PHILOSOPHICAL TRANSACTIONS

OF THE

ROYAL SOCIETY OF LONDON

Series B. Biological Sciences

No. 584 Vol. 231 pp. 329–451 31 October 1944

# ON LARGE-SCALE SAMPLE SURVEYS

By

P. C. MAHALANOBIS

LONDON

Printed and published for the Royal Society

By the Cambridge University Press

Bentley House, N.W. 1

# ON LARGE-SCALE SAMPLE SURVEYS

By P. C. MAHALANOBIS, *Statistical Laboratory, Calcutta*

(*Communicated by R. A. Fisher, F.R.S.—Received* 31 *March* 1943)

CONTENTS



In sample surveys the final estimate is prepared from information collected for sample units of definite size (area) located at random. Large-scale work involves journeys from one sample unit to another so that both cost and precision of the result depend on size (area) as well as the number (density per sq. mile) of sample units. The object of planning is to settle these two quantities in such a way that (*a*) the precision is a maximum for any assigned cost, or (*b*) the cost is a minimum for any assigned precision. The present paper discusses the solution for (1) uni-stage sampling (with randomization in one single stage) both in the abstract and in the concrete; and for (2) multi-stage sample (with randomization in more than one stage) mostly in the abstract.

The whole area is considered here as a statistical field consisting of a large number of basic cells each having a definite value of the variate under study. These values (with suitable grouping) form an abstract frequency distribution corresponding to which there exists a set of associated space distributions (of which the observed field is but one) generated by allocating the variate values to different cells in different ways. This raises novel problems which are space generalizations of the classical theory of sampling distribution and estimation. On the applied side it also enables classification of the technique into two types: (*a*) 'individual' or (*b*) 'grid' sampling depending on whether each sample unit consists of only one or more than one basic cell. For most space distribu-

tion precision of the result is nearly equal for both types of sampling; these are called fields of random type. For certain fields (including those usually observed in nature) precision depends on sampling type; these are fields of non-random type.

Application to estimating acreage under jute covering 60,000 sq. miles in Bengal in 1941–2 is described with numerical data. The margin of error of the sample estimate was about 2 %, while cost was only a fifteenth of that of a complete census made in the same year by an official agency.

## PART I. INTRODUCTORY

### 1. Introduction

1. Since 1937 I have had the opportunity of studying in the Calcutta Statistical Laboratory the problem of estimating with the help of sample surveys the area and yield of a number of crops like paddy and jute in Bengal and wheat and sugar cane in the United Provinces. The work has to be done on a large scale covering tracts of land fifty or sixty thousand square miles or more in extent. The question of costs is therefore of great importance, and my aim has been to develop a sampling technique which would supply, at any given cost, a final estimate with the lowest possible margin of error.

2. It is the object of the present paper to give a general account of the work relating to the estimation of crop areas. The basic principles are not new. These were enunciated concisely but with characteristic precision by Professor R. A. Fisher who stated that the object of such sample surveys was 'to give the maximum precision in return for the labour expended' (*J. R. Statist. Soc.* 1934, p. 615)—a point of view identical with that adopted in the present paper. In fact, in one sense what has been done is to develop Fisher's ideas in a systematic way in the light of experimental studies of large-scale field surveys.

3. As already mentioned a characteristic feature of the work is its large scale of operations. This introduces many special problems on the theoretical as well as on the applied side. Four or five hundred investigators are often employed in the field survey; and they have to work, not in a compact group, but scattered over the whole country covering fifty or sixty thousand square miles in area. Preparatory and tabulation work has to be organized on an extensive scale; in the jute survey, for example, it involved handling over 200,000 sheets of village maps. The whole task thus partakes of something of the nature of an engineering project; and this is why the present paper may be described as dealing with a problem of statistical engineering.

4. The paper thus naturally falls into two distinct portions—one concerned with the abstract theory and the other with the application to concrete problems. The first part is purely introductory, and gives a general description of the nature of the problem in non-technical language and a brief history of the jute-survey scheme which was the starting point of the investigations.

5. Part II* is concerned almost exclusively with the basic concepts and theoretical principles in an abstract form supported, however, by the results of model sampling experiments. This part is complete by itself and will be of special interest to those who desire to get acquainted with the abstract theory without entering into the details of experimental procedure. This is followed by a discussion in Part III* of the application of the abstract theory

* The paragraphs have been numbered continuously except in the Appendices to Part II, and equations have been numbered according to paragraph numbers.

to the estimation of area under crops with special reference to the work on jute in Bengal. A summary is given of relevant experimental results with numerical examples. This account of the jute survey is being given mainly for purposes of illustration. The emphasis throughout has been first on elucidating the theoretical principles with the help of concrete examples; and secondly, on the special features of work on a large scale. The successful organization of sample surveys covering fifty or sixty thousand square miles is a definite advance in the application of statistical methods to practical problems.

6. I am sorry that for unavoidable reasons I am obliged to present this paper in a rather crude form. I have not been able to discuss or refer to work done by other investigators on sample surveys or allied topics. Owing to the deterioration in the war situation in Bengal practically the whole of the Statistical Library was removed in April 1942 to Giridih, a place in the interior at a distance of over 200 miles from Calcutta, and the books are lying there either in stacks or packed in boxes. At the time of writing this paper I did not, therefore, have access to our library.

7. There are also many gaps in the paper on which work is in progress. I thought it advisable not to wait for the completion of these investigations but to put down in writing, even if in a rough and unfinished form, an account of some of the work done in Calcutta in recent years. The most compelling reason is, of course, the growing menace of war in east India which made me anxious to finish this paper without further delay. A second reason has been the publication in the March 1942 issue of the *Journal of the American Statistical Association* (which reached India in August) of an article on recent developments in sampling for agricultural statistics, in which an account has been given of methods in certain ways similar to those described in this paper, but no mention has been made of the work done in India. This made me think it desirable to have an account of the present work published at the earliest opportunity.

8. Practically the whole of Part II is entirely new, and has not been published elsewhere. The account of the jute survey given in Part III has been based on the primary data described in a series of reports prepared by me (and printed by the Indian Central Jute Committee) in the course of a five-year scheme and in other papers mentioned in the following list. The treatment adopted in the present paper is, however, entirely new, being based on a joint study of the material for different years, while the earlier reports dealt with each year separately. All the tables (with the single exception of table 25, which is taken from the Indian Science Congress address) were specially prepared for the present paper and contain much new information.

*List of Publications*

(1) *A statistical report on the experimental crop census of* 1937 (I.C.J.C., September 1938).
(2) A note on grid sampling, *Science and Culture*, 4 (5), 300, November 1938.
(3) *First report on the crop census of* 1938 (I.C.J.C., February 1939).
(4) *Second report on the crop census of* 1938 (I.C.J.C., July 1939).
(5) *Report on the sample census of jute in* 1939 (I.C.J.C., December 1939).
(6) *Statistical report on crop-cutting experiments on jute*, 1939 (I.C.J.C., 1940).
(7) A sample survey of the acreage under jute in Bengal (*Sankhyā*, 4, 511–530, March 1940)

(8) *Sample census of the area under jute in Bengal in* 1940 (I.C.J.C., 1941).
(9) *Crop-estimating experiments on jute in Bengal,* 1940 (I.C.J.C., 1941).
(10) *Sample census of the area under jute in Bengal,* 1941 (I.C.J.C., in the Press).
(11) *Sample surveys* (Presidential Address, Section of Mathematics and Statistics, Indian Science Congress, Baroda, January 1942).

## 2. Brief history of the jute-survey scheme

9. Statistics relating to crops in India have been known to be unreliable for a long time. In 1935 Sir Girja Shankar Bajpai (of the Department of Lands of the Government of India) directed my attention to this question, and a little later I prepared a tentative scheme of sample survey. Sir John Russell saw this scheme during his visit to Calcutta in January 1937 and referred to it in his *Report on the Imperial Council of Agricultural Research* (Govt. of India, 1937). A statutory body called the Indian Central Jute Committee was set up a little later; and this Committee at its very first meeting in February 1937 sanctioned a grant of five lakhs of rupees (£37,500) for a five-year scheme for the improvement of the estimate of area under jute in Bengal. The Committe at first had the idea of carrying out a detailed and complete census of each plot under jute roughly in instalments of one-fifth of the whole area each year. My opinion having been invited in the matter I opposed this proposal, and suggested that a small-scale pilot survey should be undertaken to explore the possibilities of the sampling method. The Committee accepted this proposal and provided necessary funds for this purpose.

10. This was the beginning of the jute-survey scheme. In the first exploratory survey of September and October 1937 a good deal of field material based on both complete enumeration and sample survey was collected. The Jute Comittee, was, however, not convinced about the practical usefulness of sample surveys, and was doubtful whether the scheme should be proceeded with or not. Fortunately, Professor R. A. Fisher, who came to India at this time, examined the scheme in January 1938, and recommended it in written notes as well as in personal discussions with government officials. His powerful support turned the scales in favour of the sample survey; and funds were provided for a series of gradually expanding exploratory surveys in 1938, 1939 and 1940, culminating in a full-scale survey covering about 60,000 sq. miles in Bengal in 1941. The total expenditure was Rs. 4,14,000 (a little over £31,000) in five years.

11. A special Jute Census Committee had been set up for the scheme with representatives of Government, growers, and manufacturers; I also was a member and acted as its Statistical Adviser. This Committee was in administrative charge of the scheme, but the whole of the statistical and field work was done under my technical control and guidance. I was able to undertake this heavy responsibility only because of the willing co-operation of my fellow-workers in the Calcutta Statistical Laboratory.

12. To judge the success (or otherwise) of the scheme the Jute Census Committee had laid down three tests. The reliability of the sample survey must be such that the margin of error of the final estimate of the area under jute should not exceed 5%; secondly, the results must be available sufficiently early in the jute season and preferably by the first or second

week of September; and finally, the cost of the sample survey should not be excessive. As will be seen later, the sample estimate in 1941 agreed within 2·8 % of an entirely independent official estimate based on a complete detailed census of each individual plot sown with jute which was carried out in that year by Government for the purpose of a compulsory scheme of jute regulation. The estimate based on the sample survey was ready within one week of the cessation of field work, and was submitted to Government on 27 August 1941. The cost of a sample survey, was estimated at Rs. 1,14,000 (or about £8500) per year against an expenditure of about fifteen lakhs of rupees (about £110,000) for a complete census. The Jute Census Committee therefore considered the sampling technique to have fully satisfied all three tests, and recommended to Government the adoption of the sample survey in future.

## 3. The nature of the problem: survey of crop areas

13. Sampling technique assumes a particularly simple form in the familiar urn problem. Balls are drawn from the urn, and by counting the number of balls of each colour in the sample the relative proportion of balls of different colours in the sample is ascertained, and hence the composition of the balls in the urn inferred. The margin of error of the estimate is also calculated with the help of the familiar binomial or multinomial distribution. In the crop census an analogous method would be to draw at random a suitable number of plots of land and ascertain which of these are under the particular crop under survey. In order that statistical principles may be used in a valid manner it is essential that the sample plots should be picked up strictly at random. The task would be quite easy if the plots were of equal size. If the plots were serially numbered it would then be sufficient to select the sample plots with the help of a series of random numbers. But unfortunately this method cannot be adopted, as the size of individual plots in Bengal varies widely from a tenth or twentieth of an acre to several hundreds of acres. Selection by serial number of plots would not therefore give each unit area of land the same chance of being included in the sample; and samples drawn in this way would not be truly random.

14. One way of getting over the difficulty of unequal size of plots would be to form progressive totals of the area of the plots in serial order, and then use random numbers on the basis of such progressive totals themselves instead of the serial numbers. This would supply a theoretically valid sample. The total number of revenue plots which would have to be taken into consideration in Bengal for this purpose would be, however, something like a hundred millions; and the task of compiling the area of individual plots and of forming the progressive totals would be not only expensive but difficult to carry through with accuracy in practice. But this is not all. In an appreciable number of cases (30 % for jute) the whole of the plot is not under the same crop. If such a plot is included in the sample it would be difficult to allot the plot to a particular crop in an unambiguous manner.

15. In this situation it becomes necessary to use not points but sample units of a finite size like 4-acre or 20-acre or 40-acre. In my two earliest publications in 1938 I had referred to these sample units as 'grids', and this term has been adhered to in the present paper. Other considerations also make the same procedure inescapable. In the case of large-scale surveys covering areas of the order of fifty or sixty thousand square miles only a limited number of sample units can be used within the available resources of labour or

money, and the sample units would be on the whole widely scattered. The time required for moving from one sample unit or grid to another (which is called 'journey time' for convenience of reference) would not be negligible in comparison with the time required for locating the sample units and estimating the proportion of jute in it. It would be uneconomical to examine or enumerate a single plot at each locality, and it would be obviously possible to examine or enumerate a group of plots or sample units of fairly large size at each locality visited by the investigators, as the additional time required for this purpose would be usually small. Considerations of economy thus suggest the use not of single plots but of sample units of fairly good size.

16. In this method the whole area would be divided into a suitable number of zones, each of which would be as homogeneous as possible; and within each zone a suitable number of sample units or grids of a suitable size would be located at random and examined for the proportion under jute or the crop under survey. If a sufficiently large number of grids are used it would be possible in this way to determine the average proportion of land under jute in each zone, and hence, multiplying by the area of the zone, the area under jute in each zone. Adding the figures for the different zones it would then be possible to obtain the total acreage under jute for the country as a whole.

17. Consider now what should be the size or area of each sample unit or grid. Usually the total cost (or the total number of field investigators) at the disposal of the party would be fixed. If work is done with grids of a large size there can be only a few of them, so that they will be widely scattered and the density (or number of grids per square mile) would be small. On the other hand, if the grids are of small size there can be more of them, so that they would lie fairly close together and the density would be higher.

18. The important point to be noted is that the need of keeping the total cost the same obviously places a restriction on the choice of the size (that is, the area) and the density of the sample units. Once the size (or area) of individual sample units is fixed, the total number allowable also becomes fixed. On the other hand, if the total number of sample units is fixed, then the size or area of each individual sample unit in its turn becomes determined. It is not possible to choose independently both the area of individual sample units and their number. Questions of cost thus supply one connexion between area and number of sample units. This is why a study of the 'cost function' or how the cost depends on the size (or area) and density (or number per square mile) of sample units is a matter of great importance in the present connexion.

19. But cost is not the only factor. Consideration must also be given to the precision or the margin of error of the final estimate. The variance or the margin of error for individual sample units would be large when the size or area of the sample unit is small. The variance would decrease, that is, the precision of individual sample units would increase as the area of each individual sample unit is increased. The variance of the mean value based on the whole group of sample units increases. It has already been seen that if work is done with sample units of small size, a large number of such units can be employed; the small size of sample units would mean a comparatively large variance for individual units, but their large number would tend to reduce the variance of the mean value. On the other hand, if work is done with sample units of large size (i.e. area) then only a few of them can be afforded.

Here, although the variance for individual sample units would be comparatively small, their small number would tend to increase the variance of the mean value.

20. In this situation whether there is any net gain or not in the precision of the mean value will just depend on whether the increase due to the reduction in the number of sample units is or is not compensated for by the decrease in the variance of individual sample units as their area is increased. This question can only be settled by studying experimentally how the variance changes as the area of individual sample units is increased; that is, by studying the variance as a function of the size (i.e. area) of individual sample units.

21. From the point of view of sampling technique the following questions thus arise:

(1) What should be the size of each sample unit or grid in each zone?

(2) How many such sample units should there be altogether, and how should these be distributed among different zones?

These questions must of course be settled in accordance with (*a*) the degree of precision required in the final estimate, and (*b*) the amount of labour or money which can be used for the sample survey. If the total amount of money or labour is fixed, the next object will be to settle the size and density of grids in order to obtain the final result with the minimum possible margin of error. On the other hand, if the precision of the final estimate is fixed, then the size and distribution of grids must be settled in such a way that the work may be done at a minimum cost. This is the typical problem in the estimation of the area under crops.

### *Production (or yield) surveys*

22. A second form of the sample survey on a large scale occurs in crop-cutting experiments. For purposes of crop forecasts the ultimate object is usually to obtain an estimate of the total production of a particular crop over the whole country. The whole area is, however, usually divided into a convenient number of zones or administrative divisions, and separate estimates for each zone or region are sometimes required. It is also sometimes necessary to have more detailed knowledge of the yield per acre over comparatively small pieces of land. This, however, can be conveniently considered to be a third type of problem which I shall describe a little later.

23. As compared to the area census, crop-estimating work is necessarily more complicated and more expensive. In the area census the only thing necessary is to measure or estimate the proportion of land under a particular crop or under a number of different crops for each sampling unit or grid. In crop estimating surveys the actual crop has to be harvested, and usually has also to be subjected to some kind of processing before being measured.

24. Now consider a particular zone or area under survey. A procedure analogous to that described in the case of the area census would be to locate a suitable number of sampling units or grids purely at random over the whole zone or area. As the cost of harvesting the crop for each sampling unit is usually high, it is clear that the total number of sample units or grids would have to be much smaller than that in the case of the area census at any given level of total cost. In other words, the sampling units must be necessarily more widely scattered in crop-cutting work as compared to the area census at any given level of expenditure,

which means that the time and expense required for travelling from one sampling unit to another would be necessarily greater in the case of crop-cutting work. The cost is usually prohibitive.

25. A different type of sampling procedure is, therefore, often adopted in practice. Instead of scattering all the sampling units purely at random these are grouped together in hierarchical order. For example, the whole area may be divided into a fairly large number of compact areas each of which may be called a 'block', and a suitable number of blocks out of the total number may be selected at random. Within each selected block a suitable number of villages is selected at random; within each village a suitable number of fields is selected at random and finally within each selected field a suitable number of sample units, again purely at random, is located.

26. In the area census the sampling units are located by one single stage of randomization, and hence this method may be called single or uni-stage sampling. In the case of crop-estimating surveys, on the other hand, the sample units are located by successive stages of randomization, and this may be called multi-stage sampling. In this method each and every portion of the area under survey is subjected to at least one stage of randomization, but some of the regions are subjected to more than one process of randomization.

*Mapping surveys*

27. I may now briefly refer to a third form of the sample survey. The object here is to obtain detailed information regarding the yield of a crop or some soil characteristic or some other variate for comparatively small pieces of land. In the case of yield of crops this may even be such a small piece as a revenue plot, the average size of which is somewhat less than half an acre in Bengal. For example, the object sometimes is to make an equitable assessment of the land revenue based on an objective estimate of the fertility of the land; the average yield over even a single village is not of any use for this purpose, and detailed information has to be collected for different classes of land within the village. Such problems fall under mapping surveys in which the ultimate object is to prepare a map of the whole area showing the fertility level of the land estimated in terms of the yield of selected crops.

28. To sum up, there are three broad types of problems which may be called respectively (1) area surveys, (2) yield (or production) surveys, and (3) mapping surveys. In each case work has to be done on a large scale; and in each case special types of sampling technique have to be developed in order that the work may be done with efficiency and economy. In the present paper I have explained certain basic concepts and theoretical formulations which will be found useful in the case of all three problems, but otherwise I have confined my attention mainly to uni-stage sampling which is specially appropriate for area surveys. Work is in progress on the other two problems, and I hope to be able to deal with these questions in subsequent papers.

*Exploratory work*

29. It may be useful at this stage to point out that the approach adopted in the present paper is especially suited to surveys (like crop forecasts) which are carried out every year or at fairly short intervals. It is then possible to secure information relating to the 'cost function' and the 'variance function' either by a preliminary series of exploratory surveys

or by special experiments carried out along with regular annual or periodic surveys. The planning of sample surveys thus consists of two stages, namely, (*a*) the exploratory and (*b*) the final stage. In the exploratory stage the chief object is to study the cost and variance functions, and the only way to do this is to carry out suitably designed experiments on the field. The most economical method would obviously be to proceed by gradual stages. This is what was done in the jute-survey scheme. In 1937 a pilot survey covering an area of only 124 sq. miles was made. This supplied a rough idea of the physical order of the different elements involved, and enabled a second pilot survey to be organized on a much larger scale in 1938. Thus the plan was adopted of a series of exploratory surveys on a gradually expanding scale culminating in a full-scale country-wide survey covering roughly 60,000 sq. miles in 1942.

30. *The variance function.* Information was collected for sample units or grids of widely varying sizes from a fraction of an acre to 30 or 40 acres or even several hundreds of acres in some cases, and the variance of individual grids of different sizes directly calculated. In the case of statistical variates conforming to the normal law one would expect the variance to decrease inversely as the size (in this case, the area) of the sampling units or grids. It was found, however, that the actual decrease was much smaller, so that the gain in precision by increasing the size of individual grids was appreciably less than one would ordinarily expect on the normal theory. This may be ascribed to the fact that the proportion of land sown with a particular crop (or the yield of a crop) in plots in the same neighbourhood are not statistically independent but are correlated. The absence or presence of such correlation determines whether the variance would follow the normal law or would decrease more slowly with increasing size of grids. In fact, this furnishes a convenient basis for the classification of fields into two distinct types, namely, (*a*) random type (in which the variance function is normal) and (*b*) non-random type (in which the variance decreases more slowly). The theoretical formulation is given in Part II; the point to be emphasized is that this is based not on speculative grounds but on experimental evidence and on the fact that such classifications lead to economy in the planning of sample surveys.

31. *Cost of operations.* From the very beginning both field and statistical workers were asked to keep daily records of the time spent on different types of work. This furnished the material for the cost function which was studied in the first instance in labour units, that is, in terms of man-hours or hours of work per investigator or computer which were later converted into money values. In doing this the total cost of course was taken into consideration; for example, the cost of one investigator-hour included not only the pay of the field investigator but also the pay of the inspecting and supervising staff, travelling and all contingent expenses. After much experimentation (some account of which is given in § 4 of Part III), it was found that the cost of field operations could be split up into three chief components. First of all there was the time required for going from one sampling unit to another which was called 'journey time'. This involved movements from one camp to another (camp being defined as the place where the field investigator spent the night), from camp to field and back, and from grid to grid. Next came the time required for locating the grids, enumerating the crops, and making necessary field entries; this was called 'enumeration time'. Besides this a certain amount of time was required for miscellaneous work.

32. It was found that in a large-scale survey the journey time was much greater than enumeration time; in the jute survey of 1941 it was in fact three times larger. It was also found that the journey time depended upon the number of grids per square mile (i.e. on their distance apart) but was independent of the size (or area) of the individual grids. Enumeration time, on the other hand, naturally increased with the size of grids, but was independent of their density. Time required for miscellaneous work was more or less constant and independent of both size and density of grids. A similar analysis was also made of the time required for the statistical portion of the work. Further details are given in Part III. Numerical values of the parameters naturally fluctuated to some extent from year to year, but the results were reasonably steady, and in general agreement with what one would expect from broad considerations. The experimental work thus fully confirmed the need as well as the possibility of studying the cost function on empirical lines.

