimensional systematic sampling and the associated stratified and random sampling By A. C. Das, *Sankhyā* 10(1 & 2), 1950, 95-108.

47. Age tables based on the Y-sample By P. C. Mahalanobis (Submitted to the Government of India, 1950).

48. Means of livelihood and industries tables based on the Y-sample By P. C. Mahalanobis (Submitted to the Government of India, 1950-51).

49. Survey of rural indebtedness By P. C. Mahalanobis (Report submitted to the Government of West Bengal, April 1950).

50. Survey of economic condition of agricultural labour By P. C. Mahalanobis (Report submitted to the Government of West Bengal, December 1950).

51. Estimation of parameters from incomplete data with application to design of sample surveys By Abraham Matthai, *Sankhyā*, 11(2), 1951, 145-152.

52. Some further results on errors in double sampling technique By Chameli Bose, *Sankhyā*, 11(2), 1951, 191-194.

53. On two phase sampling and sampling with varying probabilities By A. C. Das (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. II, 105-112).

54. Systematic sampling By A. C. Das (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. II, 119-132).

55. A method of sample selection providing unbiased ratio estimates By D. B. Lahiri (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. II, 133-140).

56. Experimental survey for the estimation of cinchona yield By J. M. Sengupta, I. M. Chakravarti and D. Sarkar (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. II, 313-331).

57. On the size of Sample cuts in crop-cutting experiments By P. C. Mahalanobis and J. M. Sengupta (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. II, 359-404).

58. An overall measure of precision of a sample table with applications in the study of relative efficiencies of different sampling units in population census By D. B. Lahiri and A. Ganguly (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. IV, 55-74).

59. The development of a system of agricultural statistics By S. B. Sen (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. V).

60. A note on the scrutiny of statistical results By N. C. Ghosh (International Statistical Conferences, 1951, *Bull. Int. Stat. Inst.*, vol. XXXIII, pt. V).

61. The National Sample Survey, General Report No. 1 on the first round—October 1950-March 1951 By P. C. Mahalanobis (Government of India, published in December 1952).

62. Some aspects of the design of sample surveys By P. C. Mahalanobis, *Sankhyā*, 12(1 & 2) 1952, 1-7.

63. Report of a sample survey of refugee households settled in rehabilitation colonies in West Bengal, December 1951 By S. B. Sen (Submitted to the Fact-Finding Committee in February 1953).

64. The National Sample Survey, Report No. 2. Tables with notes on the second round April-June 1951 (Submitted to the Government of India in August 1953 and published in December 1953).

65. On certain extended cases of double sampling By K. C. Seal, *Sankhyā*, 12(4), 1953, 352-362.

66. The National Sample Survey, Report No. 3. Tables with notes on the third round, August—November, 1951 (Submitted to the Government of India in August 1953 and published in March 1954).

67. On some aspects of the Indian National Sample Survey By P. C. Mahalanobis and S. B. Sen (International Statistical Conferences, September, 1953).

68. National Sample Survey, Report No. 4. Survey of persons in the 'live-register' of Delhi Employment Exchange, September 1953 By Pitambar Pant (Submitted to the Government in December 1953 and published in March 1954).

69. Sample survey of displaced persons in the urban areas of Bombay State : July—September 1953 By S. B. Sen (Submitted to the Ministry of Rehabilitation in February 1954).