33. *Optimum size and density of grids.* Having determined the variance and cost functions with sufficient accuracy for practical purposes the next thing necessary was to settle the best size and density of sample units or grids. In the jute survey the total expenditure which could be incurred was fixed by the amount of the grant sanctioned for this purpose. In this situation the size and density of grids had to be determined in such a way that the margin of error of the final estimate would be a minimum. (Alternatively, the permissible uncertainty in the final estimate being assigned, it is possible to work out a solution for doing the work at minimum cost.) Such solutions (which may be called optimum solutions) are discussed in an abstract form in Part II, and numerical examples are given in Part III.

34. As already indicated, the theoretical approach adopted in the present paper is based on the joint use of the variance and the cost functions. This has been justified by the fact that variances were found in practice to decrease more slowly than one would expect in the normal case, and also by the possibility of determining the cost function with sufficient accuracy for all practical purposes by empirical methods. The present approach has been also fully justified on grounds of economy. The cost of a survey on the same scale as the jute survey of 1941 could have been easily increased two or three times by a plausible but wrong choice of the size of grids. For annual or periodic surveys (which give scope for what I have called exploratory work) the present technique would thus appear to be the most suitable one. On the other hand for a survey which would be carried out only once or only at very long intervals and where no previous information is available a simple random sampling procedure would probably be found most convenient.

35. *The human factor.* So far attention has been confined to sampling fluctuations (which arise from the information being based on samples or limited portions of the whole population) which are amenable to statistical treatment. The exploratory work done in 1937 and 1938, however, showed clearly that, apart from such random fluctuations inherent in the sampling method, crude mistakes in locating or identifying the plots or in estimating the proportion of land under jute within each sample unit were by no means negligible. This was partly due to lack of experience on the part of the field workers, and an appreciable improvement was found after they had been given suitable training. In many cases

the inaccuracies were, however, due to false entries or gross negligence, and systematic inspection of the field work was found to be essential to maintain a minimum standard of accuracy.

36. *Subsamples.* This, however, was not enough. The method of arranging the field survey in the form of two separate but interpenetrating subsamples in each zone was therefore adopted. Linked pairs of sample units were located at random on maps; and one sample unit was allotted to half-sample ($A$) and the other to half-sample ($B$); the distance between each pair of grids was kept constant, but the orientation was settled at random. The information for all sample units belonging to half-sample ($A$) was collected by one set of investigators, while the information for the sample units belonging to half-sample ($B$) was collected independently by an entirely different set of investigators. The time programme was arranged in such a way that investigators belonging to the two different sets ($A$) and ($B$) never worked in the same region at the same time. In this way independent records for each of a pair of adjoining sample units were obtained and two separate estimates of the area sown with jute. The degree of agreement between these two estimates supplied a good idea of the precision of the survey.

37. There were many difficulties on the organizational side. There was not a single trained worker to start with. Apart from considerations of economy, one important reason why it was essential to adopt the gradually increasing scale of work was the need of building up the necessary human agency for carrying out surveys on a country-wide scale. On the statistical side also it was necessary to organize the computational work, standardize routine methods of preparing the sample units, and arrange for continuous and systematic tabulation of the field data and their final analysis.

38. In fact, apart from the study of the cost and variance functions, one of the great advantages of the exploratory method was the opportunity it gave for developing suitable methods of controlling mistakes arising from the human factor, for giving training to the workers and to build up the necessary human agency for both field and statistical work. Once this was done, and the cost and variance functions were determined, the final stage was reached when the planning of the sample survey could be undertaken on scientific lines. In one sense there is, of course, no final stage. Conditions are changing from year to year; and it would be obviously desirable to continue auxiliary work of an exploratory type (along with the main survey) with the help of which the efficiency of the survey can be continuously improved.

### *Concluding remarks*

39. As already noted, the sample survey of the kind described here may be called a project in statistical engineering. The whole scheme was essentially a co-operative undertaking; and the real credit for the successful organization of the jute survey therefore belongs to the large group of both statistical and field workers who were associated with me in this project. It would be invidious to mention particular names, and I am reserving this pleasant task for a more suitable occasion, namely, a full report of the whole undertaking. I cannot, however, conclude this paper without recording my grateful appreciation of the help I received from Mr A. P. Cliff, the first Secretary of the Indian Central Jute Committee, without whose drive and initiative the jute-survey scheme would never have come into

operation; the late Subhendusekhar Bose, who was associated with the work until his untimely death in November 1938; Mr N. C. Chakravarti, B.C.S., who first set up the field organization; and Messrs Samarendranath Roy, Sudhir Kumar Banerjee, Jitendramohan Sengupta and Purnendusekhar Bose, without whose untiring help it would not have been possible to prepare this paper in the course of seven or eight weeks.

## PART II. OUTLINE OF THE THEORY FOR UNI-STAGE SAMPLING

40. I shall consider the general principles in abstract form in this part, and discuss in outline the relevant theory in the case of uni-stage sampling which is especially appropriate for the estimation of the area under crops. Illustrative numerical examples are given in Part III of the paper.

### 1. Basic concepts

41. In this section I shall explain the basic concepts of fields or space distributions of a statistical variate and associated frequency distributions.

*Statistical variates and fields.* The fields considered in this paper are essentially geographical regions of finite areas. Use may be made, therefore, of rectangular co-ordinates in the usual way to specify any given point or location in the field. At (or rather in the neighbourhood of) each point $(x', y')$ is found a finite value of the variate under consideration, say, $z = z(x', y')$. It must be remembered, however, that the $z$-variate is essentially a statistical quantity which can be defined only as a kind of average value or density over a certain finite area in the neighbourhood of each point.

42. *Basic cells.* It is thus found that the fields considered here have an essentially discrete structure. The concept of a basic cell as the smallest area (measured in acres or square miles or any other suitable unit) for which the $z$-variate may be considered to have a sufficiently precise meaning is therefore introduced. It is not suggested that the size of this basic cell is an absolutely determinate and atomic quantity. No doubt there is a certain amount of arbitrariness in selecting a particular value as our smallest unit, but such arbitrariness is inescapable, and fortunately does not affect the general argument. It is also convenient to think of these basic cells as having a square shape. This again is arbitrary, but the general argument will not be affected by making this assumption. The field can thus be visualized as being made up of a definite number of ultimate or basic cells of square shape and finite area. The symbol □ (to be called 'quad'*) can be used to represent the area of a basic cell measured in some suitable unit like acre, sq. mile, etc. If $A$ is the total area of the field under consideration, then $A/\square = N_0$ will give the total number of ultimate basic cells. This $N_0$ will be generally large in the case of fields considered in the present paper.

43. *Co-ordinate numbers* $(i', j')$. Having introduced the discrete structure, strictly speaking, it is not permissible to use continuously variable co-ordinates $x'$ and $y'$. In this situation a series of co-ordinate numbers $i' = 1, 2, 3, \ldots, l$ and $j' = 1, 2, 3, \ldots, m$ may be used; so that the location of any particular ultimate cell may be specified by a pair of values for $(i', j')$. In terms of the quad or □ which serves as the scale unit for length the total area $A = lm$.

* As suggested by Professor F. W. Levi of Calcutta University, who informed me that Hilbert had used this name for the symbol □ in a course of lectures.

Therefore, when it is necessary to emphasize the discrete structure of the field, the notation $z(i', j')$ for the variate will be used.

44. *Multi-variate fields.* This notation can also be extended easily to cover more than one entity; for example, say the respective area under different crops which may be represented by different variates $z_1(i', j')$, $z_2(i', j')$, ..., $z_p(i', j')$, etc. A 'field' in the sense of the present paper thus consists of a finite number, say $N_0$, basic cells arranged in a definite space or geographical order together with a single value (or a set of values in the multi-variate case) of $z$ for each basic cell.

*Abstract and space distributions*

45. *Abstract set and abstract frequency distribution of z.* Now consider a particular field consisting of $N_0$ cells in which the value of $z$ is uniquely determined for each cell, that is, for each pair of values of $(i', j')$ or in the neighbourhood of each point specified by $(x', y')$. The set of $N_0$ values of $z$ then constitute the abstract set of $z$. Using any suitable set of ranges or class intervals, which may be equal or unequal in length, a histogram or finite frequency distribution of the $N_0$ values of $z$ can also easily be constructed. This will be called the abstract distribution of $z$, and will, in general, depend upon the detailed specifications, namely, the number and lengths of the class intervals which may be denoted by $(z'_0 - z'_1)$, $(z'_1 - z'_2)$, ..., $(z'_{c-1} - z'_c)$. The above set of values of $(z'_0 - z'_1)$, etc., which determine the framework of the classification may be concisely represented by $I(c)$.

46. *Space distribution of z.* In the case of an actual field, the $N_0$ values of $z$, however, have a definite space distribution, and the statistical properties of this space distribution are of great importance in the present problem. A clear distinction must be made, therefore, between what I have called the 'abstract distribution of $z$', and the corresponding 'space distribution'. Consider a field consisting of $N_0$ basic cells, and $N_0$ (or $N_0$ sets of) values of $z$. Corresponding to any particular abstract distribution there are $(N_0)!$ different ways of arranging the $N_0$ different values of $z$ in $N_0$ cells. Each of these $(N_0)!$ distributions may be considered to be a micro-distribution or a micro-state in space corresponding or belonging to the given abstract distribution of $z$. (I am using the phrase 'micro-state' in the sense in which it is used in statistical mechanics.) It will be assumed that these $(N_0)!$ micro-distributions or fields may be considered to be 'equally likely' in the sense in which this phrase is generally used in the theory of probability. The field which is actually observed will be only one particular micro-distribution or micro-state, and as such is as likely to occur as any other micro-distribution by pure chance.

47. It may happen that some of the values of $z$ are identical. When the abstract set of $z$ is classified into a finite number of class ranges, and all values of $z$ falling within the same class range by the same symbol are labelled, replicated values of $z$ will then necessarily occur. If $N_1, N_2, \ldots, N_c$, etc., are the frequencies in the different classes, then the total number of micro-states which can be physically distinguished will be given by

$$(N_0)!/(N_1)!\,(N_2)!\ldots(N_c)!,$$

where the sum of $N_1 + N_2 + \ldots + N_c = N_0$. If, however, the values of $z$ falling within the same class range are considered to be distinguishable in a statistical sense, then the total number of micro-states will be $(N_0)!$. Usually consideration will be given to expectation values or

statistical or probabilistic properties of the distributions, and for such purposes, so far as the abstract argument is concerned, it will be in most cases immaterial whether the values of $z$ within the same class range are considered to be distinguishable or not. In either case the appropriate weights would automatically enter into the calculations and leave mean values and probabilistic results formally unchanged. The total number of exhaustive (and mutually exclusive) micro-states will be written as $N$, which will be either equal to $(N_0)!$ or some other appropriate number determined by the multinomial distribution which will be of the order of $(N_0)!$ and large in comparison with $N_0$.

*Different types of sampling procedure*

Before proceeding further it will be convenient to consider different types of sampling procedure.

48. *Unitary and zonal sampling.* To fix ideas, let it be supposed that there is a field consisting of $N_0$ basic cells, and out of these it is desired to select $n$ cells as a sample. Broadly speaking there are two different procedures which can be adopted. The $n$ sample cells may be drawn out of the whole lot of $N_0$ cells without dividing the field into smaller subdivisions. This is the first type, which may be called unitary sampling over the whole field. This, of course, is a familiar procedure often adopted in statistical practice.

49. There is a second broad type, which may be called zonal sampling. In this method the whole field is divided into a suitable number of, say, $k$ compartments, strata or zones, say $A_k$, where $k = 1, 2, \ldots, l$, and then a certain number of samples is allotted to each compartment. $n_k$ is written as the number of samples allotted to the $k$th compartment or zone; summing $n_k$ for all values of $k$ will naturally give the total number of sample cells $n$ for the whole field. A familiar example of this type of sampling is the one called 'stratified sampling' by Professor J. Neyman.

50. *Unrestricted and configurational sampling* (this so far as the field itself is concerned). As regards the actual procedure of choosing the individual basic cells there are also two broad types. The $n$ basic cells over the whole field in unitary sampling may be chosen (or $n_k$ cells in the $k$th zone in the zonal sampling) individually at random, each individual basic cell being given the same chance of being included in the sample. This may be called the *unrestricted* type of random sampling, and its two subclasses as (1) unrestricted unitary, and (2) unrestricted zonal, depending on whether the field is treated as an undivided whole or is subdivided into compartments. In unrestricted random sampling no restriction whatever is imposed on the individual basic cells forming the sample of $n$ basic cells in the case of the whole field (or $n_k$ basic cells in the $k$th zone).

51. *Configurational or 'grid' sampling.* In zonal sampling it is seen that certain space restrictions are imposed on the field itself, but none whatever on the sample cells. However, geometrical restrictions can be imposed, not on the field, but on the sample itself. For example, the sample in compact blocks or groups of basic cells may be collected. Thus compact blocks of two, three or more cells occurring in a column may be used. Or square-shaped blocks of $4, 9, 16, \ldots, m^2$ adjoining basic cells may be collected; or blocks of a rectangular shape consisting of $m \times n$ cells, etc. Instead of taking compact blocks of adjoining cells,

groups of cells arranged in any particular geometrical configuration may also be collected, for example, four cells at the four corner points of a square of a particular size, etc.*

52. *Grids.* It will be convenient to give a name to such sample units consisting of groups of cells. In the *Statistical report on the crop census of* 1937, I had used the word 'grid' for this purpose, and it may be retained here. From this point of view, therefore, a single grid may be defined as a sample unit consisting of a number of basic cells arranged in a standard geometrical pattern or configuration. In this sense one may speak of grids of a definite size or shape (like square or rectangular grids) or having a definite pattern.

53. The essential point to be noted is that each such 'grid' functions as a complete integral unit for purposes of sampling, and has to be located as a whole in a purely random manner over the field. In order to do this, appropriate rules of procedure must be adopted. For example, in the case of a square block of cells, i.e. in the case of a grid of a square shape, one may locate the lower left-hand corner point at random, and then proceed to build up the whole block by taking the required number of cells along adjoining columns and rows. It will be noticed that in configurational sampling restrictions are imposed on the geometrical arrangement of the individual basic cells forming each sample unit or 'grid', but each sample unit or 'grid' as a whole is located at random. Within specified restrictions the principle of randomization is thus preserved intact so that statistical methods may be used on a valid basis.

54. Configurational sampling itself may be of either (*a*) unitary or (*b*) zonal type. Thus there are four different types: (1) unitary unrestricted, (2) zonal unrestricted, (3) unitary configurational, and (4) zonal configurational.

55. *Grid notation.* As will be seen later it is often necessary to vary the size (that is, the area) of the grids from zone to zone, but it is usually possible to keep the size of the grid constant within each zone. Let $\square_k$ be the area of each basic cell in the $k$th zone—usually this will be the same in all zones, in which case $\square$ may be used instead of $\square_k$. Let each grid in the $k$th zone consist of $n_{0k}$ basic cells arranged either in the form of a compact square or rectangle, or otherwise in any geometrical configuration or pattern in which the basic cells are not necessarily contiguous; the first is, of course, a special case of the second. The $n_{0k}$ basic cells arranged in any geometrical configuration will be denoted by $Gr_k(\square_k, n_{0k})$, the shape depending on the nature of the particular configuration chosen. The area occupied by each individual grid in the $k$th zone, which may be written as $a_k$, is then given by $a_k = \square_k n_{0k}$ (or $\square n_{0k}$ when the size of basic cells is the same in all zones). In unitary sampling (in which the field is treated as an integral whole without subdivision into zones) the suffix $k$ may be dropped, and $a = \square n_0$ written as the size of individual grids.

56. Very often it is convenient to use grids which are simply compact blocks of adjoining cells arranged in a rectangular or square shape, each consisting of say $i_k$ and $j_k$ basic cells in the two directions in the $k$th zone. In this case $n_{0k} = i_k j_k$ and $a_k = \square_k i_k j_k$ in the $k$th zone,

* It is to be noted that such sampling introduced a 'bias' in the neighbourhood of the boundary of the field, inasmuch as basic cells in the interior of the field appear a certain number of times in the totality of all samples while those near the boundary appear less often. In the present study, the effect of the 'bias' on the field, which will be called the border effect, has been discussed in a separate note, No. 6, in the appendix to this part.

or simply $a = \square ij$ in the case of unitary sampling. The area *covered* by the grid (which is the maximum area which can be bounded by straight lines joining any two constituent cells of the grid) will in general be different from $\square_k n_{0k}$ and will depend upon the shape of the configuration; in the particular case of a compact grid the two would, of course, be equal.

*Uni-stage and multi-stage sampling*

57. It should be noted that all the four types of sampling described above have one common characteristic, inasmuch as the process of randomization is carried out only once in each and every region of the field. This may be called uni-stage sampling. More complicated methods in two or more stages may also be adopted in which the process of randomization is used at least once in each region of the field, but certain regions are subjected to a second, third, or further stages of randomization. For example, in crop-cutting experiments a common procedure is to select at random a certain number of villages; then to select, again at random, a certain number of fields within each of the villages already included in the sample in the first stage; and then to select at random certain portions of each selected field. Here the first stage of randomization refers to the selection of villages, and this process covers the whole of the area under survey. But once the first stage of randomization is completed, the second stage of randomization is restricted only to fields lying within selected villages, so that fields belonging to excluded villages have no further chance of being included in the sample, that is, are not subjected to the second stage of randomization. In the same way the third stage of randomization is restricted to only those fields which have been already included in the sample at the second stage. In multi-stage sampling more and more basic cells or larger and larger areas are excluded at each stage, and the process of randomization becomes more and more restricted in coverage or extent.

58. The type of sampling adopted at different stages of sampling may of course be different. Zonal unrestricted may be adopted in the first stage, unitary unrestricted in the second stage, and finally unitary configurational in the third stage. For example, in crop-cutting work the whole area may be divided into a number of zones and villages selected purely at random (zonal unrestricted) in the first stage; within each village the fields may be selected purely at random (unitary unrestricted) in the second stage; and finally within each field compact blocks of square shape may be selected at random (unitary configurational) in the third stage. There are four fundamental types of sampling, namely, (1) unitary unrestricted, (2) zonal unrestricted, (3) unitary configurational, and (4) zonal configurational, and any one of these may be used at any stage. If there are $s$ stages of sampling then the number of possible combinations of different types of sampling procedure will be $4^s$. This will indicate the wide range of choice in the selection of suitable types of sampling procedure.

*Difference between unrestricted and configurational sampling*

59. Consider now the difference between unrestricted and configurational sampling. It will be sufficient if these two methods are compared in the case of the unitary field. Corresponding to any abstract distribution of $N_0$ values of $z$ it is seen that there are $N$ micro-states generated by the allotment of the $N_0$ values of $z$ to the $N_0$ basic cells of the field in all possible

ways. Serial numbers $1, 2, \ldots, s, \ldots, N$ may be applied to these $N$ micro-states in any manner necessary for purposes of identification; from this the $s$th micro-state is derived.

60. Now consider any particular, say, the $s$th micro-state. Consider two grids (each consisting of $n_0$ basic cells arranged in a definite geometrical pattern) located at random on this $s$th micro-state or field. These two grids may be considered to have an identical position if they consist of the same identical set of $n_0$ basic cells. On the other hand, they will be considered to have different positions when there is at least one basic cell which is not common to both the grids. As the whole field consists of a finite number of cells $N_0$ it is clear that the total number of different positions of any grid of specified pattern will be finite. Call this number $N'$. This number $N'$ will, of course, depend not only on what is known as the abstract distribution of $z$, but also on the size and shape of the field and on the number $n_0$ of basic cells and the pattern in which they are arranged to form each grid. It will, however, be identical for all micro-states for any assigned shape and size of grid. Serial numbers $(1, 2, 3, \ldots, t, \ldots, N')$ may be allotted to the set of $N'$ grids in any form necessary.*

61. Now consider the $t$th grid. There are $n_0$ values of $z$ each of which corresponds to one of the $n_0$ basic cells which constitute the grid. Any statistic (in the Fisherian sense) in terms of these $n_0$ values of $z$ belonging to the same $t$th grid may now be constructed. Any such statistic may be written as, say, $u_{st}\{Gr(\square, n_0)\}$, where $Gr(\square, n_0)$ indicates that $n_0$ basic cells each of size $\square$ are arranged in a certain definite configuration, and $u_{st}\{Gr(\square, n_0)\}$ indicates that the statistic in question has to be constructed from the $n_0$ values of $z$ drawn in the form of the specified grid. For the $s$th micro-state there will be $N'$ such values (given by $t = 1, 2, \ldots, N'$) which forms a frequency distribution of the given statistic for the $s$th micro-state. The fact that all these values belong to the $s$th micro-state is explicitly indicated by the suffix $s$ in the form $u_{st}\{Gr(\square, n_0)\}$, where $t$ goes from 1 to $N'$ for any particular value of $s$, and $s$ goes from 1 to $N$. These $NN'$ values of $u_{st}\{Gr(\square, n_0)\}$ form a complete set of sample values of the statistic under consideration.

62. Consider now unrestricted sampling; and to fix ideas attention must be focused on the $s$th micro-state. A sample consisting of $n_0$ basic cells can now be formed, each of which is located separately at random over the field, and as in the case of configurational sampling, the value of the statistic calculated from the $n_0$ values of $z$ drawn separately at random in this way. Such a value of the statistic will be written as $u_{st}\{R(\square, n_0)\}$ to distinguish it from $u_{st}\{Gr(\square, n_0)\}$. As $\square$, or the size of the basic cell, will be fixed in any given situation it is not necessary to mention it explicitly on each occasion. Therefore, henceforth $u_{st}\{R(n_0)\}$ and $u_{st}\{Gr(n_0)\}$ will be written to denote the values of the statistic in question for unrestricted and configurational sampling respectively. In $u_{st}\{R(n_0)\}$ for unrestricted sampling the value of $t$ can obviously assume ${}_{N_0}C_{n_0}$ values, the number of different ways in which $n_0$ can be selected out of $N_0$ values of $z$. This number will be called $N''$. It is clear that $N'' \geqq N'$, where $N'$ is the corresponding number for configurational sampling, as any group of $n_0$ cells which occurs in $N'$ must occur in $N''$ but not vice versa. Now take up some other, say, the $p$th micro-state. If the procedure of drawing samples of $n_0$ basic cells is repeated, each of which is separately

* This may be called the method of overlapping grids which is analogous to sampling from an urn with replacement. There may also be a system of exclusive grids which may be defined as a system in which no two grids have even one single basic cell in common. This would be analogous to sampling from an urn without replacement, and may be appropriate in special problems not discussed in the present paper.

located at random, obviously the same identical set of $N''$ values of the statistic will be obtained, that is, $u_{pt}\{R(n_0)\} = u_{st}\{R(n_0)\}$, which was obtained in the case of the $s$th micro-state. Here, as already explained, $R(n_0)$ would mean $n_0$ basic cells taken at random, and $u_t\{R(n_0)\}$ would mean a statistic formed from the corresponding $n_0$ values; and obviously the suffix $p$ or $s$ may be dropped. It should be noted that once the $n_0$ values are given—no matter whether they come from a random sample or grid—$u_t\{R(n_0)\}$ or $u_{st}\{Gr(n_0)\}$ are the same functionally and numerically; $R(n_0)$ and $Gr(n_0)$ denote how the $n$ values have come. In fact, in unrestricted sampling all micro-states would yield the same identical distribution of the statistic $u_t\{R(n_0)\}$, where $t$ goes from 1 to $N''$. It is obvious that this is also the same distribution as that which is obtained from the abstract distribution of $z$. Thus the important and practically axiomatic result is reached that in unrestricted sampling the distribution of any sample statistic based on samples of size $n_0$ is identical for all micro-states and also for the abstract distribution. In other words, in unrestricted sampling it is not possible to distinguish between different micro-states or between any micro-state and the corresponding abstract set of $z$. In this situation the theory of sampling distribution for the abstract distribution is sufficient for all purposes.

63. Now consider configurational sampling. Here there are $N$ separate bundles each consisting of $N'$ values; and each such set of $N'$ values belong to a particular micro-state. Each set of $N'$ values of the sample statistic $u_{st}\{Gr(n_0)\}$, where $t$ goes from 1 to $N'$, constitutes a frequency distribution. For the $s$th micro-state write this as $F_s[u_{st}\{Gr(n_0)\}]$, where $s$ of course goes from 1 to $N$. The corresponding distribution in the case of unrestricted sampling is written as $F_0[u_t\{R(n_0)\}]$, where $t$ goes from 1 to $N''$ This abstract distribution may be derived (at least in a formal manner, or with sufficient approximation for purposes of numerical work) when the form of the abstract distribution of $z$ as also of the sample statistic $u_t(n_0)$ are known. Call $F_s[u_{st}\{Gr(n_0)\}]$ the space frequency distribution of $u_{st}\{Gr(n_0)\}$ for the $s$th micro-state and write it more concisely as $F_s[Gr(n_0)]$, and call $F_0[u_t\{R(n_0)\}]$ the corresponding abstract distribution of $u_t\{R(n_0)\}$ and write it as $F_0[R(n_0)]$. It has been seen that

$$F_s[R(n_0)] \equiv F_0[R(n_0)].$$

*Random and non-random fields*

64. Now consider whether $F_s[Gr(n_0)]$ and $F_0[R(n_0)]$ are always identical for all values of $s$, or whether these two distributions can be distinguished in the case of certain micro-states. The problem can be approached in many different ways. One method would be to compare directly any particular space distribution $F_s[Gr(n_0)]$ with $F_0[R(n_0)]$. This would obviously depend on setting up suitable criteria for distinguishing between two frequency distributions on a statistical basis. It is not necessary to attempt a rigorous development of tests of difference between two frequency distributions. For the present purpose all that is necessary is to assume the possibility of judging whether two frequency distributions should be considered to be distinguishable or not. Once this is granted, then compare one by one each of the $N$ space distributions $F_s[Gr(n_0)]$ with $F_0[R(n_0)]$. In this way the frequency distributions $F_s[Gr(n_0)]$ can be divided into two classes, one of which (*a*) consists of all space-frequency distributions of $u_{st}\{Gr(n_0)\}$, $t = 1, 2, \ldots, N'$, which are indistinguishable from the corresponding abstract distribution $F_0[R(n_0)]$ of $u_t\{R(n_0)\}$, $t = 1, 2, \ldots, N''$, while the other class consists of (*b*) all space-frequency distributions which have to be considered different

from the abstract-frequency distribution as judged by the particular test of hypothesis and assigned level of significance which is adopted for this purpose.

65. Corresponding to the space-frequency distributions the actual fields or micro-states will also naturally fall into two groups, namely: (*a*) one consisting of all fields or micro-states for which the corresponding space-frequency distributions of $u_{st}\{Gr(n_0)\}$ are indistinguishable from the abstract distribution; and (*b*) all fields or micro-states for which the corresponding space-frequency distributions have to be considered statistically different from the abstract distribution. All fields falling in class (*a*) are now defined as fields of a random type, and all fields falling in class (*b*) as fields or micro-states of a non-random type. Without entering into the details of the procedure for such classification it is sufficient for the present purpose to point out that, as a matter of empirical fact, fields observed in nature have often been found to be of a non-random type as defined above. Further, it is possible to construct fields or micro-states for which the corresponding space-frequency distributions are clearly different from the corresponding abstract distribution.

*Mean values and variance*

66. Instead of comparing directly the frequency distribution of $u_{st}\{Gr(n_0)\}$ with $u_t\{R(n_0)\}$, where in the first case $t = 1, 2, \ldots, N'$, and in the second case $t = 1, 2, \ldots, N''$, consideration may also be given to their expectation or other moments over the whole range of variation which, for the first statistic, would mean variation over the $s$th micro-state. Again, instead of considering any general statistic $u_{st}\{Gr(n_0)\}$ or $u_t\{R(n_0)\}$, the mean value or variance of $z$ based on the $n_0$ values of $z$ in each grid in the case of configurational or in each random sample in the case of unrestricted sampling may be considered. Fixing ideas, consider now the $t$th grid in the $s$th micro-state and the $t$th random sample. Write respectively $z_{st}\{Gr(n_0)\}$ and $z_t\{R(n_0)\}$ for the mean values of the $n_0$ values of $z$ in the $t$th grid and the $t$th random sample. Without difficulty the mean values of $z_{st}\{Gr(n_0)\}$ may be defined (the mean being taken over all values of $t$). This will be called $z_s\{Gr(n_0)\}$ and can be written in the following form:

$$z_s\{Gr(n_0)\} = \frac{1}{N'}\sum_{t=1}^{N'}[z_{st}\{Gr(n_0)\}]. \tag{66·1}$$

The corresponding variance of $z_{st}$ can be defined in the following way:

$$\sigma_s^2\{Gr(n_0)\} = \frac{1}{N'}\sum_{t=1}^{N'}[z_{st}\{Gr(n_0)\} - z_s\{Gr(n_0)\}]^2. \tag{66·2}$$

Here the sample or grid statistic is $z_{st}\{Gr(n_0)\}$, which is the mean of $n_0$ values constituting the $t$th grid in the $s$th micro-state. For the $s$th micro-state this has a distribution $F_s[z_{st}\{Gr(n_0)\}]$ over $t = 1, 2, \ldots, N'$. Instead of considering and comparing these $F_s$'s ($s = 1, 2, \ldots, N$), consider here the first and corrected second moments of these $F_s$'s which are denoted respectively by $z_s\{Gr(n_0)\}$ and $\sigma_s^2\{Gr(n_0)\}$, with $s = 1, 2, \ldots, N$.

67. Now consider the corresponding abstract distribution of $z$. The mean value of $z$ and the variance of $z$ are of course quite determinate and can be written in the following form:

$$\xi(1) = \frac{1}{N_0}\sum_{i=1}^{N_0}(z_i), \tag{67·1}$$

$$\sigma^2(1) = \frac{1}{N_0}\sum_{i=1}^{N_0}[z_i - \xi(1)]^2. \tag{67·2}$$

68. If $n_0$ values of $z$ at a time are drawn out of $N_0$ values, then it is clear that this can be done in ${}_{N_0}C_{n_0}$ different ways which are called $N''$. The mean value of $z_t\{R(n_0)\}$ and its variance may then be written in the following form:

$$\xi\{R(n_0)\} = \frac{1}{N''}\sum_{t=1}^{t=N''}[z_t\{R(n_0)\}], \tag{68·1}$$

$$\sigma^2\{R(n_0)\} = \frac{1}{N''}\sum_{t=1}^{N''}[z_t\{R(n_0)\}-\xi\{R(n_0)\}]^2. \tag{68·2}$$

69. So far as the mean value of the abstract distribution is concerned the position is clear. The mean value $\xi\{R(n_0)\}$ based on random samples each consisting of $n_0$ values of $z$ is equal to the mean value $\xi(1)$ of individual values of $z$ except for a correcting term for cells in the neighbourhood of the boundary of the field. I have discussed this correction in a separate note attached as Appendix 6 to this Part.

70. The variance $\sigma^2(n_0)$ for samples each consisting of $n_0$ values of $z$ drawn at random is also determinate, and can be easily calculated for any particular abstract distribution of $z$. The variance $\sigma^2(1)$ of the individual values of $z$ is also, of course, known or can be easily calculated for any particular distribution of $z$. In fact, under certain mild and well-known restrictions, usually

$$\sigma^2(n_0) = \sigma^2(1)/n_0, \tag{70·1}$$

provided that $n_0$ is small compared to $N_0$, the total number of cells; otherwise there is a correcting term. The above equation may also be written in the form

$$\sigma^2(n_0)/\sigma_2(1) = 1/n_0. \tag{70·2}$$

71. Now go back to the variance for any particular, say, the $s$th micro-state or space-frequency distribution belonging to the above abstract distribution of $z$. For configurational sampling in the $s$th micro-state, $\sigma_s^2\{Gr(n_0)\}$ has been used to represent the variance of individual grids each consisting of $n_0$ basic cells to distinguish it from the corresponding variance $\sigma^2\{R(n_0)\}$ for samples each consisting of $n_0$ values of $z$ drawn purely at random, that is, for unrestricted sampling. For individual basic cells or individual values of $z$ it is, however, clear that the variance in the case of the space-frequency distribution $\sigma_s^2\{Gr(1)\}$ is identical with $\sigma^2(1)$, the variance in the case of the corresponding abstract distribution for all micro-states. Thus

$$\sigma_s^2\{Gr(1)\} = \sigma^2(1) \quad \text{for} \quad s = 1, 2, \ldots, N. \tag{71·1}$$

72. The two values $\sigma_s^2\{Gr(n_0)\}$ and $\sigma^2\{R(n_0)\}$ are, however, not in general equal. At this stage this may be accepted as a matter of empirical observation. In order to compare these two variances it is convenient to consider their ratio, namely, $\sigma_s^2\{Gr(n_0)\}/\sigma^2\{R(n_0)\}$. This ratio will depend on (1) the nature of the abstract distribution of $z$; (2) the nature of the particular space-frequency distribution, that is, of the particular $s$th micro-state under consideration; and (3) the value of $n_0$. (It will also of course depend on the value of the quad $\square$; but, as already mentioned, such dependence is implied throughout and is not being explicitly stated.) In the case of sample surveys there will always be an upper limit to the size of the grid or the sampling unit $n_0$. That is, in actual sampling practice $n_0$ will have a maximum value in any actual situation; and the ratio $\sigma_s^2\{Gr(n_0)\}/\sigma^2\{R(n_0)\}$ can be adopted,

where $n_0$ is such a maximum value as a criterion for purposes of classification. On the other hand, it is clear that for configurational or grid sampling the minimum value of $n_0$ is two; and the most compact form of a grid is simply a sample unit consisting of two adjoining cells. Obviously this also may be adopted as the standard basis for comparison. That is,

$$\sigma_s^2\{Gr(2)\}/\sigma^2\{R(2)\}$$

may be adopted as the criterion for the purpose of classification—it being understood that the grid is to consist of two adjoining basic cells. In the case of two-dimensional fields, for any given network of basic cells, the junction of the two will be along either of two standard orthogonal directions. Thus $Gr(2)$ will involve two possibilities which may be written $Gr(0, 1)$ and $Gr(1, 0)$, so that $\sigma_s^2\{Gr(2)\}$ for any $Gr$ will have two alternative values.

73. Call this ratio $\sigma_s^2\{Gr(2)\}/\sigma^2\{R(2)\}$ say $\theta_s$. It is clear that $\theta_s$ will have a definite value for each space-frequency distribution, that is, for each micro-state. In this way a series of $N$ values of $\theta_s$ may be obtained which can be ordered in, say, ascending magnitude. For purposes of classification it becomes necessary at this stage to adopt two suitable critical values of $\theta$ on two sides of unity which may be called $\theta_0$ and $\theta_0'$ respectively. $\theta_0$ and $\theta_0'$ may be chosen as definite magnitudes, or $\theta_0$ and $\theta_0'$ may be chosen in such a way that a definite proportion of values of $\theta_s$ (such as 5% or 1% or 1‰) lie outside the range $\theta_0-\theta_0'$. It is clear that in this way the $N$ different values of $\theta_s$ can be separated into two classes: (*a*) one consisting of all values of $\theta_s$ falling outside the critical range $\theta_0-\theta_0'$, and (*b*) the other class consisting of all values such that $\theta_0<\theta_s<\theta_0'$. In this way the corresponding micro-states would also be separated into two distinct classes: (*a*) those micro-states for which $\theta_s$ or the ratio

$$\sigma_s^2\{Gr(2)\}/\sigma^2\{R(2)\}$$

lies outside the critical range $\theta_0-\theta_0'$, and (*b*) those micro-states for which $\theta_0<\theta_s<\theta_0'$. All micro-states or fields falling in the first group (*a*) may now be defined to be of a non-random type at the assigned level of significance. On the other hand, all micro-states or fields falling in the second group (*b*) may be considered to be of a random type at the assigned level of significance.

74. It is worth while explaining at this stage one point on which depends the success or failure of the present methods of differentiation between random and non-random fields. The notation for grid statistic for the $s$th micro-state has already been introduced, namely, $u_{st}\{Gr(n_0)\}$, $(t = 1, 2, \ldots, N')$, and random-sample statistic for the same micro-state or random-sample statistic for the abstract distribution $u_t\{R(n_0)\}$, $(t = 1, 2, \ldots, N'')$; the associated frequency distributions have been called $F_s[u_{st}\{Gr(n_0)\}]$ and $F_0[u_t\{R(n_0)\}]$, where the first $t$ varies from 1 to $N'$ and the second from 1 to $N''$. Considering all the $NN'$ values of $u_{st}\{Gr(n_0)\}$ with $s$ varying from 1 to $N$ and $t$ from 1 to $N'$ (that is, summing over any micro-state and then over all such micro-states), another abstract distribution of $u_{st}\{Gr(n_0)\}$ is obtained which is called $F[u_{st}\{Gr(n_0)\}]$. Apart from the question of a border effect (which is negligible when the area covered by the grid is small compared to the total area and which otherwise merely introduces a correcting term), $F$ may be identified with $F_0$, but is usually slightly or largely different from $F_s$ $(s = 1, 2, \ldots, N)$.

75. Instead of considering the distributions $F_s$ ($s = 1, 2, \ldots, N$), $F$ or $F_0$, consider their moments (of any order) or any other property or characteristic of the distribution which may be denoted by $C_s$ ($s = 1, 2, \ldots, N$), $C$ or $C_0$ respectively. In this case $C$ is to be identified with $C_0$; but the $C_s$'s will usually differ from $C$ or $C_0$, some slightly and others largely. These $C_s$'s ($s = 1, 2, \ldots, N$) form a frequency distribution. Depending on (*a*) the form of the abstract-frequency distribution, (*b*) the nature of the physical field, (*c*) the statistic $u_{st}\{Gr(n_0)\}$ chosen, and (*d*) the particular characteristic $C_s$, $C$ and $C_0$ selected to describe the frequency distributions $F_s[u_{st}\{Gr(n_0)\}]$, $F[u_{st}\{Gr(n_0)\}]$ and $F_0[u_t\{R(n_0)\}]$, the distribution of $C_s$'s (i) will usually give a heaped curve, (ii) will become more and more heaped as $N_0$ (and hence $N$) is increased, and (iii) the value which $C_s$ approaches, that is, the point about which the peak grows with increasing $N_0$ and $N$, is usually the value $C_0$ or $C$. If (i), (ii) and (iii) happen to be true then $C_s$ may be spoken of as stochastically or probabilistically converging to $C_0$ or $C$. It is likely that in most cases (ii) and (iii) will follow from (i), but this cannot be assumed in the abstract set up. Under conditions (i), (ii) and (iii) for any value of $N_0$ (and $N$) the centiles of the distribution of $C_s$'s can usefully be plotted ($s = 1, 2, \ldots, N$), and it is possible to judge whether at any assigned level of significance a given field is of random or non-random type, fields for which the values of $C_s$'s are near enough $C_0$ (or $C$) being considered random and those for which $C_s$'s are farther off being considered non-random. The success of this method of differentiation essentially depends upon (i), (ii) and (iii) being true, which again depend on (*a*), (*b*) and (*c*) and (*d*). The important point to be noted is that concrete examples have actually been found (e.g. variance function) of (*a*), (*b*) and (*c*) for which (i), (ii) and (iii) hold good.

76. The formulation adopted here may be briefly described in the following way. In the case of what has been called the abstract distribution, there are ${}_{N_0}C_{n_0}$ or $N''$ samples; and corresponding to each sample there is a statistic in the Fisherian sense. That is, there are altogether $N''$ statistics. Corresponding to the abstract distribution there are $N$ micro-states or space distributions. For configurational sampling with grids, each of which consists of $n_0$ cells, there are $N'$ different locations of grids for each micro-state. Corresponding to each such location there is a value of the statistic which may be called the grid value. There are thus $N'$ grid values for each micro-state, so that altogether there are $NN'$ grid values for the complete set of $N$ micro-states. This number will be usually much larger than $N''$ that is, many of the grid values would occur in more than one micro-state. In addition to the sample statistics, the expectation of grid values over each particular micro-state which may be considered to be a certain characteristic of the distribution of the grid values over that particular micro-state, deserve attention. Thus there are $N$ values of the particular characteristic for the $N$ different micro-states. The point to be emphasized is that these characteristics are neither statistics in the direct Fisherian sense nor parameters. The characteristics are expectation values from one point of view, and yet have their own distribution over the different micro-states.

77. It is worth noting that here is found an analogue of Bernoulli's theorem in probability or the law of large numbers. The functions $F_0$, $F$ and $F_s$'s ($s = 1, 2, \ldots, N$) together with values of the characteristics $C_s$ ($s = 1, 2, \ldots, N$) associated with the micro-states raise a whole body of new problems of sampling distribution, of estimation, and of testing of

hypothesis which may be regarded as space generalizations of the corresponding classical theory for abstract distributions.

78. Coming back to $\sigma_s^2\{Gr(n_0)\}$ and $\sigma^2\{R(n_0)\}$, it is observed that the ratio of

$$\sigma_s^2\{Gr(n_0)\}/\sigma^2\{R(n_0)\} \equiv \theta_s$$

supplies a simple procedure for the separation of random and non-random micro-states or fields which is adequate for the present purpose. It is thus seen that any given observed field which occurs in nature may be classified as belonging to either the random or the non-random type at an assigned critical level of significance. This distinction has several important consequences which are now stated without proof:

(1) For unrestricted random sampling, that is, for any sampling procedure in which each individual value of $z$ or each individual basic cell is drawn at random, the variance of the mean value of $z$ based on samples of $n_0$ is the same for random and non-random fields and is equal to the variance of the corresponding abstract distribution of $z$.

(2) For configurational or grid sampling, in which each grid (consisting of a group of $n_0$ basic cells arranged in a definite pattern) is located at random, the variance $\sigma_s^2\{Gr(n_0)\}$ in fields of a random type is equal to the variance $\sigma^2\{R(n_0)\}$ for the corresponding abstract distribution.

79. In terms of our criterion $\theta_s \equiv \sigma_s^2\{Gr(n_0)\}/\sigma^2\{R(n_0)\}$, the above results may be stated in the following form:

(1) For unrestricted random sampling $\theta_s = 1$ for fields of both random and non-random types.

(2) For configurational or grid sampling $\theta_0 < \theta_s < \theta_0'$ for fields of a random type.

(3) For configurational or grid sampling $\theta_s$ lies outside the critical range $\theta_0 - \theta_0'$ for fields of a non-random type.

80. These results are based partly on empirical and partly on logical considerations—considerations which will be elaborated at a later stage. As regards (3) it is sufficient to note that fields for which the variance of grids is statistically different from the variance of the corresponding abstract distribution have been actually observed to occur in nature. Model or artificial fields can also be easily constructed to illustrate this. As regards propositions (1) and (2) a proof under certain mild restrictions for linear or uni-dimensional fields is being given in a Note in Appendix 1 to this Part. Theoretical aspects of the question are being investigated in the Statistical Laboratory, Calcutta, and the results of these investigations will be published in due course.

81. The choice of the critical region $\theta_0 - \theta_0'$ is no doubt arbitrary in the same sense as the choice of the critical level in tests of significance; in the ultimate analysis the choice has to be guided by considerations of practical usefulness. In the present case also the choice of $\theta_0 - \theta_0'$ would ultimately depend on what difference is caused to the cost of operations (or alternatively to the precision of the final estimate) by the random or non-random character of the field. Fields for which no appreciable difference in either cost or error is caused by treating them as either of a random or a non-random type for purposes of sampling technique, may be obviously classified as of random type. On the other hand, for fields for which greater

economy can be secured by treating them as being of a non-random type should naturally be considered to belong to this type. The basic idea is clear; and this question will be discussed further in a later section.

*Zonal sampling*

82. So far consideration has been given to the case of the unitary field or a single zone. It is, however, possible to extend the treatment without difficulty to the case of zonal sampling. Here the whole field is divided into a suitable number of zones. A little consideration will show that each of these zones may be either of a random or of a non-random type in the sense explained above. All the zones may be of a random type, or all the zones may be of a non-random type. These are the two forms which usually occur in nature. It is, however, possible in the same field for certain zones to be of a random and certain other zones to be of a non-random type; but such cases are probably rare. From the point of view of sampling technique, methods of zonal unrestricted and zonal configurational methods have to be considered. But it is not necessary at this stage to develop the abstract formulation in greater detail.

## 2. The variance function

83. Consideration is now given in greater detail to the sampling variance of the estimated mean value of $z$ based on all possible samples. The uni-stage case is first considered. It will be remembered that for an abstract distribution neither unitary configurational nor zonal unrestricted nor zonal configurational sampling has any meaning or relevance. In the case of the abstract distribution $\sigma^2(1)$ was used to denote the variance of individual values of $z$. The notation will be changed slightly and written as $V(1)$. This will always exist in the case of a finite population consisting of $N_0$ values, but may or may not exist for an infinite population.

84. Now consider samples of $n_0$ values of $z$ drawn at random. The total number of ways in which such samples of $n_0$ may be drawn out of $N_0$ values is ${}_{N_0}C_{n_0}$, which may be written as $N''$; $z_t\{R(n_0)\}$ has been used to denote the mean value of $z$ based on such $n_0$ values, $R(n_0)$ indicating that they are drawn at random; $V[z_t\{R(n_0)\}]$ is written to represent the sampling variance of $z$ over all possible values of $t$ from 1 to $N''$. When there is no chance of confusion this will sometimes be written more simply as $V[R(n_0)]$ or even $V(n_0)$. This also will always exist for finite populations but may or may not exist for infinite populations. For a finite population of $N_0$ basic cells or values of $z$ it follows that

$$V[z_t\{R(n_0)\}] \equiv V[R(n_0)] \equiv V(n_0) = \frac{V(1)}{n_0}\left\{1 - \frac{n_0 - 1}{N_0 - 1}\right\}, \tag{84·1}$$

and for an infinite population when $V(1)$ and $V(n_0)$ both exist

$$V(n_0) = V(1)/n_0, \tag{84·2}$$

which will be called the normal form.

85. Now consider physical fields. It has been seen that from any given abstract distribution it is possible to generate a system of space distributions or micro-states. Now consider the $s$th micro-state, and also samples of $n_0$ basic cells drawn either individually at random or in the form of a grid. Consider the $t$th sample for the $s$th micro-state. The mean

value of the $n_0$ values of $z$ in the unrestricted case can be written as $z_{st}\{R(n_0)\}$, while the corresponding mean values in the case of grid sampling is written as $z_{st}\{Gr(n_0)\}$. Now write the variance of $z_{st}\{Gr(n_0)\}$ over the $s$th micro-state, i.e. by summing over all possible values of $t$ from 1 to $N'$, as $V_s[z_{st}\{Gr(n_0)\}]$, which will sometimes be written as $V_s[Gr(n_0)]$. In the case of unrestricted sampling the corresponding variance will be written as $V_s[z_{st}\{R(n_0)\}]$, where the summation is over all values of $t$ from 1 to $N''$ in this case.

When $n_0 = 1$, then

$$V_s[z_{st}\{Gr(1)\}] = V_s[z_{st}\{R(1)\}] = \sigma^2(1) = V(1) \tag{85·1}$$

for finite populations, and also for infinite populations when $V(1)$ exists.

86. Now compare the results for two, say, the $s$th and $p$th micro-states, in the case of unrestricted sampling in the first instance. In each micro-state there are $N_0$ basic cells out of which $n_0$ cells are drawn individually at random in each sample. The total number of ways of doing this is ${}_{N_0}C_{n_0}$, which has been written as $N''$. This number $N''$ is the same for all micro-states, as well as for the abstract distribution. Every possible combination of $n_0$ values selected out of $N_0$ values of $z$ would thus occur in all these cases. From this it follows that for unrestricted sampling

$$\begin{aligned} V_s[z_{st}\{R(n_0)\}] &= V_p[z_{pt}\{R(n_0)\}] = V[z_t\{R(n_0)\}] \\ &= \frac{V(1)}{n_0}\left\{1-\frac{n_0-1}{N_0-1}\right\} = V(n_0). \end{aligned} \tag{86·1}$$

That is, for unrestricted sampling the variance of the estimated mean value of $z$ over all possible samples for each micro-state is the same and is equal to the corresponding variance in the case of the abstract distribution. Another line of proof (which is also applicable in the case of infinite populations) is indicated in the section on the correlation function.

87. Now consider configurational or grid sampling. Here there is one definite value of $V_s[z_{st}\{Gr(n_0)\}]$ (obtained by summing for all values of $t$ from 1 to $N'$ the possible number of different grids in each micro-state) for each micro-state. $N$ such values are obtained as $s$ goes from 1 to $N$. Now write $\mu_4$ as the corrected fourth moment of the abstract distribution of $z$, and $B(N_0)$ as the expectation of

$$N_0^2(V_s[z_{st}\{Gr(n_0)\}]^2) - V[z_t\{R(n_0)\}] \tag{87·1}$$

(over all possible micro-states, i.e. for all values of $s$ from 1 to $N$). If $\mu_4$ is finite for indefinitely large values of $N_0$, then it can be shown in the case of an endless linear field that

$$P[|\,V_s[z_{st}\{Gr(n_0)\}] - V[z_t\{R(n_0)\}]\,| < \epsilon] > 1 - \frac{B(N_0)}{\epsilon^2 N_0^2}, \tag{87·2}$$

where $P$ denotes the probability of $V_s[z_{st}\{Gr(n_0)\}]$ and $V[z_t\{R(n_0)\}]$ differing by less than $\epsilon$, the probability having reference to the frequency distribution of $V_s$ over all possible micro-states, that is, for all values of $s$ from 1 to $N$. (The proof is given in Appendix 1.)

Since $V[z_t\{R(n_0)\}] = V(1)/n_0$ (when $n_0$ is small compared to $N_0$) the above result may be written in the following form:

$$P\left[\left|V_s[z_{st}\{Gr(n_0)\}] - \frac{V(1)}{n_0}\right| < \epsilon\right] > 1 - \frac{B(N_0)}{\epsilon^2 N_0^2}. \tag{87·3}$$

This shows that when $\mu_4$ and $V(1)$ exist the variance of the estimated mean value of $z$ based on $n_0$ basic cells drawn from any micro-state will probabilistically (over the micro-states) converge to $V(n_0)$ or $V(1)/n_0$ by sufficiently increasing $N_0$ (the total number of basic cells) and hence $N$ (the total number of micro-states). In other words it is found that the frequency distribution of $V_s[z_{st}\{Gr(n_0)\}]$ for $s = 1, 2, \ldots, N$ is a heaped up distribution which becomes more and more heaped round the value $V(n_0)$ or $V(1)/n_0$ as $N_0$ (and hence $N$) are increased.

88. It is clear that the values of the variance for different micro-states can be arranged in ascending order, and thus form accumulated frequencies and percentile points. Although most of the values of $V_s[z_{st}\{Gr(n_0)\}]$, which may be written more concisely as $V_s[Gr(n_0)]$ or even as $V_s$, will be heaped up near the value $V(n_0)$ or $V(1)/n_0$, it is clear that some of the values in either tail end will differ by large amounts from $V(n_0)$ or $V(1)/n_0$. This supplies a concrete basis for the classification of the fields into random and non-random types. Suitable critical intervals may be chosen say, $V_c$ and $V_c'$, on either side of $V(1)/n_0$, either as absolute magnitudes or in terms of centile points. The $N$ values of $V_s$ will now fall into two classes: ($a$) those which fall within the range of $V_c$ or $V_c'$ on either side of $V(1)/n_0$, and ($b$) those which fall outside this range. At the assigned critical level all values of $V_s$ falling into class ($b$) must be treated as statistically distinguishable from $V(1)/n_0$. All micro-states corresponding to these values of $V_s$ in class ($b$) are then defined to be of a non-random type. On the other hand, all values of $V_s$ falling into class ($a$) may be considered to be statistically indistinguishable from $V(1)/n_0$, and the corresponding micro-states are considered to be of a random type.

89. On the basis of such classification it is seen that for space distributions of a random type

$$V_s[z_{st}\{Gr(n_0)\}] = V_s[z_t\{R(n_0)\}] = V(1)/n_0, \tag{89·1}$$

where the sign of equality is to be interpreted as indicating statistical indistinguishability at the assigned critical level.

90. In the case of fields of a non-random type the sampling variance of the estimated mean value based on configurational sampling is, on the other hand, statistically different (at the assigned critical level) from the corresponding variance for the abstract distribution.

91. Extensive sampling experiments on practically every natural field studied so far have shown that the above result is true for such variates as crop acreage or crop yields. In most cases it was also found that the variance of the mean value in configurational sampling could be graduated by an equation of the form

$$V_s[z_{st}\{Gr(n_0)\}] = V_s[Gr(n_0)] = b/(\square n_0)^g, \tag{91·1}$$

where $b$ and $g$ are constants implicitly supposed to depend on $s$.

*Zonal sampling*

92. Consideration may now be given to zonal sampling, taking up the uni-stage case in the first instance. Here there are $l$ different abstract distributions based on different numbers $N_{0k}$ ($k = 1, 2, \ldots, l$), and corresponding to each there is one micro-state or zone (out of $N_k$ possibilities with $k = 1, 2, \ldots, l$) in the actual field under survey. For the given set of $l$ abstract distributions there are thus $N_1 \times N_2 \times \ldots \times N_l$ (or, say $N$) micro-states. Let any such micro-state be called $s$ with $s = 1, 2, \ldots, N$. Let $z_k$ be the value of the variate per unit area estimated

in the $k$th zone. Then the estimated total value for the whole area under investigation is given by

$$\bar{z} = \sum_{k=1}^{l} A_k z_k. \tag{92·1}$$

Since $n_k$ grids are taken *at random* in any $k$th zone and independently of a similar number of grids in any other zone, the sampling variance or the corrected second moment of $z$ over all possible samples will evidently be given by

$$V_s = \sum_{k=1}^{l} A_k \psi_k \{Gr_k(\square_k, n_{0k}), w_k\} \tag{92·2}$$

in the most abstract formulation, where $Gr_k\{\square_k, n_{0k}\}$ indicates that samples are drawn in the form of grids, each of which consists of $n_{0k}$ basic cells each of area $\square_k$, and $w_k = n_k/A_k$ is written so that $w_k$ is the density of grids per unit area. This $\psi_k$ will be called the variance function. It is suggested by experience that in the same region the basic cell $\square_k$ and the grid pattern $Gr_k$ may be conveniently taken to be the same for all zones and may, therefore, be replaced by $\square$ and $Gr$. Our experience has also been that $\psi_k$ has the same functional form $\psi$ (though the numerical values of the parameters do vary in this case from zone to zone) for different zones. Further, in actual practice $\square$ is usually replaced by some conventional unit, and its explicit mention may be dropped although its presence is implied throughout. Hence (92·2) may be replaced by

$$V_s = \sum_{k=1}^{l} A_k \psi\{Gr(n_{0k}), w_k, d_k\}, \tag{92·3}$$

where $d_k$ now collectively stands for a group of parameters which might differ from zone to zone, being functions of $z_k$ and other zonal characteristics, and where the form of the function $\psi$ depends on the random or non-random nature of the field and, among other factors, on the particular grid pattern chosen.

93. I have already mentioned that in many cases it was found that a special form of the function $\psi$, namely, $b/(n_0\square)^g$, gave good graduations. Adopting this form, and introducing the suffix $k$ to denote the different zones, it follows that

$$V_s = \sum_{k=1}^{l} A_k b_k / w_k (n_{0k})^{g_k}, \tag{93·1}$$

where $b_k$ depends on $z_k$ for the zone in question, and $g_k$ also is a zonal constant involving $z_k$ and the grid pattern $Gr_k(\square_k, n_{0k})$ or in the simpler case $Gr(\square, n_{0k})$ or more simply $Gr(n_{0k})$. It may be mentioned here that $b_k$'s and $g_k$'s will implicitly depend upon the particular micro-state $s$ chosen ($s = 1, 2, \ldots, N$), but this dependence is not here explicitly stated. Now put

$$V_0 \equiv \sum_{k=1}^{l} A_k b_{-k} / (n_{0k} w_k), \tag{93·2}$$

where $b_k$'s ($k = 1, 2, \ldots, l$) are supposed to be constants depending upon the $l$ abstract distributions but independent of the space distributions. It is surmised (and can be also proved —but this will be discussed in a later paper) that $V_s$'s will form a frequency distribution heaped up about $V_0$—the heaping up being sharper and sharper the more we increase $N_{0k}$'s ($k = 1, 2, \ldots, l$). In technical language $V_s$'s stochastically converge to $V_0$ by sufficiently

increasing $N_{0k}$'s ($k = 1, 2, \ldots, l$). For any set of $N_{0k}$'s the frequency distribution of $V_s$'s can be cut at any convenient assigned level; and any given space distribution (over $l$ zones) is to be called random or non-random at the assigned level of significance on the basis of its calculated $V_s$.

94. In a more general set-up for the random sample, $V_0$ of (93·2) is replaced by

$$V_0 \equiv \sum_{k=1}^{l} A_k \psi_k \{\square_k, (n_{0k} w_k)\}, \tag{94·1}$$

where the $(Gr)$ symbol is omitted to indicate that the variance is independent of the grid configurations. Assuming a constant value for the basic cell size and the same general form $\psi$ for the general variance function then reduces to

$$V_0 \equiv \sum_{k=1}^{l} A_k \psi \{\square, (n_{0k} w_k), d_{0k}\}. \tag{94·2}$$

95. Certain results regarding the variance function have already been stated in the case of unitary unrestricted and unitary configurational sampling for the abstract distribution as well as for space distributions of both random and non-random types. Consideration may now be given to zonal fields. As mentioned before, a zonal random field would consist of a number of different zones with different mean values of the variate $z$ but each zone being separately a random-space distribution. Similarly, a zonal non-random field would consist of a number of zones each of which is a space distribution of a non-random type. Certain results will now be stated for zonal fields which appear to be plausible in the light of actual experimental studies and partly also on logical grounds.

(1) Under unitary unrestricted sampling, zonal fields classified into both random and non-random types (under the criterion already mentioned and at any assigned critical level) will have the same variance function of form $V(n_0)$, where $n_0$ is the total number of basic cells or units of conventional area included in the sample.

(2) Under unitary configurational sampling zonal fields of random and non-random types, classified in the same way as in (1), will have different variance functions, each being, however, different from $V(n_0)$.

(3) Under zonal unrestricted sampling, zonal fields of both zonal random and non-random types will have, under certain broad restrictions, variance functions of the form (94·1) or (94·2) or (93·2).

(4) Under zonal configurational sampling, fields of the zonal random type will have variance functions of the form (94·1) or (94·2) or (93·2) as under (3) for zonal unrestricted sampling; but fields of a non-random type will have variance functions of the form (92·2), (92·3) or (93·1).

96. As observed earlier this variance function is but one of the various possible means to distinguish between space distributions of random and non-random types. Other methods will be considered in later sections, and the consistency between the different methods will be discussed.

97. Thus far for uni-stage sampling. There is an obvious extension to multi-stage sampling which involves really nothing new in principle and which will be considered later.

But brief consideration may be given to the case of several variates say $p$ in number. As before $i = 1, 2, \ldots, p$ will be used to represent the different variates, say $p$ different crops. $z_{ki}$ will be written as the mean value of the $i$th variate in the $k$th zone. The estimated total value for the $i$th variate would be given by

$$\sum_{k=1}^{l} (A_k z_{ki}) \quad \text{for} \quad i = 1, 2, \ldots, p. \tag{97·1}$$

98. It is possible of course to have different grid patterns $Gr_{ki}$ and basic cells $\square_{ki}$, and different values of $n_{0ki}$ and $w_{ki}$ for different crops in the same zone. That is, in the most general case the variance function for zonal non-random fields under zonal configurational sampling may have to be written in the form

$$\sum_{k=1}^{l} A_k \psi_{ki}\{Gr_{ki}(\square_{ki}, n_{0ki}), w_{ki}\}, \tag{98·1}$$

where $i = 1, 2, \ldots, p$. But usually it will be possible to keep the grid pattern $Gr_{ki}$ and the values of $\square_{ki}$ and $w_{ki}$ the same for all variates in the same zone, in which case the variance function for zonal non-random fields under zonal configurational sampling will be given by

$$\sum_{k=1}^{l} A_k \psi_{ki}\{Gr_k(\square_k, n_{0k}), w_k\}. \tag{98·2}$$

This is the abstract set up. The more concrete one (based on the generally adopted procedure) will be

$$\sum_{k=1}^{l} A_k \psi_i\{\square, (n_{0k} w_k), d_{ki}\}, \tag{98·3}$$

where the group of constants $d_{ki}$ involve not merely the physical peculiarities of the zone but also the mean value $z_{ki}$ for the particular variate $i$ and the particular zone $k$. As in the case of one variate so here also in many situations, and (mostly with grids of compact shape) it has been found that (98·3) takes the form

$$\sum_{k=1}^{l} A_k b_{ki} / w_k (n_{0k})^{g_{ki}} \quad (i = 1, 2, \ldots, p), \tag{98·4}$$

where $b_{ki}$ and $g_{ki}$ depend on $z_{ki}$ and the physical peculiarities of the zone.

## 8. The correlation function

99. The efficiency of uni-stage sampling depends on the variance function. In actual planning of such surveys the determination of the variance function is, therefore, a problem of great importance. This question can be studied in two ways—by direct experimental work on the field, and secondly, by model sampling experiments in the Laboratory on the basis of material in the form of a complete enumeration or inventory of particular regions. In this method an area of a convenient size is surveyed in detail and a map is prepared; and, with the help of this map, model sampling experiments are carried out with various sizes of grids. In making these studies it was found that the use of certain auxiliary methods is often convenient.

100. One such auxiliary tool, namely, the correlation function which is closely linked to the variance function and throws a good deal of light on it, will now be described. The

general approach can be explained very briefly. In a non-random field the decrease in the variance with increasing size of samples is less than that in a random field. A little consideration will show that this may be ascribed to the existence of correlation between the values of the variates in neighbouring cells. This is the basic idea which will now be developed.

101. Suppose now that the field under survey, the $s$th micro-state, is rectangular and consists of $N_0 = N_1 \times N_2$ cells, there being $N_1$ columns each consisting of $N_2$ cells, and $N_2$ rows each consisting of $N_1$ cells. (This particular shape of the field is convenient but does not affect the generality of the results.) Each cell can be identified with the help of a pair of co-ordinate numbers $(i', j')$, where $i'$ goes from 1 to $N_1$ and $j'$ from 1 to $N_2$. As usual, the value of the variate for any basic cell will be represented by $z(i', j')$ for any micro-state. (The suffix $s$ is dropped as the results are true for all micro-states.) Consider a second cell separated from the above cell by a gap of $(i, j)$ cells; the value of the variate for this second cell would then be $z(i'+i, j'+j)$. If the size and shape of the gap remains constant, that is, $(i, j)$ is constant, then the value of $i'$ goes from 1 to $(N_1-i)$, and $j'$ goes from 1 to $(N_2-j)$. The co-variance between $z(i', j')$ and $z(i'+i, j'+j)$ can now easily be written down. $V\{\ \}$ will be written for the variance, $\sigma\{\ \}$ for the standard deviation of the variate mentioned within brackets, and $\theta\{\ \}$ for the covariance and $\rho\{\ \}$ for the correlation between the two variates mentioned within brackets. Thus

$$\theta\{z(i', j'), z(i'+i, j'+j)\} = \sigma\{z(i', j')\}\, \sigma\{z(i'+i, j'+j)\}\, \rho\{z(i', j'), z(i'+i, +j'+j)\}, \quad (101{\cdot}1)$$

and hence

$$\rho\{z(i', j'), z(i'+i, j'+j)\} \equiv \rho(i, j) = \frac{\Sigma[\{z(i', j') - \overline{z(i', j')}\}\{z(i'+i, j'+j) - \overline{z(i'+i, j'+j)}\}]}{[\Sigma\{z(i', j') - \overline{z(i', j')}\}^2]^{\frac{1}{2}} \times [\Sigma\{z(i'+i, j'+j) - \overline{z(i'+i, j'+j)}\}^2]^{\frac{1}{2}}}, \quad (101{\cdot}2)$$

where the summation is to be taken over $i' = 1, 2, \ldots, (N_1-i)$, and $j' = 1, 2, \ldots, (N_2-j)$; and $\overline{z(i', j')}$ and $\overline{z(i'+i, j'+j)}$ represent mean values over domains indicated by the $(i', j')$ summation. The function defined by the right-hand side of the above equation (101·2) will be called the correlation function of a pair of cells separated by a fixed gap $(i, j)$ and will be written as $\rho(i, j)$.

102. The variance of the mean value of $z(i', j')$ and $z(i'+i, j'+j)$ would be given by

$$V[\tfrac{1}{2}\{z(i', j') + z(i'+i, j'+j)\}] = \tfrac{1}{4}V\{z(i', j')\} + \tfrac{1}{4}V\{z(i'+i, j'+j)\} + \tfrac{1}{2}\theta\{z(i', j'), z(i'+i, j'+j)\}$$
$$= \tfrac{1}{4}V\{z(i', j')\} + \tfrac{1}{4}V\{z(i'+i, j'+j)\} + \tfrac{1}{2}\sigma\{z(i', j')\}\, \sigma\{z(i'+i, j'+j)\}\, \rho(i, j). \quad (102{\cdot}1)$$

103. When the gap $(i, j)$ is small compared to the size of the total field the $(i', j')$ summation can be taken over the entire field, and in such a case

$$V[\tfrac{1}{2}\{z(i', j') + z(i'+i, j'+j)\}] = \tfrac{1}{2}V\{z(i', j')\}[1 + \rho(i, j)], \quad (103{\cdot}1)$$

where $V\{z(i', j')\}$ is, of course, the variance of individual cells and is what has previously been called $V(1)$; the left-hand side of (103·1) is the variance function for grids each consisting of a pair of cells at a gap $(i, j)$ which may be written more concisely as $V(i, j)$. Thus for any micro-state it follows that

$$V(i, j) = \tfrac{1}{2}V(1)\,[1 + \rho(i, j)]. \quad (103{\cdot}2)$$

Thus a relation between the variance function and the correlation function which is really of one-to-one correspondence is established.

104. In particular, for any micro-state $i = 0, j = 1$, or $i = 1, j = 0$ may be written, and in this case $V(0,1)$ or $V(1,0)$ may be expressed as

$$\begin{bmatrix} V(0,1) \\ V(1,0) \end{bmatrix} = \tfrac{1}{2}V(1)\left[1+\begin{matrix}\rho(0,1) \\ \rho(1,0)\end{matrix}\right]. \tag{104·1}$$

or

$$\begin{bmatrix} V(0,1) \\ V(1,0) \end{bmatrix} \Big/ V(1) = \tfrac{1}{2}\left[1+\begin{matrix}\rho(0,1) \\ \rho(1,0)\end{matrix}\right], \tag{104·2}$$

where $\rho(0,1)$ or $\rho(1,0)$ represent the correlation between adjoining cells in two different orthogonal directions. This gives us a one-to-one relation between the variance and correlation function at unit gap, i.e. for adjoining cells.

105. For a grid pattern with $n_0$ cells forming a certain configuration let one cell be at $(i', j')$ and other cells be at $(i'+i, j'+j)$ with certain given values of $i$ and $j$; $(i', j')$ will of course run over nearly the whole field subject to certain restrictions near the boundary necessitated by the condition that no cell of the grid may go out of the field. Then

$$V\left[\frac{1}{n_0}\sum_{i,j} z(i'+i, j'+j)\right] = \frac{1}{n_0^2}\left[\sum_{i,j} V\{z(i'+i, j'+j)\}\right]$$
$$+\frac{1}{n_0^2}\sum_{i_1, i_2, j_1, j_2} [\sigma\{z(i'+i_1, j'+j_1)\}\,\sigma\{z(i'+i_2, j'+j_2)\}\,\rho\{z(i'+i_1, j'+j_1), z(i'+i_2, j'+j_2)\}], \tag{105·1}$$

where $$i_1 \neq i_2, j_1 \neq j_2, \text{ and } \rho\{z(i'+i_1, j'+j_1), z(i'+i_2, j'+j_2)\}$$

may be changed into $\rho(i_1-i_2, j_1-j_2)$, the summation being taken over all possible values of $i_1, i_2, j_1, j_2$ subject to the condition that $i_1 \neq i_2, j_1 \neq j_2$.

106. In particular, when the area covered by the grid pattern is small compared to that of the total field under investigation $V\{z(i'+i, j'+j)\}$ can be taken to be the same as $V\{z(i', j')\}$, and furthermore in $V\{z(i', j')\}$ itself $(i', j')$ may be supposed to have run over the whole field, the usual boundary effect being ignored; in such a case $V\{z(i'+i, j'+j)\}$ can be replaced by $V\{z(i', j')\}$, which is now the same for all micro-states. In such a case $V\{z(i', j')\}$ reduces to $V(1)$; and equation (105·1) may be replaced by

$$V\left[\frac{1}{n_0}\sum_{i,j}\{z(i'+i, j'+j)\}\right] = \frac{V(1)}{n_0^2}\left[n_0 + \sum_{i_1, i_2, j_1, j_2}\{\rho(i_1-i_2, j_1-j_2)\}\right], \tag{106·1}$$

the summation for $\rho$ being over all values of $i_1, j_1, i_2, j_2$, subject to restrictions already indicated. There are, therefore, $n_0(n_0-1)$ terms in the $\rho$-summation, or with a duplication really ${}_{n_0}C_2$ terms. The correspondence between the variance function for a grid pattern and the correlation function for different possible pairs taken out of the cells constituting the grid pattern is thus not one-to-one correspondence. The above relation, however, supplies a convenient and labour-saving device for studying the variance function with the help of the correlation function.

107. The correlation function also opens out the possibility of studying questions of optimum size and shape of blocks in agricultural field trials from a new point of view. The saving in computational work in any case would be considerable. But this is not all, the

correlation function is likely to throw some light on the theory of design of experiments in the case of non-random fields.

108. It will be noticed that the treatment adopted here is a kind of generalization for two dimensions of the method of serial correlation in the case of time series on which a large volume of work is already in existence. In fact, in many instances results for serial correlation are capable of being extended in two dimensions without difficulty. But, as can be easily seen from the above discussion, many new problems arise in the case of the two-dimensional correlation function which have no analogue in the case of serial correlation.

109. For simplicity the concept of correlation function has been developed with reference to the unitary configuration type of sampling. A corresponding development (involving, however, nothing fundamentally new in principle) may be easily given with reference to other types of sampling.

110. It is clear that in the same manner as the variance function the correlation function may be used as a tool for differentiation between random and non-random space distributions. The general principles have been explained earlier. It may be useful to recapitulate the procedure with special reference to the correlation function. Consider an abstract distribution and the totality of all possible $N$ associated space distributions or micro-states which can be generated from the abstract distribution by all possible distributions of the values of $z$ over the different cells. For any given abstract distribution there are thus a bundle of $N$ micro-states. For each micro-state there is a definite value of $\rho_s(i,j)$, and these $\rho_s$'s $(s = 1, 2, \ldots, N)$ for all micro-states may be now arranged in the form of a frequency distribution with centile points for $\rho_s(i,j)$.

111. On partly experimental and partly intuitive grounds there are reasons for believing that the frequency distribution of $\rho_s$, for any given gap $(i,j)$ over different micro-states $(s = 1, 2, \ldots, N)$, will have a form such that there is a heaping up near zero and a falling off at the tail ends $\pm 1$. Furthermore, by sufficiently increasing $N_0$ or $N$ this frequency distribution of $\rho_s$ (over $s = 1, 2, \ldots, N$) becomes more and more peaked up about the value zero. The theoretical discussion of these properties of the frequency distribution of $\rho_s$ (for $s = 1, 2, \ldots, N$) is being reserved for a subsequent paper. On positive and negative sides of zero suitable critical levels are now chosen, say, $\rho_0$ and $\rho_0'$, cutting at, say, 5 % by the tail ends of the above-mentioned frequency distribution of $\rho_s$. If an observed micro-state or space distribution is found to have a value of $\rho_s(i,j)$ not lying between $\rho_0$ and $\rho_0'$, then at the assigned level of significance (which must be necessarily arbitrary depending upon the stringency with which the classification is made) it can be asserted that the micro-state in question is non-random. If the observed $\rho_s(i,j)$ lies between $\rho_0$ and $\rho_0'$, then the observed micro-state may be considered to be of random type at the assigned level of significance. The choice of the critical level in this case (as in the case of the variance function) will depend, as already pointed out, ultimately on whether any material difference is made or not in the cost of operations by treating the field under survey as belonging to the random or non-random type.

112. One point requires to be emphasized at this stage. Attention is here confined (as in the case of the variance function) to fields which are fundamentally non-periodic in character. The implications in the case of the correlation function may be indicated in the following way. Consider any particular, say, the $s$th micro-state, then $\rho_s(i,j)$ has a deter-

minate value for a pair of the gap co-ordinates $(i, j)$. Consideration may now be given to the variation in the value of $\rho_s(i, j)$ as $(i, j)$ are increased from $(0, 1)$ or $(1, 0)$. By a non-periodic field is meant a space distribution in which the value of $\rho_s(i, j)$ on the whole (that is, possibly with minor fluctuations) decreases as the values of $i$ and/or $j$ are increased. In a fuller theory it is necessary of course to take into consideration periodic or quasi-periodic fields, but for present purposes this is not essential, as fields which occur in nature in the case of large-scale sample surveys have been found so far to be practically of the non-periodic type.

113. Now study a little more closely the nature of the correlation function $\rho_s(i, j)$. It is found from (102·1) for any, say, the $s$th micro-state

$$\sigma_s\{z(i', j')\}\sigma_s\{z(i'+i, j'+j)\}\rho_s(i, j) = \text{Expectation } [\{z_s(i', j') - \overline{z_s(i', j')}\}\{z_s(i'+i, j'+j) - \overline{z_s(i'+i, j'+j)}\}], \tag{113·1}$$

the expectation and the mean values (indicated by the bars) being taken over appropriate domains. Now relax the condition that $(i, j)$ refers to a fixed gap and consider $(i, j)$ to denote any possible gap whatsoever, then $\rho_s(i, j)$ would be replaced by $\rho_s(2)$ (which is of course different from $\rho_s(0, 1)$ or $\rho_s(1, 0)$, which refer to adjoining cells) and would simply mean the correlation between any possible pair of cells (identical cells also not being excluded)—the summation being taken over all parts of the given space distribution. In such a case the right-hand side of (113·1), which is

$$\sum_{i', j'} \sum_{i, j} \{z_s(i', j') - \overline{z_s(i', j')}\}\{z_s(i'+i, j'+j) - \overline{z_s(i'+i, j'+j)}\},$$

may be conveniently written as

$$\sum_{i', j'} \{z_s(i', j') - \overline{z_s(i', j')}\} \sum_{i, j} \{z_s(i, j) - \overline{z_s(i, j)}\}. \tag{113·2}$$

But $\sum_{i, j} \{z_s(i, j) - \overline{z_s(i, j)}\}$ is evidently zero. Hence (113·2) is zero.

114. The following statements can therefore be made, which can be justified either on empirical grounds or on grounds partly logical and partly intuitive:

(1) Under unitary unrestricted sampling the correlation function is (statistically) zero for all micro-states. In our symbolic language this can be expressed as

$$\rho_s(2) = 0, \quad (s = 1, 2, \ldots, N).$$

(2) Under unitary configurational sampling the correlation functions over different micro-states will be heaped up about the value zero and will be more and more heaped up about that value the more $N_0$ (the number of basic cells) and hence $N$ (the number of possible micro-states) is increased. For any $N_0$ the distribution of $\rho_s$'s over $s = 1, 2, \ldots, N$ yields percentile points and thus enables one to judge whether, at any preassigned level of significance, a given micro-state is to be called random or non-random on the basis of its correlation function. For a number of space distributions, either observed in nature or artificially generated by processes (functional or otherwise) different from those associated with random numbers, values of $\rho_s$ have been observed which are greater than any reasonable critical value. Such space distributions or micro-states may be called non-random, the

criterion being, in the present case, the correlation function for a fixed gap which may be in particular (1, 0) or (0, 1), or the average value of the correlation between adjoining cells. In our symbolic language this can be expressed as

$$P[|\rho_s(i,j)-\rho_s(2)|<\epsilon]>1-\delta, \tag{114·1}$$

where $\epsilon$ is any arbitrary small number and $\delta$ is a small number depending on $\epsilon$ and $N_0$. Since $\rho_s(2) = 0$ for all values of $s$, this can be otherwise expressed as

$$P[|\rho_s(i,j)|<\epsilon]>1-\delta. \tag{114·2}$$

115. It has been seen, however, that in the case of any pair of cells with a fixed separation, and hence for adjoining cells, there exists a one-to-one correspondence between the variance and the correlation functions. It is possible therefore to make the classification of micro-states or space distributions into random and non-random types consistent by a suitable choice of the critical levels in the two cases.

116. From the results given above it is easy to generalize to properties of the correlation function under (3) zonal unrestricted, or (4) zonal configurational sampling for (*a*) zonally random or (*b*) zonally non-random space distributions—just as was done for variance function. The differentiation between random and non-random will, of course, be made by cutting the frequency distribution of $\rho_s$'s (over $s = 1, 2, \ldots, N$) at any conventional percentile point. This, however, need not be further pursued here.

117. Certain properties of the correlation function for any fixed gap $(i, j)$ will now be discussed. For analytic convenience the field is supposed to be made up of an infinite number of infinitesimal basic cells; and it is also supposed that $z$ is a continuous variate distributed over the field. The further assumption is made that the field is of non-random type and another assumption to be presently indicated. Under such assumptions $\rho(i, j)$ for any micro-state may be given with sufficient approximation by

$$\rho(i,j) = 1-\frac{1}{2V(z)}\left[\iint\left(i\frac{\partial z}{\partial x}+j\frac{\partial z}{\partial y}\right)^2 dx\,dy\right] \tag{117·1}$$

$$= 1-\frac{1}{2V(z)}\iint(\vec{t}\,\mathrm{grad}\,z)^2\,dx\,dy$$

$$= 1-\frac{1}{2V(z)}\left[i^2\iint\left(\frac{\partial z}{\partial x}\right)^2 dx\,dy+j^2\iint\left(\frac{\partial z}{\partial y}\right)^2 dx\,dy+2ij\iint\left(\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\right)dx\,dy\right], \tag{117·2}$$

where $\vec{t}$ denotes the gap $(i, j)$ regarded as a vector, $V(z)$ is the variance of $z$, and where it is assumed that in the Taylor's expansion of $z(x+i, y+j)$ in terms of $z(x, y)$ and its derivatives, the remainder term after the third, is negligible compared to the sum of the foregoing terms. This would be approximately true (*a*) if $(i, j)$ is small compared to our total area under study, and if further (*b*) either $(i, j)$ is small compared to what is conventionally called unit area or length, or (*c*) the field is so slowly fluctuating that approximately

$$z(i'+i, j'+j) = a_0+(a_1 i+a_2 j)+(a_{11}i^2+a_{22}j^2+2a_{12}ij), \tag{117·3}$$

where the parameters $a_0, a_1, a_2, a_{11}, a_{22}, a_{12}$ are expressible in terms of $z$ and its differential coefficients at $(i', j')$.

118. For a non-continuous field (cut up into a finite number of irreducible basic cells) the sign of integration $\iint(\ )\,dx\,dy$ will be replaced by $\Sigma$ which is the sign of summation, and $\partial z/\partial x$ and $\partial z/\partial y$ will be replaced by $(\Delta_x z)$ and $(\Delta_y z)$. The formulae (117·2) will then be replaced by

$$\rho(i,j) = 1-\frac{1}{2V(z)}\left[i^2\Sigma(\Delta_x z)^2+j^2\Sigma(\Delta_y z)^2+2ij\Sigma(\Delta_x z)\,(\Delta_y z)\right]. \qquad (118\cdot1)$$

Formulae (117·2), (117·3) and (118·1) will hold under certain restrictions already mentioned. Denoting $\iint\left(\frac{\partial z}{\partial x}\right)^2 dx\,dy$ or $\Sigma(\Delta_x z)^2$ by $G_{xx}$, $\iint\left(\frac{\partial z}{\partial y}\right)^2 dx\,dy$ or $\Sigma(\Delta_y z)^2$ by $G_{yy}$ and $\iint\left(\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\right)dx\,dy$ or $\int(\Delta_x z)\,(\Delta_y z)$ by $G_{xy}$, the general formula may be written

$$\rho(i,j) = 1-\frac{1}{2V(z)}\left(i^2G_{xx}+j^2G_{yy}+2ijG_{xy}\right), \qquad (118\cdot2)$$

where $V(z)$ is the variance of $z$ which has also been written as $V(1)$ or $\sigma^2(1)$.

119. A proof of formulae (117·2) or (118·1) under restrictions already mentioned is given in Appendix 2 at the end of this Part. It may be noted here that these restrictions are not satisfied in the case of random-space distributions. Hence for such random-space distributions the above formulae would not be valid; and in fact (117·2) would have no significance, as $\rho(i,j)$ would be statistically zero for all values of $(i,j)$. For a non-random (or patterned) field these formulae may or may not hold, depending upon the nature of distribution of the variate over the field and the size of the gap $(i,j)$. For fields and gaps in which the above formulae are valid, the quantities $G_{xx}$, $G_{yy}$ and $G_{xy}$ may be regarded as fundamental parameters in the characterization of the field in question. Having calculated these parameters the correlation function for a gap $(i,j)$ can be obtained by any of the foregoing formulae and hence the variance function by (105·1) or (106·1). This, of course, considerably simplifies the task of calculation of either the variance function for any grid pattern or correlation function for any gap.

120. For a non-random field and a gap not subject to the restrictions under which the foregoing formulae are valid, the correlation function is given by another group of approximate formulae which will be stated and proved in a later paper. Those formulae, too, will be valid under certain restrictions which, however, are much milder than those already indicated. Natural fields seldom conform to the first set of stringent restrictions but often do conform to the milder restrictions just referred to. Artificial fields can be constructed which conform to the first set of restrictions given above. Thus artificial fields exist for which the previous correlation formula will hold, and experimental data in confirmation are given in Appendix 4. Natural fields are also met with for which some of the results discussed here are found to be true. In the same Appendix illustrative numerical results are given to indicate the nature of the empirical evidence.

## 4. Cost function

121. Associated with each of the four broad types of sampling considered in the foregoing section there is a total cost of operations for the survey and a sampling error for that quantity (in the first instance only one statistical variate is considered) which it is proposed

to estimate by the sampling procedure. Consider first for simplicity the case of uni-stage sampling which as already mentioned is usually more suitable for the estimation of area under a particular crop or crops. In this method a number $n_k$ of grids (each composed of $n_{0k}$ basic cells of area $\square_k$ arranged in a particular geometrical pattern) are located at random in the $k$th out of the $l$ zones into which the whole area under investigation is divided, each zone being broadly homogeneous in respect of the character under survey. It is considered that the most general or zonal configurational case where the area $\square_k$ and the shape of the basic cell, as also the shape and nature of the geometrical pattern constituting the grid, may vary from zone to zone. In actual practice, however, in most cases the basic cell will be taken to be of square shape and of the same area $\square$ for all zones, and the geometrical configuration for the grid will be taken to be the same for all zones and will be supposed to be predetermined on physical grounds. Thus for the general case the grid configuration will be denoted by $Gr_k(\square_k, n_{0k})$, where $k = 1, 2, \ldots, l$, and in actual practice this will usually reduce to $Gr(\square, n_{0k})$. As observed earlier, the grid might in a particular case be a compact block of $n_0$ basic cells placed side by side in the form of, say, a square or a rectangle. It will be noticed that (*a*) by putting $k = 1$, $n_k = n$, $n_{0k} = 1$ and $\square_k = \square$ in the zonal configurational case the unitary unrestricted is obtained; (*b*) by putting $k = 1$, $n_k = n$ and $\square_k = \square$ the unitary configurational; and (*c*) by putting $n_{0k} = 1$ the zonal unrestricted cases.

122. Let $A_k$ be the area in some convenient unit (say, a square mile) of the $k$th zone. As in the section on the variance function $w_k$ denotes the number of grids per unit area (say, per square mile). Then it is clear that the total cost of survey will depend on $\square_k$, $A_k$, $n_{0k}$ and $w_k$, on the nature of the configuration constituting the grid $Gr_k(\square_k, n_{0k})$, and also on certain physical characteristics of the field under inquiry which might be different for the different zones and which might differentially affect the cost of operations. Symbolically the total cost of operations for the most general uni-stage case may be written as

$$T = \sum_{k=1}^{l} A_k \phi_k \{Gr_k(\square_k, n_{0k}), w_k\}. \tag{122·1}$$

This $\phi_k$ is a function of $w_k$ and $n_{0k}$, and the form of this function depends on the geometrical pattern making up the grid; moreover, the functional form itself may vary from zone to zone, and not merely the numerical values of algebraic parameters occurring in the same type of function. This is the most general formulation, and this function will be called the 'Cost Function'.

123. Its form and nature cannot be determined on *a priori* grounds; it must be settled empirically by suitable experiments. A good deal of work has been done on this subject, of which an account will be given in the next Part. It has been found in actual practice that $\square_k$ may be conveniently taken to be the same in all zones, and may thus be called $\square$; the geometrical pattern of the grid may also be taken to be the same for all zones and may be settled from physical considerations; furthermore, the form of the function $\phi_k$ remains steady from zone to zone, and even the numerical values of the algebraic parameters entering into the form usually remain steady over zones, though it would be safe to keep open the possibility of variation in the numerical values of the parameters from zone to zone. Hence with the form of geometrical pattern predetermined the 'Cost Function' may in most actual cases be taken as $\phi(\square, n_{0k}, w_k, c_k)$, where the form of $\phi$ depends on the nature of the

geometrical pattern of the grid and involves a group of algebraic parameters collectively called $c_k$ which might vary from zone to zone. Thus the total cost will be given by

$$T = \sum_{k=1}^{l} A_k \phi(\square, n_{0k}, w_k, c_k). \qquad (123{\cdot}1)$$

124. When the inquiry concerns several variates (for example, a number of crops) an abstract scheme could be set up in which the grid pattern is different for different crops, and in general the $i$th crop could be taken as $\square_{ki}$, $n_{0ki}$ and $w_{ki}$, the grid denoted by $Gr_{ki}(\square_{ki}, n_{0ki}), \ldots$ $(i = 1, 2, \ldots, p)$ and the cost replaced (122·1) by

$$T = \sum_{k=1}^{l} A_k \phi_k\{Gr_{k1}(\square_{k1}, n_{0k1}),\ Gr_{k2}(\square_{k2}, n_{0k2}), \ldots Gr_{kp}(\square_{kp}, n_{0kp}), w_{k1}, w_{k2}, \ldots, w_{kp}\}. \qquad (124{\cdot}1)$$

But experience suggests that this is unnecessary. Even in the case of several crops or variates it is ultimately economical and convenient to work with the same grid pattern and basic cell for all crops. Thus, even in the more general case, work will actually be done with (123·1). The only point to be remembered is that for several crops the $\phi_k$ of the more general formulation and $\phi$ of the concrete formulation would be different from those of (122·1) or (123·1) for one crop alone; but even here experience suggests that the functional form of $\phi_k$ or $\phi$ remains the same no matter whether one crop or several is under consideration—only the numerical values of the algebraic parameters change. But here, however, in the case of several crops, the more general possibility of a different functional form is kept open.

125. Thus far about the cost function in the abstract. One or two points about the concrete contents of the functions (122·1) or (123·1), brought out by extensive field experiments in Bengal, may be briefly explained here; a fuller discussion is reserved for a later section. In the case of a sample survey the work falls into two broad groups, namely, field operations and statistical work (which includes the planning of the survey, preparation of grids and making preliminary arrangements, tabulation and analysis of the field material). For studying the cost function arrangements were made for both field and statistical workers to keep diaries showing the amount of time spent on different kinds of work. For the field survey the cost is in the first instance reckoned in labour units or man-hours—one investigator working for one hour constituting one man-hour. These labour units are later converted into money value. In doing this it is necessary of course to include not only the pay of the field investigator but also the pay of the inspecting and supervising staff, travelling expenses, and all incidental and contingent charges. In the same way the cost of the statistical branch is measured in the first instance in terms of computer-hours—one computer working for one hour constituting a single computer-hour. Here also the cost would include the pay of the computers as well as the pay of the inspecting and supervising staff, the cost of calculating machines, accommodation, stationery, etc.

126. After a good deal of experimentation it was found convenient to study the time-cost of field operations under three different heads. First of all there is the time required for travelling from one sampling unit to another, which, for convenience of reference, is called the 'journey time'. This naturally would depend on how far apart the sample units are located, that is, on the density of the grids, but should be independent of the size of the grids; and this has been fully corroborated by actual field records. Next comes the time

required for locating the grids, identifying the plots, examining the crops and making necessary field records. This has been called the 'enumeration time'; it should increase with the size of the grids but should be independent of their density. This also has been corroborated by experimental studies. In large-scale sample surveys the 'journey time' is by no means negligible in comparison with the 'enumeration time'. In fact, in the case of the jute survey the 'journey time' is actually three times larger than 'enumeration time'. This is the special feature in large-scale work which makes necessary the approach adopted in the present paper. Besides the 'journey' and the 'enumeration' time there is also the time required for miscellaneous work which consumed practically a constant fraction of the day. In the case of the statistical work also the total number of computer-hours could be split up into different components. One portion of the work depended merely on the total area covered and was independent of both size and density of grids. Certain items depended on the size but not on the density of grids and certain items on both size and density of the sample units. In the case of uni-stage sampling the general scheme has become fairly clear and has been discussed in Part III. The cost function for multi-stage sampling naturally would be more complicated, but may be developed on similar lines, both in the abstract as well as in the concrete. No new principles are, however, involved and further consideration is reserved for a subsequent paper.

## 5. Solution of the optimum size-density distribution of grids

### A. *Abstract scheme*

127. The optimum size and density of grids will now be investigated, and in the first instance a single variate is considered. Here a change over is effected from basic cells $\square_k$ to any conventional unit of area (like acre or square mile), which of course may be taken to be the same over all zones. In the uni-stage case it will be supposed that $n_{0k}$ such conventional units of area arranged in any definite manner constitute a grid—the possibility is kept open of fractions of this conventional unit area functioning as individuals in the composition of the grid. All the functions and solutions will, of course, depend upon the absolute size of this conventional unit area; but in the subsequent sections the area of the conventional unit is dropped, the dependence of the functions and solutions on this being, however, throughout implied. Each grid consists of $n_{0k}$ such units, and $A_k w_k$ such grids are located at random in the $k$th zone. In actual practice in the uni-stage case grid patterns forming a compact square or rectangle have been used, but this will not be assumed in the general set-up. For generality a zonal non-random field or space distribution under the zonal configurational type of sampling with the grid pattern varying from zone to zone is considered. The estimated total value of the variate will be given by

$$z = \sum_{k=1}^{l} [A_k z_k]. \tag{127·1}$$

The appropriate variance function and the cost function respectively may be written in the following forms:

$$V = \sum_{k=1}^{l} [A_k \cdot \psi_k\{Gr_k(n_{0k}), w_k\}], \tag{127·2}$$

$$T = \sum_{k=1}^{l} [A_k \cdot \phi_k\{Gr_k(n_{0k}), w_k\}]. \tag{127·3}$$

The problem now can be solved in either of the two equivalent alternative forms:

(A) Given a value of $T$, to choose $n_{0k}, w_k$ and also the grid pattern $Gr_k$ for any $k$th zone $(k = 1, 2, \ldots, l)$ so as to minimize $V$; or alternatively

(B) Given $V$ or $V/z$ to choose the foregoing entities so as to minimize $T$.

128. The two questions will, as can be easily shown, yield the same answer, that is, the same grid pattern and the same set of values of size $n_{0k}$ and density $w_k$ of grids over the different zones. For the first problem the solution is

$$\delta V + \lambda \delta T = 0, \tag{128·1}$$

with a given value $T$; and for the second problem the solution is

$$\delta T + \mu \delta V = 0, \tag{128·2}$$

with a given value for $V$, using in either case Lagrange's principle of undetermined multipliers. In these equations $\delta$ refers to variation not merely of $n_{0k}, w_k$ but also of the nature of the grid pattern and the basic cell in the different zones. Further, it is also evident that (128·1) and (128·2) would lead to the same solutions, which shows that questions (A) and (B) yield the same answer. The possibility of a variation in the grid pattern introduces complications, and really links up this problem with the calculus of variations. However, this will not be discussed in the present paper.

129. For $p$ variates (or crops) some additional assumption must be introduced to make the problem definite. Attaching arbitrary weights $\lambda_i$ $(i = 1, 2, \ldots, p)$ (determined by, say, the money value of the different variates or by some such meta-statistical consideration), the total value of the variates estimated is given by

$$\sum_{i=1}^{p} \sum_{k=1}^{l} A_k \lambda_i z_{ki}. \tag{129·1}$$

The variance function for this quantity is

$$\sum_{k=1}^{l} A_k \Big[ \sum_{i=1}^{p} \lambda_i^2 \psi_{ki}\{Gr_k(n_{0k}), w_k\} + \sum_{i,j=1}^{p} \lambda_i \lambda_j \theta_{kij}\{Gr_k(n_{0k}), w_k\} \Big], \tag{129·2}$$

where $\theta_{kij}\{Gr_k(n_{0k}), w_k\}$ may be regarded as a covariance function for the $i$th and $j$th variates by analogy with the variance function $\psi_{ki}$ for the $i$th variate, and where $i \neq j$. The cost function, as noted in the previous section, is of the same general form

$$T = \sum_{k=1}^{l} A_k \phi_k\{Gr_k(n_{0k}), w_k\}. \tag{129·3}$$

Hence for optimum distribution it is seen that either

$$\delta V + \lambda \delta T = 0 \tag{129·4}$$

with a given value of $T$, or

$$\delta T + \mu \delta V = 0 \tag{129·5}$$

with a given value of $V$. Here also the symbol $\delta$ carries the same implications as formerly.

130. For several variates, however, a different set-up is possible. For instance, the optimum distribution at given (different) levels of error for the different variates might be

required—the optimum having reference to minimum total cost. In that case the solution would be

$$\delta T+\sum_{i=1}^{p}\mu_i\delta\left[\sum_{k=1}^{l}A_k\psi_{ki}\{Gr_k(n_{0k}),w_k\}\right]=0, \tag{130·1}$$

where $\mu_i$'s are undetermined multipliers. Other formulations are possible for the $p$-variate problem. But the above two are the most fruitful.

131. This is so far as the abstract scheme is concerned. Under actual conditions of sampling the cost and variance functions for the uni-stage, uni-variate take simpler forms:

$$V=\sum_{k=1}^{l}A_k\psi(n_{0k},w_k,d_k), \tag{131·1}$$

$$T=\sum_{k=1}^{l}A_k\phi(n_{0k},w_k,c_k). \tag{131·2}$$

For simplicity assume that the conventional unit area, as also the nature of the grid pattern, is settled beforehand. This would facilitate the discussion without, however, unduly restricting the generality of the formal solution. For convenience of mathematical analysis assume also that both $w_k$ and $n_{0k}$ are continuous variates. A little reflexion will show that this assumption is not unjustifiable. With this set-up it is seen that

$$\delta V+\lambda\delta T=0, \quad \text{or} \quad \delta T+\mu\delta V=0 \tag{131·3}$$

with a given value of $T$ in the first case, and a given value of $V$ in the second case. These lead to

$$\left.\begin{aligned}\frac{\partial\psi}{\delta w_k}+\lambda\frac{\partial\phi}{\partial w_k}&=0,\\ \frac{\partial\psi}{\partial n_{0k}}+\lambda\frac{\partial\phi}{\partial n_{0k}}&=0,\end{aligned}\right\}\ (k=1,2,\ldots,l) \tag{131·4}$$

with

$$\sum_{k=1}^{l}A_k\phi(n_{0k},w_k,c_k)=T=\text{a given value}, \tag{131·41}$$

or

$$\left.\begin{aligned}\frac{\partial\phi}{\partial w_k}+\mu\frac{\partial\psi}{\partial w_k}&=0,\\ \frac{\partial\phi}{\partial n_{0k}}+\mu\frac{\partial\psi}{\partial n_{0k}}&=0,\end{aligned}\right\}\ (k=1,2,\ldots,l) \tag{131·5}$$

with

$$\sum_{k\ 1}^{l}A_k\psi(n_{0k},w_k,d_k)=V=\text{a given value}. \tag{131·51}$$

In both cases there are as many equations as there are unknowns; the equations are usually independent and compatible; and usually there is also one solution giving a true minimum.

132. If the variance function is of the special form

then

$$\left.\begin{aligned}\psi&=b_k/w_k(n_{0k})^{g_k},\\ \frac{\partial\psi}{\partial w_k}&=-\frac{b_k}{w_k^2(n_{0k})^{g_k}},\\ \frac{\partial\psi}{\partial n_{0k}}&=-\frac{g_kb_k}{w_k(n_{0k})^{g_k+1}}.\end{aligned}\right\}\ (k=1,2,\ldots,l) \tag{132·1}$$

Particularizing the form of the cost function $\phi$ the solution can easily be carried through to the stage of exact determination of $w_k$ and $n_{0k}$ for different zones, that is, the optimum distribution of size and density over the zones at a given level of cost or variance. Some such solutions based on particular forms of the cost function will be given at the end of this section.

133. Equations like (131·4) or (131·5) will be called 'optimum equations', the meaning being quite clear. In most practical cases, however, a functional or algebraic solution cannot be obtained of these equations giving $w_k$ or $n_{0k}$ as explicit functions of the parameters $(c_k, d_k)$ regarded as algebraic quantities. In practice, it is necessary to be content with a numerical solution of the problem (giving the size and density distribution in terms of numerically known parameters), which is usually obtained by a combination of numerical and graphical methods. This will be discussed in a later section. For certain simple forms of the cost function and variance function it may be possible, of course, to obtain explicit algebraic solution of the size-density distribution.

134. Now turn to the case of $p$-variates. Starting with a linear compound of the variates (the weights $\lambda_i$ being determined on meta-statistical grounds), the following total estimated value, as already noted, is obtained:

$$\sum_{i=1}^{p} \sum_{k=1}^{l} A_k \lambda_i z_{ki}, \tag{134·1}$$

The cost and variance functions are given respectively by

$$V = \sum_{k=1}^{l} A_k \Big[ \sum_{i=1}^{p} \{\lambda_i^2 \psi(n_{0k}, w_k, d_{ki})\} + \sum_{i,j=1}^{p} \{\lambda_i \lambda_j \theta(n_{0k}, w_k, d_{kij})\} \Big], \tag{134·2}$$

$$T = \sum_{k=1}^{l} A_k \phi(n_{0k}, w_k, c_k), \tag{134·3}$$

where $\theta$ might be regarded as the covariance function for the $i$th and $j$th variates by analogy with the variance function $\psi$, and where $i \neq j$ and $d_{kij}$ are zonal parameters depending upon both the variates $(i, j)$. The optimum size-density distribution will be given by either

$$\left.\begin{aligned} \frac{\partial}{\partial w_k}\Big\{ \sum_{i=1}^{p} \lambda_i^2 \psi(i) + \sum_{i,j=1}^{p} \lambda_i \lambda_j \theta(i,j) \Big\} + \mu \frac{\partial \phi}{\partial w_k} = 0, \\ \frac{\partial}{\partial n_{0k}}\Big\{ \sum_{i=1}^{p} \lambda_i^2 \psi(i) + \sum_{i,j=1}^{p} \lambda_i \lambda_j \theta(i,j) \Big\} + \mu \frac{\partial \phi}{\partial n_{0k}} = 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{134·4}$$

with a given $T$, or

$$\left.\begin{aligned} \frac{\partial \phi}{\partial w_k} + \lambda \frac{\partial}{\partial w_k}\Big\{ \sum_{i=1}^{p} \lambda_i^2 \psi(i) + \sum_{i,j=1}^{p} \lambda_i \lambda_j \theta(i,j) \Big\} = 0, \\ \frac{\partial \phi}{\partial n_{0k}} + \lambda \frac{\partial}{\partial n_{0k}}\Big\{ \sum_{i=1}^{p} \lambda_i^2 \psi(i) + \sum_{i,j=1}^{p} \lambda_i \lambda_j \theta(i,j) \Big\} = 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{134·5}$$

with given $V$. In the summations there is, of course, the previous restriction $i \neq j$.

135. As already noted, for the $p$-variate case there is another set-up where the cost is minimized, subject to the sampling error of the different variates being kept at different given levels. For such a case

$$\left.\begin{aligned}\frac{\partial\phi}{\partial w_k}+\sum_{i=1}^{p}\mu_i\frac{\partial\psi(i)}{\partial w_k}&=0,\\ \frac{\partial\phi}{\partial n_{0k}}+\sum_{i=1}^{p}\mu_i\frac{\partial\psi(i)}{\partial n_{0k}}&=0,\end{aligned}\right\}\quad(k=1,2,\ldots,l;\ i=1,2,\ldots,p)\qquad(135\cdot1)$$

with given values of $\sum_{k=1}^{l}A_k\psi(n_{0k},w_k,d_{ki})$.

In (134·4) and (134·5) $\mu$ and $\lambda$ respectively are Lagrange's undetermined multipliers to be determined by the equations, while the $\lambda_i$'s are assumed to be determined by meta-statistical considerations. In the case of (135·1) $\mu_i$'s $(i=1,2,\ldots,p)$ are Lagrange's undetermined multipliers to be determined by the equations. In either case there are as many equations as there are unknowns; and, in actual practice, the equations are usually independent and compatible.

## B. *Special forms of solution*

136. Having discussed the optimum solution in a general form consideration will now be given, one by one, to all four types of sampling, assuming, however, a variance function of the following special form which has been found to give satisfactory results in practice:

$$\sum_{k=1}^{l}\left[\frac{A_kb_k}{w_k(n_{0k})^{g_k}}\right].\qquad(136\cdot1)$$

In each case a solution will be given with a general form for the cost function as well as a particular form. The solution with a particular form again will be given both for $(a)$ $g_k\neq1$ generally, and for $(b)$ $g_k=1$, which will be called conventionally the 'normal' case.

137. A start may be made from the variance and cost equations (93·1) and (123·1)

$$V=\sum_{k=1}^{l}\frac{A_kb_k}{w_k(n_{0k})^{g_k}},\qquad(137\cdot1)$$

$$T=\sum_{k=1}^{l}A_k\phi(w_k,n_{0k},c_k),\qquad(137\cdot2)$$

where the size of the conventional unit area (e.g. one acre) in terms of the quad □ or basic cell has been dropped. The optimum size-density distributions would be given by

$$\left.\begin{aligned}-\frac{b_k}{w_k^2(n_{0k})^{g_k}}+\lambda\frac{\partial\phi}{\partial w_k}&=0,\\ -\frac{g_kb_k}{w_k(n_{0k})^{g_k+1}}+\lambda\frac{\partial\phi}{\partial n_{0k}}&=0,\end{aligned}\right\}\quad(k=1,2,\ldots,l)\qquad(137\cdot3)$$

with a given value $T$ in the equation (137·2). With this set-up consideration may be given now to the cases one by one.

(1) *Unitary-unrestricted* ($k = 1$, $n_{0k} = 1$)

138. $c_k$ or $c$ may be dropped, it being understood that the cost function may involve certain parameters which involve regional peculiarities which can be settled beforehand. In such a case the cost function takes the simple form $\phi(w)$, and (137·3) would now reduce to

$$\left.\begin{aligned} -\frac{b}{w^2}+\lambda\frac{\partial\phi}{\partial w} &= 0, \\ A\phi(w) &= \text{constant.} \end{aligned}\right\} \tag{138·1}$$

But here the first equation is superfluous, giving as it does $\lambda$ in terms of $w$; from the second equation $w$ can be found, and substituting this in the variance $b/w$ it is easy to find the error for the given cost. No question of optimum therefore arises here; also no special function need be considered. (If, of course, $\phi$ be such that there are several positive values of $w$ for each cost, then one of these will lead to minimum error, in which case an optimum will have some relevance, but this is unlikely to occur in practice. In any case, the first equation serves no useful purpose.)

(2) *Unitary configurational* ($k = 1$, $n_{0k} = n_0$)

139. Dropping $c$ the cost function reduces to $\phi(n_0, w)$ and (97·3) reduces to

$$\left.\begin{aligned} -\frac{b}{w^2(n_0)^g}+\lambda\frac{\partial\phi}{\partial w} &= 0, \\ -\frac{gb}{w(n_0)^{g+1}}+\lambda\frac{\partial\phi}{\partial n_0} &= 0, \\ A\phi(n_0, w) &= \text{constant.} \end{aligned}\right\} \tag{139·1}$$

From the first two equations there is at the optimum point

$$n_0\frac{\partial\phi}{\partial n_0} = gw\frac{\partial\phi}{\partial w}. \tag{139·12}$$

If further $g = 1$, this reduces to

$$n_0\frac{\partial\phi}{\partial n_0} = w\frac{\partial\phi}{\partial w}. \tag{139·13}$$

For any $\phi$ of the form $c_0n_0w+c_1w$, then

$$\frac{\partial\phi}{\partial w} = c_0n_0+c_1; \quad \frac{\partial\phi}{\partial n_0} = c_0w, \tag{139·14}$$

and the equations now reduce to

$$\left.\begin{aligned} -\frac{b}{w^2(n_0)^g}+\lambda(c_0n_0+c_1) &= 0, \\ -\frac{gb}{w(n_0)^{g+1}}+\lambda c_0w &= 0, \\ c_0n_0w+c_1w &= T/A = T'. \end{aligned}\right\} \tag{139·2}$$

Then from the first two equations

$$\frac{n_0}{g} = \frac{c_0n_0+c_1}{c_0} = n_0+\frac{c_1}{c_0}. \tag{139·3}$$

This and the last equation would easily enable us to determine $n_0$ and $w$. With a cost function of the form

$$T = Ac_0 n_0 w \tag{139·31}$$

(that is, when the total cost is proportional to the product of the grid size and the total number of the grids), the general procedure is unnecessary and misleading. Thus, more simply,

$$n_0 w = T/Ac_0. \tag{139·4}$$

Substituting in the variance formula it is seen that

$$V = Ab/w(n_0)^g = A^2 bc_0/T(n_0)^{g-1}. \tag{139·5}$$

In practical situations that have so far arisen, $0<g\leqq 1$. With this restriction on $g$ and at a particular cost level, the above equation shows that the smaller $n_0$ is made (that is, the larger $w$ is made) the smaller will be the variance or error. If now $g = 1$,

$$V = \frac{A^2 bc_0}{T}. \tag{139·6}$$

Here for a given $T$ there is a unique value of $V$. But both $n_0$ and $w$ are indeterminate, their product alone being fixed by $T = Ac_0 n_0 w$. It will also be seen that $VT$ is constant. As $A^2/V$ is the amount of information (in respect of the estimated proportion of land sown with the crop under survey) in the Fisherian sense, this shows that in the present case the amount of information is simply proportional to the cost.

(3) *Zonal unrestricted* ($n_{0k} = 1$)

140. Equation (137·3) would now reduce to

$$-\frac{b_k}{w_k^2} + \lambda \frac{\partial \phi}{\partial w_k} = 0 \quad (k = 1, 2, \ldots, l), \tag{140·1}$$

$$\sum_{k=1}^{l} A_k \phi_k(w_k, c_k) = T = \text{constant}. \tag{140·2}$$

Here each $w_k$ $(k = 1, 2, \ldots, l)$ is theoretically determinate in terms of $\lambda$. In practice, unless the form of $\phi$ is simple, this determination can only be made by graphical-numerical methods explained in Part III. However, determining (by whichever method) $w_k$'s in terms of $\lambda$, these values are substituted in the cost equation and $\lambda$ obtained in terms of $T$ (again in most cases by graphical numerical methods). Having obtained $\lambda$, the $w_k$'s are found and hence the optimum distribution at the given cost level. In particular, when $\phi(w_k, c_k) = c_k w_k + c_{0k}$, then

$$\left.\begin{aligned} \frac{\partial \phi}{\partial w_k} &= c_k, \\ -\frac{b_k}{w_k^2} + \lambda c_k &= 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{140·2}$$

$$\sum_{k=1}^{l} A_k (c_k w_k + c_{0k}) = T. \tag{140·21}$$

This leads to $w_k^2 = b_k/\lambda c_k$; substituting in the cost equation it follows that

$$\lambda = \frac{\sum_{k=1}^{l} A_k \sqrt{(c_k b_k)}}{(T - \sum_{k=1}^{l} A_k c_{0k})^2}, \tag{140·3}$$

whence all the $w_k$'s can be determined.

141. *Stratified sampling.* Putting $c_{0k} = 0$, Neyman's result for stratified sampling is immediately obtained, which may be stated in the present notation in the following way:

$$w_k = \frac{\sqrt{b_k}}{\sqrt{c_k}} \frac{T}{\sum_{k=1}^{l} A_k \sqrt{(c_k b_k)}}. \tag{141·1}$$

If, in addition, $c_k = c$ (same for all zones), then

$$T = \sum_{k=1}^{l} (A_k w_k) = cn, \tag{141·2}$$

where $n$ is the total number of grids. In this case for any given value $T$ the value of $n$ is prescribed, and the optimum density is given by

$$w_k = \frac{n\sqrt{b_k}}{\sum_{k=1}^{l} A_k \sqrt{b_k}}. \tag{141·3}$$

(4) *Zonal-configurational case*

142. The general solution for this case is provided by (137·3). Consideration will be given to two special cases with two particular forms for the cost function, one constructed artificially and the other based on graduation of field data collected in the course of sample surveys of acreage under jute in Bengal during 1939–41. The variance function also will be of the special form which has been found to give fairly satisfactory results. For the artificial function it is assumed that

$$T = \sum_{k=1}^{l} A_k (c_{0k} + c_{1k} n_{0k} w_k + c_{2k} w_k). \tag{142·1}$$

Then the optimum equations reduce to

$$\left.\begin{aligned} -\frac{b_k}{w_k^2 (n_{0k})^{g_k}} + \lambda(c_{1k} n_{0k} + c_{2k}) &= 0, \\ -\frac{g_k b_k}{w_k (n_{0k})^{g_k+1}} + \lambda c_{1k} w_k &= 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{142·2}$$

$$\sum_{k=1}^{l} A_k (c_{0k} + c_{1k} n_{0k} w_k + c_{2k} w_k) = \text{constant}. \tag{142·3}$$

This again has to be solved by graphical-numerical methods explained in Part III. By eliminating $n_{0k}$ a relation is obtained between $\lambda$ and $w_k$ $(k = 1, 2, \ldots, l)$ from which a $(\lambda, w_k)$ table or graph is constructed, $n_{0k}$ being determinate in terms of $w_k$ from the first two equations. There will be $l$ such tables or graphs. For any $\lambda$ it can be found out from the $l$ different $(\lambda, w_k)$ tables or graphs the $l$ values of $w_k$'s $(k = 1, 2, \ldots, l)$ and the corresponding $n_{0k}$'s;

substituting in the cost equation a value of $T$ is obtained; thus for every $\lambda$ there is a value of $T$. Then a $(\lambda, T)$ graph and table is constructed; now for any $T$ find out the corresponding $\lambda$, and thence the $w_k$'s and $n_{0k}$'s, which would constitute the optimum solution. When $g_k = g$ for $k = 1, 2, \ldots, l$, the procedure becomes more simplified.

143. For the actual jute material the following graduation for the cost function has been used:

$$T = \sum_{k=1}^{l} [A_k(c_0 + c_1 w_k + c_2 n_{0k} w_k + c_3 w_k^2)]. \tag{143·1}$$

For simplicity, using in the variance function a pooled value of $g$ for $g_k$'s, the optimum equations are given by

$$\left.\begin{aligned} -\frac{b_k}{w_k^2 (n_{0k})^g} + \lambda(c_1 + c_2 n_{0k} + 2c_3 w_k) &= 0, \\ -\frac{g b_k}{w_k (n_{0k})^{g+1}} + \lambda(c_2 w_k) &= 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{143·2}$$

with $T$ in (143·1) = constant. Solving these equations (the graphical numerical method for doing this is explained in Part III) for any $T$ a minimum $V$ is obtained, that is, a maximum $\vartheta$, if by $\vartheta \equiv A^2/V$ is denoted the 'information' in the Fisherian sense, that is, the reciprocal of variance of the estimated proportion of land under the crop in question. The minimum $V$, as has been observed, depends upon a proper choice being made of sizes and densities. Call this the optimum information at a particular cost level, and assume for simplicity that there is only one zone for which $\vartheta$, $\left(\frac{\partial \vartheta}{\partial T}\right)$, $\left(\frac{1}{\vartheta}\frac{\partial \vartheta}{\partial T}\right)$ and $\left(\frac{T}{\vartheta}\frac{\partial \vartheta}{\partial T}\right)$ denote respectively the optimum information, the rate of change of information with regard to cost, the proportional change in information due to a small change in cost, and finally the proportional change in information due to a proportional change in cost. Ultimately each is a function of the cost level, the value of '$g$' in the variance function, the value of '$p$' or proportion of area under the crop in question, and finally on the parameters of the cost function. Assuming the parameters to be given (and fixed), it is worth while investigating how the four quantities $\vartheta$, $\left(\frac{\partial \vartheta}{\partial T}\right)$, $\left(\frac{1}{\vartheta}\frac{\partial \vartheta}{\partial T}\right)$ and $\left(\frac{T}{\vartheta}\frac{\partial \vartheta}{\partial T}\right)$ change with a change in either of the quantities $T$, '$g$' and '$p$'.

144. Taking the cost function from the 1941 material for the jute survey and assuming that there is only one zone of area $A$, and solving the equation (143·2) for $k = 1$, it follows that

$$\vartheta = \frac{A^2}{V} = \frac{A}{2b(1+g)c_3}\left\{\frac{g}{(1-g^2)c_2}\right\}^g (U + c_1 g)^g (U - c_1), \tag{144·1}$$

$$\frac{\partial \vartheta}{\partial T} = \left(\frac{1-g^2}{b}\right)\left\{\frac{g}{(1-g^2)c_2}\right\}^g (U + c_1 g)^{g-1}, \tag{144·2}$$

$$\frac{1}{\vartheta}\frac{\partial \vartheta}{\partial T} = 2(1-g^2)(1+g)c_3/A(U-c_1)(U+c_1 g), \tag{144·3}$$

$$\frac{T}{\vartheta}\frac{\partial \vartheta}{\partial T} = 2(1-g^2)(1+g)c_3 T/A(U-c_1)(U+c_1 g), \tag{144·4}$$

where $$U^2 \equiv c_1^2 - 4c_0c_3(1-g^2) + \frac{4c_3(1-g^2)\,T}{A}. \qquad (144{\cdot}5)$$

It should be noted that '$b$', which is involved in $\vartheta$ and $\left(\frac{\partial\vartheta}{\partial T}\right)$, is really a function of $p$ and of other zonal characteristics. From the above equations it is easy to investigate by graphical numerical methods how $\vartheta$, $\left(\frac{\partial\vartheta}{\partial T}\right)$, $\left(\frac{1}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ and $\left(\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ depend on $T$, '$g$' and '$p$'. A numerical table is given in Part III.

*Minimum error or minimum percentage error in each zone*

145. There might be physical situations where, instead of trying to minimize the variance of the estimated total area under a crop, it might be necessary to obtain an optimum size-density distribution over different zones such that either (*a*) variance of the estimated total area in each zone is the same, and the common value is made minimum, or (*b*) $V_k/(A_kz_k)^2$ for each zone is the same and the common value is minimized (which means that the percentage error for each zone is made a minimum). For purpose of illustration, taking a variance function of the form

$$V = \sum_{k=1}^{l}\left[\frac{A_kb_k}{w_k(n_{0k})^g}\right], \qquad (145{\cdot}1)$$

and a cost function of the form (143·1), the solution for both (*a*) and (*b*) will be given. In the first case (*a*), in which the error in each zone is the same and a minimum, it is seen that

$$\left.\begin{aligned} V_k &= \frac{A_kb_k}{w_k(n_{0k})^g} \equiv \frac{1}{\mu} \text{ suppose.} \\ w_k &= \frac{\mu A_kb_k}{(n_{0k})^g}, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \qquad (145{\cdot}2)$$

Then

$$V = \sum_{k=1}^{l}[V_k] = \frac{l}{\mu}, \qquad (145{\cdot}21)$$

$$T = \sum_{k=1}^{l}\left[A_k\left\{c_0 + \frac{\mu A_kb_k(c_1+c_2n_{0k})}{(n_{0k})^g} + \frac{\mu^2(A_kb_k)^2}{(n_{0k})^{2g}}\right\}\right]. \qquad (145{\cdot}22)$$

Now $V$ must be minimized subject to $T$ = constant, the variables being $n_{0k}$'s ($k = 1, 2, \ldots, l$) and $\mu$. The optimum equations come out as

$$-\frac{l}{\mu^2} + \lambda\sum_{k=1}^{l}\left[A_k\left\{\frac{A_kb_k(c_1+c_2n_{0k})}{(n_{0k})^g} + \frac{2\mu(A_kb_k)^2}{(n_{0k})^{2g}}\right\}\right] = 0, \qquad (145{\cdot}3)$$

$$gc_1 + (g-1)\,c_2n_{0k} + \frac{2\mu gA_kb_k}{(n_{0k})^g} = 0. \quad (k = 1, 2, \ldots, l) \qquad (145{\cdot}4)$$

From equations (145·22), (145·3) and (145·4) determination is made of the $(l+2)$ quantities $n_{0k}$'s ($k = 1, 2, l \ldots,$), $\mu$ and $\lambda$ (for the optimum conditions).

146. In the second case ($b$), where the percentage error in each zone is to be made equal and a minimum, it is seen that

$$\left.\begin{aligned} \frac{V_k}{A_k^2 z_k^2} &= \frac{b_k}{A_k z_k^2 w_k (n_{0k})^g} \equiv \frac{1}{\mu} \text{ suppose,} \\ w_k &= \frac{\mu b_k}{A_k z_k^2 (n_{0k})^g}, \end{aligned}\right\} \quad (k = 1, 2, \ldots, l) \tag{146·2}$$

$$V = \sum_{k=1}^{l} [V_k] = \frac{1}{\mu} \sum_{k=1}^{l} [A_k^2 z_k^2], \tag{146·21}$$

$$T = \sum_{k=1}^{l} A_k \left\{ c_0 + \frac{\mu b_k (c_1 + c_2 n_{0k})}{A_k z_k^2 (n_{0k})^g} + \frac{\mu^2 b_k}{A_k^2 z_k^4 (n_{0k})^{2g}} \right\}. \tag{146·22}$$

$V$ must now be minimized subject to $T =$ constant, the variables again being $n_{0k}$'s ($k = 1, 2, \ldots, l$) and $\mu$. The optimum equations come out as

$$-\frac{1}{\mu^2} \sum_{k=1}^{l} [A_k^2 z_k^2] + \lambda \sum_{k=1}^{l} \left[ \frac{b_k (c_1 + c_2 n_{0k})}{z_k^2 (n_{0k})^g} + \frac{2\mu b_k^2}{A_k z_k^4 (n_{0k})^{2g}} \right] = 0, \tag{146·3}$$

$$g c_1 + (g-1) c_2 n_{0k} + \frac{2\mu g b_k^2}{A_k z_k^2 (n_{0k})^g} = 0 \quad (k = 1, 2, \ldots, l). \tag{146·4}$$

The $(l+2)$ quantities $n_{0k}$'s ($k = 1, 2, \ldots, l$), $\mu$ and $\lambda$ (for the optimum conditions) are determined from equations (146·22), (146·3) and (146·4).

147. The above examples sufficiently illustrate the general principles for uni-stage sampling in the case of a single variate. Solutions (with special forms for cost and variance functions) for multi-variate and uni-stage sampling or for multi-stage uni-variate or multi-stage multi-variate sampling are not considered here. Material for satisfactory graduation of appropriate cost functions and variance or co-variance functions is accumulating for certain crops, and it is hoped to deal with the problem in a later paper. Some simple artificial functions might have been discussed here, but for purposes of elucidation that is scarcely worth while, since the uni-variate uni-state examples provide sufficient illustration of the general principles.

## 6. Multi-stage sampling

148. Explanation has already been made in an earlier section of the procedure of multi-stage sampling, but as pointed out in the sections on both cost and variance functions, multi-stage sampling really does not introduce any fundamentally new principles other than those already considered in connexion with uni-stage work. The optimum size-density distribution would also thus offer no new difficulties so far as the abstract theory is concerned. As has been already noted, multi-stage sampling has certain advantages for estimating the yield of crops (be it one crop or several) as distinguished from the problem of estimating the area under crops (one or several) for which the uni-stage sampling would appear to be more appropriate. As mentioned earlier, if there are $s$ stages altogether in a multi-stage sampling then there are $4^s$ possible types of composite sampling, since at each stage the sampling may be of any of the four possible types. It is of course possible to develop an abstract theory of cost function, variance function and optimum distribution for each of these types on the

general lines already described for uni-stage sampling. For purposes of illustration it is, however, enough to consider one particular type of multi-stage sampling.

149. This type may be described as follows. Assume that the total area $A$ under investigation is divided into $N_1$ blocks $A(k_1)$ $(k_1 = 1, 2, \ldots, N_1)$; any block into $N(k_1)$ subblocks $A(k_1, k_2)$ $(k_2 = 1, 2, \ldots, N(k_1))$; any subblock into $N(k_1, k_2)$ finer subblocks $A(k_1, k_2, k_3)$ $(k_3 = 1, 2, \ldots, N(k_1, k_2))$; and so on till we get to $A(k_1, k_2, k_3, \ldots, k_s)(k_s = 1, 2, \ldots, N(k_1, \ldots, k_{s-1}))$. The $A(k_1, k_2, \ldots)$ will indicate both the zones themselves as well as their areas, the sense being read against the context. Now in an area $A$ divided and subdivided in such a manner, select at random $n_1$ out of $N_1$ zones $A(k_1)$. This is the first stage of sampling which gives us a random sample of $n_1$ zones $A(k_1)$ $(k = 1, 2, \ldots, n_1)$. The ordering of $k_1$ in the $n_1$ randomly chosen zones $A(k_1)$ is different from the ordering of $k_1$ in the complete set of $N_1$ zones $A(k_1)$. In any such zone $A(k_1)$ choose at random $n(k_1)$ subzones $A(k_1, k_2)$ $(k_2 = 1, 2, \ldots, n(k_1))$; similarly, in any subzone $A(k_1, k_2)$ choose further $n(k_1, k_2)$ finer subzones $A(k_1, k_2, k_3)$ with $k_3 = 1, 2, \ldots, n(k_1, k_2)$; and continue this process till the last stage is reached of the $s$th order subzones, where in the subzone $A(k_1, k_2, k_3, \ldots, k_s)$ select at random $n(k_1, k_2, \ldots, k_s)$ grids, each grid consisting of $n_0(k_1, k_2, \ldots, k_s)$ basic cells of uniform size. (This uniformity of basic cell size is a simplifying assumption which will not affect the general argument.) Thus there is unitary unrestricted sampling throughout, except in the last stage which is unitary configurational. This case corresponds to the procedure often adopted for crop-cutting experiments on a large scale. As regards ordering, what has been noted in the case of the first stage will hold good throughout, namely, not only will $n_1, n(k_1), n(k_1, k_2), \ldots$, etc., be different from $N_1, N(k_1), N(k_1, k_2), \ldots$, etc., but the ordering of $k$'s and $n$'s and $N$'s would be different, that is, even for the same values of $k_1, k_2, \ldots$, etc., the random subzone $A(k_1, k_2, \ldots)$, etc., associated with $n$'s will be different from the subzone $A(k_1, k_2, \ldots)$ as occurring in the complete set associated with $N$'s. Of course the subzone $A(k_1, k_2, \ldots)$ coming from a random sample will refer to some subzone $A(k_1, k_2, \ldots, k_s)$ of the complete set, but the same values of $k_1, k_2, \ldots, k_s$ will not refer to identical subzones in the two cases.

150. The cost function for such sampling will evidently be of the form

$$T = \phi[\square, R(A), n_0(k_1, k_2, \ldots, k_s), n(k_1, k_2, \ldots, k_s), c(k_1, k_2, \ldots, k_s)], \qquad (150{\cdot}1)$$

where $n(k_1, k_2, k_3, \ldots, k_s)$ stands collectively for the totality of all grids in the different (last order) subzones $A(k_1, k_2, \ldots, k_s)$; $n_0(k_1, k_2, \ldots, k_s)$ for the associated number of basic cells for any grid pattern; $R(A)$ for those zones and subzones that come into the sample; and $c(k_1, k_2, \ldots, k_s)$ for the group of parameters for the particular subzone $A(k_1, k_2, \ldots, k_s)$ and $\square$ is the basic cell size. As already indicated in the equation the functional form of the cost function is considered to be the same over all the subzones, only the parameters differing from zone to zone. The other simplifying assumptions (which do not, however, affect the general argument or set-up) have been already mentioned. It should be remembered that this cost function will hold good only for the area under survey, and also only when that area has been subdivided into zones and subzones and subjected to multi-stage random (with configurational in the last stage) sampling in the manner indicated. For any other zoning and any other type of sampling the formula (150·1)—even in its general abstract form—will have to be changed.

151. The total estimated value of the yield will be given by

$$z = \frac{A}{\sum_{k_1=1}^{n_1} A(k_1)} \sum_{k_1=1}^{n_1} \frac{A(k_1)}{\sum_{k_2=1}^{n(k_1)} A(k_1,k_2)} \sum_{k_2=1}^{n(k_1)} \frac{A(k_1,k_2)}{\sum_{k_3=1}^{n(k_1,k_2)} A(k_1,k_2,k_3)} \cdots$$
$$\cdots \sum_{k_{s-1}=1}^{n(k_1,k_2,\ldots,k_{s-2})} \frac{A(k_1,k_2,\ldots,k_{s-1})}{\sum_{k_s=1}^{n(k_1,k_2,\ldots,k_s)} A(k_1,k_2,\ldots,k_s)} \sum_{k_s=1}^{n(k_1,k_2,\ldots,k_s)} z[Gr\{n_0(k_1,k_2,\ldots,k_s)\}], \quad (151{\cdot}1)$$

where $z[Gr\{n_0(k_1,k_2,\ldots,k_s)\}]$ is the total yield in the $s$th order zone $A(k_1,k_2,\ldots,k_s)$ estimated from $n(k_1,k_2,\ldots,k_s)$ grids of the pattern indicated; and summations everywhere will be taken over zones and subzones constituting the random sample. The variance function associated with the estimated total yield will be given by

$$V = \psi[A, \square, R(A), n_0(k_1,k_2,\ldots,k_s), n(k_1,k_2,\ldots,k_s), d(k_1,k_2,\ldots,k_s)], \quad (151{\cdot}2)$$

where, as before, $A$ denotes the total area; $R(A)$ stands for those zones and subzones that come out into the sample; $\square$, $n_0(k_1,k_2,\ldots,k_s)$ have the same meaning and $d(k_1,k_2,\ldots,k_s)$ a meaning similar to $c(k_1,k_2,\ldots,k_s)$ as in the cost function; and in $n(k_1,k_2,\ldots,k_s)$, $k_s = 1,2,\ldots, n(k_1,k_2,\ldots,k_{s-1})$, in $n(k_1,k_2,\ldots,k_{s-1})$, $k_{s-1} = 1,2,\ldots,n(k_1,k_2,\ldots,k_{s-2})$ and so on until we come to $k_1 = 1,2,\ldots,n_1$. In a purely random space distribution this variance would in general be given by

$$V = \frac{A^2}{\left[\sum_{k_1=1}^{n_1} A(k_1)\right]^2} \sum_{k_1=1}^{n_1} \frac{A^2(k_1)}{\left[\sum_{k_2=1}^{n(k_1)} A(k_1,k_2)\right]^2} \cdots \sum_{k_{s-2}=1}^{n(k_1,k_2,\ldots,k_{s-3})} \frac{A^2(k_1,k_2,\ldots,k_{s-2})}{\left[\sum_{k_{s-1}=1}^{n(k_1,k_2,\ldots,k_{s-2})} A(k_1,k_2,\ldots,k_{s-1})\right]^2}$$
$$\times \sum_{k_{s-1}=1}^{n(k_1,k_2,\ldots,k_{s-2})} \frac{A^2(k_1,k_2,\ldots,k_{s-1})}{\left[\sum_{k_s=1}^{n(k_1,k_2,\ldots,k_{s-1})} A(k_1,k_2,\ldots,k_s)\right]^2} \sum_{k_s=1}^{n(k_1,k_2,\ldots,k_{s-1})} \frac{A^2(k_1,k_2,\ldots,k_s)\, V(1)}{n_0(k_1,k_2,\ldots,k_s)\, n(k_1,k_2,\ldots,k_s)}, \quad (151{\cdot}3)$$

the summation being a multiple one over $k_s = 1,2,\ldots,n(k_1,k_2,\ldots,k_{s-1})$, $k_{s-1} = 1,2,\ldots,n(k_1,k_2,\ldots,k_{s-2})$, etc., and finally $k_1 = 1,2,\ldots,n_1$. The more general form would be of course

$$V = \frac{A^2}{\left[\sum_{k_1=1}^{n_1} A(k_1)\right]^2} \sum_{k_1=1}^{n_1} \frac{A^2(k_1)}{\left[\sum_{k_2=1}^{n(k_1)} A(k_1,k_2)\right]^2} \cdots \sum_{k_{s-1}=1}^{n(k_1,k_2,\ldots,k_{s-2})} \frac{A^2(k_1,k_2,\ldots,k_{s-1})}{\left[\sum_{k_s=1}^{n(k_1,k_2,\ldots,k_{s-1})} A(k_1,k_2,\ldots,k_s)\right]^2}$$
$$\times \sum_{k_s=1}^{n(k_1,k_2,\ldots,k_{s-1})} \{A^2(k_1,k_2,\ldots,k_s)\, V[\square, n_0(k_1,k_2,\ldots,k_s), n(k_1,k_2,\ldots,k_s)]\}, \quad (151{\cdot}4)$$

with the same summation notation.

152. For the optimum solution there would be either

$$\delta V + \lambda \delta T = 0 \quad (152{\cdot}1)$$

with a given value of $T$, or

$$\delta T + \mu \delta V = 0 \quad (152{\cdot}2)$$

with a given value of $V$. Either set would give the same solution, and in either case there are as many equations as there are unknowns—the variation having reference to all the unknowns. As in the uni-stage case so here also there is the possibility of the grid pattern itself varying, in which case the problem would be one of calculus of variations. But, as in the case of uni-

stage sampling, if it is assumed that the grid pattern is settled beforehand, the problem would be one of ordinary differential calculus; and in actual practice the equations would be usually independent and compatible.

153. In the foregoing set-up consideration has been given to the case of a unitary unrestricted sampling at all stages except in the last stage where the sampling is unitary configurational. If, on the other hand, the sampling is at any earlier stage unitary configurational, or at any stage zonal unrestricted or zonal configurational, then (1) the cost function, (2) the estimated total value of the variate, and (3) the variance may be written down in the same abstract form as (150·1), (151·1) and (151·2), (151·3) or (151·4) respectively, with the proviso that the nature of the sampling pattern or configuration at any stage would enter implicitly in the variance function, and might also partly influence the form of the cost function (150·1) and the estimated total value (151·1).

154. This is so far as multi-stage sampling for the uni-variate case is concerned. For multi-stage sampling for the multi-variate case an abstract set-up can be made similar to the one for multi-variate sampling. There is nothing fundamentally new in principle about this, and the case need not be considered in detail at the present stage. I have already mentioned that for both uni-variate and multi-variate uni-stage sampling several concrete forms have been obtained for cost and variance functions based on a series of field experiments conducted over a number of years. The position is less advanced so far as the multi-stage (either uni-variate or multi-variate) case is concerned, but work is proceeding, and it is hoped to be able to discuss these questions in later papers.*

## 7. Zoning and statistical controls

### *Problem of zoning*

155. I shall now briefly refer to two questions which have some connexion with the abstract solution of the optimum size-density distribution discussed in the previous section. The first is the problem of zoning. It will be noticed that throughout the discussion given above it has been tacitly assumed that the number as well as the demarcation of different zones have been settled beforehand. The underlying principle is simple. Obviously the area must be divided into zones (that is, the number as well as the shape and size of each zone must be decided) in such a way that, at any given level of cost, the final estimate would be obtained with minimum error; or alternatively, for any assigned level of error, the work would be done at minimum cost. The solution in any particular case would depend on the nature of the physical field and would involve not merely the cost and variance functions but also a knowledge of the contour levels of the variate over the whole area as well as the randomness or degree of non-randomness of different regions. In fact, the question of zoning falls under what has been called the third type of problem, namely, 'mapping surveys', further consideration of which must be postponed until certain investigations are completed.

* [*Footnote added in proof on* 24 *July* 1944.] Since writing this present paper in November 1942 a good deal of work has been done in the Calcutta Statistical Laboratory on the theory of multi-stage sampling, giving it a concrete basis which will be discussed in other papers.

*Types of error*

156. The second question is connected with the different types of error which occur in the case of large-scale sample surveys. These may be broadly divided into three distinct groups. First of all there are the fluctuations inherent in the method of random sampling which are dealt with in the theory of probability and in the theory of sampling distributions. The use of the word 'error' in this connexion is to some extent misleading, and is merely a historic remnant of the influence of the classical theory of errors on the development of modern statistical methods. In order to prevent any misunderstanding I shall consistently refer to this particular kind of so-called 'error' as 'sampling fluctuations'. In the outline of the abstract theory our attention has been confined exclusively to such sampling fluctuations.

157. Apart from sampling fluctuations, errors also arise from the fallibility of human observers. Such errors may and do arise at all stages of sample survey work: identifying the individual plots in the field, estimating the area under jute (or other crop) and making entries of crop records, measuring the area of individual plots, etc. As the number of workers is large—of the order of three or four hundred in the case of a large-scale survey like the jute census—these errors require special attention. In so far as these observational errors arise from unconscious bias they may be presumed to conform more or less to the classical theory of errors, that is, to follow at least approximately the normal distribution so that positive and negative errors would tend to cancel increasingly as the number of observations is increased. A special study has been made of this question. The method adopted was to repeat the field observations for the same region or village or individual grids by more than one observer and to compare the results. Details are given in Part III, but I may state here that in many cases this has been found to be broadly true. Although the absolute discrepancies (irrespective of the sign) were often large, positive and negative values usually occurred in equal proportion so that they tended to cancel. The algebraic discrepancy was thus much smaller, and decreased as the number of observers or the number of observations was increased.

158. In sharp contrast to 'sampling fluctuations' and 'observational errors' of the type described above, inaccuracies have also arisen from false entries or deliberate failure to carry out instructions. (As already mentioned, in the case of the jute survey the field work has to be carried out under trying weather conditions when moving about in villages is difficult. In this situation some of the field investigators put down entries by pure guess work without taking the trouble of going round the field. This risk is increased by the fact that most of the investigators are employed only for ten or twelve weeks so that there is no permanent hold on them.) The theoretical distinction between the second type of observational errors and the third type of gross inaccuracies is quite clear. Observational errors are amenable to statistical treatment. False entries and inaccuracies arising from gross negligence, on the other hand, are not amenable to statistical or probabilistic treatment. In actual practice, however, it is difficult to separate these two groups, and it is necessary to pool together the second and the third types under one common head which may be called 'recording mistakes' arising from the human factor. Thus there are two broad groups, namely, (*a*) 'sampling fluctuations', which come under the theory of statistical distribution, and (*b*) 're-

cording mistakes', which fall partly within the scope of the classical theory of errors but also contain inaccuracies due to false entries or gross negligence on the part of the investigators.

159. The margin of error of the final estimate would, of course, involve both types of errors', that is, sampling fluctuations as well as recording mistakes. One way of reducing recording mistakes would be to give systematic training to the workers and to eliminate such of them as are found to be careless or negligent in their work. Even with all possible care a certain amount of residual 'recording mistakes' is bound to persist. In large-scale sample surveys employing several hundreds of workers, it is thus practically impossible to bring the whole survey under what is usually called 'statistically controlled conditions' by eliminating all recording mistakes arising from the human factor. In this situation it becomes essential to provide statistical controls for detecting and guarding against such recording mistakes. One way of doing this would be to organize the same survey in the form of two or more interpenetrating subsamples. These subsamples may be entirely independent or may be partially linked together in suitable ways. As already mentioned, in the jute survey in 1941 we used two subsamples, a description of which is given in Part III.

160. Such a simple control may not, however, be always adequate. Assuming, for example, that there are reasons for believing that the field investigators are extremely unreliable, it may be advisable to arrange that the same set of sample grids should be surveyed either wholly or partly by two or more different sets of investigators. A comparison of the different sets of records would then disclose the magnitude of recording mistakes with complete certainty. In case more than two sets of records are available it would also be possible to identify unreliable workers. Eliminating unreliable entries it would then be possible to secure a certain portion of field data which would be comparatively free from recording mistakes.

161. The effective number of grids would no doubt be reduced by this method, and the magnitude of sampling fluctuations would increase. But the reduction in the magnitude of recording mistakes may more than compensate for the loss of information in the purely technical sense. The general principle is clear. One portion of the money must be devoted to reducing the uncertainty in the final estimate arising from sampling fluctuations. Another portion must be used for controlling and reducing or eliminating the uncertainty arising from recording mistakes which have their origin entirely in the human observers. What should be the correct proportion must, of course, be settled in each case from practical considerations. In the ultimate analysis the decision would depend on the judgement of the person responsible for preparing the plan and his assessment of the reliability of the workers, the risk of recording mistakes, and the amount of money he is prepared to spare for controlling such mistakes.

## 8. The method of contour levels

162. The approach adopted in the present paper is relevant only in dealing with fields of what have been called the non-random type. It is, therefore, of considerable practical importance to be able to decide quickly whether any particular field under study is of non-random type. The variance function or the correlation function can be used for this purpose

on lines already indicated. In the present section I shall consider a different approach which is sometimes convenient in practice.

163. *Contour levels or patches.* Consider an abstract set of $N_0$ values of $z$, and suppose that these $N_0$ values are classified into a finite number of say $c$ class intervals. The actual class ranges may be specified (which may or may not be equal in length) as $(z'_0-z'_1)$, $(z'_1-z'_2)$, ..., $(z'_{c-1}-z'_c)$. Such detailed specification of class ranges or class intervals may be briefly written as $I(c)$. In the usual way all values of $z$ lying in the different class ranges can be labelled by a suitably chosen set of values $z_1, z_2, \ldots, z_{c-1}, z_c$ where $z_t$ corresponds to the interval $(z'_{t-1}-z'_t)$ for all values of $t$ from 1 to $c$. Now consider any particular field or micro-state belonging to the above abstract distribution of $z$, that is, any particular space distribution of the $N_0$ values of $z$ in $N_0$ basic cells arranged in a definite space order. A map of the field may now be constructed by entering the appropriate value of $z$, or by painting each basic cell with an appropriate colour, using $c$ different colours to represent the $c$ different class values of $z$. A contour line then demarcates each aggregate of one or more adjoining cells giving a contour level, or 'patch'. In counting the number of patches two different methods are available: (*a*) if diagonal contact is recognized then two cells are said to be adjoining when they have either diagonal or lateral contact; (*b*) if diagonal contact is not recognized then two cells will be called adjoining when they have lateral contact but not when they have diagonal contact. In either case two cells $(Q)$ and $(Q')$ filled with identical values of $z$ or painted with the same colour will be said to belong to the same patch, if and only if cells $(Q_1), (Q_2), \ldots, (Q_k)$ can be found each filled with the same value of $z$ or painted with the same colour such that two members of the chain $(Q), (Q_1), (Q_2), \ldots, (Q_k), (Q')$ are adjoining. The number $k$ may be zero. Thus if $(Q)$ and $(Q')$ are themselves adjoining then they belong to the same patch. (It is clear that the number of patches in any sample when diagonal contact is recognized will be less than or at the most equal to the number of patches when diagonal contact is not recognized.)

164. In this way it is possible to break up any field or any space distribution of $z$ into a number of contour levels or patches. Corresponding to any abstract distribution of $N_0$ values of $z$ there is a set of $N$ (of the order of $N_0!$) associated micro-states; and for each of these $N$ micro-states there is a definite map together with a definite number of contour levels or patches as defined above. This patch number may, therefore, be treated as a characteristic of the micro-state and a frequency distribution obtained of such patch numbers over the $N$ micro-states. (It should be noted that unlike the other micro-state characteristics, e.g. variance function and correlation function, which depend essentially on the particular sampling procedure chosen, this patch number is absolutely independent of any sampling procedure and might thus be regarded as an *intrinsic* characteristic of the field and, of course, of the abstract distribution behind it.) Write $\nu_s$ as the patch number for the $s$th micro-state, and let $\bar{\nu}$ be its expectation value averaging over all possible micro-states. It may now be presumed that, under mild restrictions, the patch numbers $\nu_s$'s will cluster round this expectation value, and such heaping up would become more and more pronounced as the value of $N_0$ (and hence of $N$) is increased. (It is easy to see that this is true in the case of a linear field filled with two values of $z$ say, 'success' and 'failure', which is considered a little later.)

165. Consider now all micro-states which have patch numbers less than $\bar{\nu}$. A critical level can now be chosen, say $\nu_0$, either as a centile point or on other considerations; and all micro-states having patch numbers less than this critical value $\nu_0$ may be considered to be of non-random type. In the same way a second critical level $\nu_0'$ may be chosen (for the classification of all micro-states having patch numbers greater than $\bar{\nu}$), and all micro-states having patch numbers greater than $\nu_0'$ may also be defined to be of non-random type. All micro-states having patch numbers lying between $\nu_0$ and $\nu_0'$ ($\nu_0 < \bar{\nu} < \nu_0'$) would then be considered to be fields of random type. As in the case of the variance and the correlation functions the choice of the critical levels $\nu_0$ and $\nu_0'$ is in one sense arbitrary. But this would be ultimately decided on purely pragmatic grounds, namely, whether the proposed classification would make any material difference in the cost of operations of sample surveys by treating the micro-state (in a given situation) as being of either random or non-random type.

166. From the method of contour levels or patch numbers a decision is reached as to whether any observed field may be considered to be of a random or a non-random type at any assigned level of significance. This classification, however, has intrinsic reference to the manner in which the abstract set of values of $z$ is arranged in different class intervals. If the number of class intervals is altered the centile values will in general change. Also, even when the total number of class intervals, say $c$, of the abstract distribution of $z$ is kept the same, any alteration in the values of the end-points of the individual class intervals will in general change the centile values. For any particular way of constructing the abstract frequency distribution of $z$ the accumulated frequencies and the centile values would be, however, unique and unambiguous; and the general argument can be developed without difficulty. The effect of modifying the class ranges of the abstract-frequency distribution is, however, of considerable interest and is being studied in the Statistical Laboratory; and results will be published in due course.

167. It is further worth noting that the classification into random and non-random types has intrinsic reference to the quad, $\square$, the size of the ultimate basic cells. For example, in a given field if sixteen small basic cells are lumped together to give new unit cells of a larger size, then the values of the patch numbers and hence of their centile values may also change.

168. Speaking generally the patch number thus depends upon (*a*) the nature of the field under consideration, (*b*) the size of the basic cell, and (*c*) the abstract-frequency distribution of $z$ and the manner in which it is arranged in class intervals. The size of the basic cell $\square$, and the specifications of class ranges or intervals of the abstract-frequency distribution of $z$, can be modified at our discretion, and hence the patch number will in general be functions of these two entities. For any given field and any given abstract frequency distribution the patch number may therefore be written $\nu_s = F_s[I(c), \square]$, where $I(c)$ stands for the exact specifications of the $c$ different intervals on the basis of which the frequency distribution is constructed and $\square$ is the size of the basic cells.

169. This opens up various possibilities. A field may remain random in character even when either $\square$ or $I(c)$ is changed within certain limits. In this case the field may be called uniformly random within the appropriate limits of $\square$ and $I(c)$. A field may have random (or non-random) characteristics up to certain critical values of $\square$ and/or $I(c)$ and then become

non-random (or random) in character. In more complicated cases the character of the field may change in an irregular manner for different ranges of values of □ and of $I(c)$. The problem is intricate, but it is not necessary to enter into greater details at this stage. It is worth noting, however, that in the present problem, as in other statistical situations, the macro-properties are intimately connected with and depend in a fundamental manner on the postulated nature of the micro-structure.

*Model sampling experiments*

170. The subject is being studied on both theoretical and experimental lines in the Calcutta Statistical Laboratory, and the results will be published separately. I may, however, briefly explain the experimental procedure and refer to a few typical results. Consider any actual space distribution based on field observations. From this an abstract frequency distribution can easily be constructed using a suitable specification of intervals $I(c)$. Using pairs of random numbers $(i, j)$ or $(x, y)$ the different values of $z$ can be distributed into the different cells in a purely random manner, and a random micro-state in space thus obtained. Proceeding in this way a large number of micro-states can be acquired. It is then possible to calculate directly the average value of the number of contour levels or patch numbers, the standard deviation of the patch numbers, etc., or a study may be made on usual lines of the approximate frequency distribution of $\nu_s$ by graduation. Instead of using actual observational fields it is also possible to construct fields of various kinds in accordance with different mathematical models and study their properties.

171. One empirical result is worth mentioning. In studying natural fields it was found that the observed number of contour levels or patches was *less* than the expectation value, showing that natural fields quite often possess patterned or non-random characters and cannot be treated as random fields in the sense in which this work is being used here. On the other hand, and again speaking generally, space distributions with contour or patch numbers appreciably greater than the corresponding random expectation number usually appear to have arisen through deliberate human design, or from some kind of periodicity in the field of which a perfect example is furnished by crystal structure. But periodicity appears to be rare on the scale on which work is usually done in agricultural sampling.

*Linear patches*

172. The simplest type of a statistical field (in the sense in which this work is used in the present study) is supplied by a series of adjoining basic cells arranged in one-dimensional linear order. The time series is a familiar example. Here there is a succession of values of the statistical variate arranged in one dimension.

173. In the simplest case the values of the statistical variate $z$ are given in alternate categories which may be thought of as either 'success' or 'failure' of a certain event occurring in time, or of contours of two different colours, say black and white, occupying a series of adjoining cells arranged in linear order. In other words, this is the familiar case of samples from a binomial population with the restriction that successive events have an intrinsic linear arrangement. A run of 'success' or a run of 'failure' (or alternatively a run of black or white counters) constitute a patch as defined above. The number of patches is thus merely

the number of runs of either 'success' or 'failure'. The subject of sampling from a binomial time series has received a good deal of attention, and it is not necessary to go over the whole thing again. I shall merely refer to certain experimental studies which have the closest analogy to similar problems in two dimensions.

174. In model or experimental sampling the operational procedure must be specified in detail. In the work on space sampling three different procedures were used which may be conveniently called (*a*) free, (*b*) non-free and (*c*) partially free sampling. I shall first explain these terms.

175. Consider a long series consisting of say $N_{00}$ cells arranged in a single row and filled at random with either black or white counters. Let the proportion of black counters be $p$. Now consider the proportion of black counters in samples of $N_0$ successive cells. From these different proportions will be obtained, each of which may be considered as a sample value of $p$. When $N_{00}$ is large in comparison with $N_0$ this method of sampling may be called free sampling. A chain of adjoining cells filled with black counters may be called a black patch. Similarly, a chain of adjoining cells filled with white counters can be called a white patch. It can be shown that the expectation of the number of black patches in our sample of $N_0$ is given by $(N_0-1)\,pq+p$, whereas the expectation of the variance of this number is given by $(N_0-1)\,(1-3pq)\,pq+(1-2p^2)\,pq$. Instead of considering only the number of black patches, the total number of patches, both black and white, may be considered. The expectation of this number is $2(N_0-1)\,pq+1$ with variance equal to $4pq\{(N_0-1)\,(1-3pq)+2pq\}$.

176. In an actual experiment $N_{00}$ was taken to be 10,000, and from this $M = 100$ samples of $N_0 = 100$ adjacent cells were formed. The mean, standard deviation and the coefficient of variation of the observed number of black patches and the total number of patches in these 100 samples are compared in tables 1 and 2 with the theoretical values obtained according to the above formulae.

177. Now consider what is termed non-free samples. Consider a single sample consisting of $N_0$ successive cells, and let the fixed proportion of black counters be $p$. Now fill the cells by $pN_0$ black counters and $qN_0$ white counters purely at random. Thus in each sample of size $N_0$, the proportion of black counters is kept fixed and equal to $p$. This method of sampling may be called non-free sampling.

178. The probability of obtaining exactly $x$ black patches in a sample size $N_0$ can be easily calculated, and is given by

$$f(x) = \frac{1}{{}_{N_0}C_{pN_0}}\,({}_{qN_0+1}C_x \times {}_{pN_0-1}C_{x-1}). \qquad (178{\cdot}1)$$

The expectation of the number of black patches is in this case $N_0pq+p$, whereas the variance of this number is $pq(qN_0+1)\,(pN_0-1)/(N_0-1)$. Instead of considering the number of black patches only, consideration may be given to the total number of patches. The expectation of this number is $2N_0pq+1$ and the variance is $4N_0pq/(N_0-1)$. $M = 50$ samples of size $N_0 = 100$ were obtained in this manner. The mean, standard deviation and coefficient of variation of the observed number of black patches and the total number of patches in these 100 cells are compared with the theoretical values in tables 1 and 2.

TABLE 1. OBSERVED AND EXPECTED MEAN, STANDARD DEVIATION AND COEFFICIENT OF VARIATION OF THE NUMBER OF BLACK PATCHES

| binomial proportions | mean | | standard deviation | | coefficient of variation | |
|---|---|---|---|---|---|---|
| | observed | expected | observed | expected | observed | expected |
| (1) | (2·1) | (2·2) | (3·1) | (3·2) | (4·1) | (4·2) |
| | (*a*) free sampling: $N_{00}=10{,}000$, $N_0=100$, $M=100$ | | | | | |
| 0·075 | 6·73 | 6·95 | 2·22 | 2·35 | 32·91 | 33·77 |
| 0·175 | 14·54 | 14·47 | 2·85 | 2·87 | 19·60 | 19·83 |
| 0·275 | 19·95 | 20·02 | 2·73 | 2·85 | 13·68 | 14·22 |
| 0·375 | 23·65 | 23·58 | 1·95 | 2·66 | 8·25 | 11·26 |
| 0·475 | 25·17 | 25·17 | 2·08 | 2·52 | 8·33 | 10·01 |
| | (*b*) non-free sampling: $N_0=100$, $M=50$ | | | | | |
| 0·075 | 6·48 | 7·01 | 0·54 | 0·65 | 8·39 | 9·31 |
| 0·175 | 14·20 | 14·61 | 1·31 | 1·42 | 9·22 | 9·71 |
| 0·275 | — | — | — | — | — | — |
| 0·375 | 23·72 | 23·81 | 2·36 | 2·34 | 9·95 | 9·84 |
| 0·475 | 26·10 | 25·46 | 2·51 | 2·50 | 9·62 | 9·83 |
| | (*c*) partially free sampling: $N_{00}=400$, $N_0=100$, $M=48$ | | | | | |
| 0·075 | 6·94 | — | 2·43 | — | 34·97 | — |
| 0·175 | 14·60 | — | 2·43 | — | 18·08 | — |
| 0·275 | 20·17 | — | 2·48 | — | 12·03 | — |
| 0·375 | 23·58 | — | 2·88 | — | 12·21 | — |
| 0·475 | 23·54 | — | 2·59 | — | 10·55 | — |

TABLE 2. OBSERVED AND EXPECTED MEAN, STANDARD DEVIATION AND COEFFICIENT OF VARIATION FOR THE NUMBER OF PATCHES OF BOTH COLOURS COMBINED

| proportions | | mean | | standard deviation | | coefficient of variation | |
|---|---|---|---|---|---|---|---|
| black | white | observed | expected | observed | expected | observed | expected |
| (0·1) | (0·2) | (1·1) | (1·2) | (2·1) | (2·2) | (3·1) | (3·2) |
| | | (*a*) free sampling: $N_{00}=10{,}000$, $N_0=100$, $M=100$ | | | | | |
| 0·075 | 0·925 | 13·46 | 14·75 | 4·44 | 4·65 | 32·91 | 31·53 |
| 0·175 | 0·825 | 29·08 | 29·59 | 5·70 | 5·58 | 19·60 | 18·86 |
| 0·275 | 0·725 | 39·90 | 39·48 | 3·46 | 5·52 | 13·68 | 13·98 |
| 0·375 | 0·625 | 47·26 | 46·41 | 3·90 | 5·08 | 8·25 | 10·95 |
| 0·475 | 0·525 | 50·02 | 49·38 | 4·16 | 4·78 | 8·33 | 9·68 |
| | | (*b*) non-free sampling: $N_0=100$, $M=50$ | | | | | |
| 0·075 | 0·925 | 12·96 | 14·88 | 1·08 | 1·39 | 8·39 | 9·34 |
| 0·175 | 0·825 | 28·40 | 29·88 | 2·62 | 2·60 | 9·22 | 8·70 |
| 0·275 | 0·725 | 39·56 | 40·88 | 5·16 | 4·11 | 13·04 | 10·05 |
| 0·375 | 0·625 | 47·44 | 47·88 | 4·72 | 4·81 | 9·95 | 10·05 |
| 0·475 | 0·525 | 52·20 | 50·88 | 5·02 | 5·11 | 9·62 | 9·87 |
| | | (*c*) partially free sampling: $N_{00}=400$, $N_0=100$, $M=48$ | | | | | |
| 0·075 | 0·925 | 13·88 | — | 4·86 | — | 34·97 | — |
| 0·175 | 0·825 | 29·28 | — | 4·86 | — | 18·08 | — |
| 0·275 | 0·725 | 40·34 | — | 4·96 | — | 12·03 | — |
| 0·375 | 0·625 | 47·16 | — | 6·02 | — | 12·21 | — |
| 0·475 | 0·525 | 47·08 | — | 5·18 | — | 10·55 | — |

179. It is also possible to arrange conditions which are neither fully 'free' nor entirely 'non-free'. For example, consider a chain of $N_{00} = 400$ successive cells with a certain proportion of $p$ of black counters. Suppose samples of $N_0 = 100$ successive cells are taken out of the total field of 400 cells. It is obvious that the conditions of free sampling would not be

satisfied. On the other hand, the observed proportion of black counters would not remain absolutely constant from sample to sample, but would vary to some extent. Such samples are called 'partially free' samples. In an actual experiment twelve fields of $N_{00} = 400$ were taken (for each value of $p$) and forty-eight samples of $N_0 = 100$ obtained from each field. The mean and standard deviation of the forty-eight samples thus obtained were found. The results obtained are given in tables 1 and 2.

180. It will be noticed that the results of experimental free and non-free sampling are in quite satisfactory agreement with expected values. A comparison of tables 1 and 2 reveals that the mean values of the number of patches in 'partially free' samples are roughly of the same order as mean values in the case of 'free' and 'non-free' samples. Mean values are thus fairly steady, and may be treated as independent of conditions of sampling for all practical purposes. The standard deviations (and hence coefficients of variations) are, however, entirely different for 'free' and 'non-free' samples. In fact, for 'free' sampling the coefficient of variation decreases with increasing values of the binomial proportion, while it is practically constant in the case of 'non-free' samples. The position is something intermediate in the case of 'partially free' samples, and it is found that the coefficient of variation decreases with increasing values of $p$, but more slowly than in the case of 'free' samples.

181. The above results make it possible to test the random or non-random nature of binomial linear fields or binomial time series without any difficulty. It is simply necessary to count the number of patches of runs of 'success' and 'failure' and compare the observed value with the expected value. The choice of the standard deviation would naturally depend on the conditions of sampling, that is, on the assumption on which the investigation is based. In case of doubt it is possible of course to use the standard deviations for 'free' sampling, as these are numerically greater than corresponding standard deviations for 'non-free' sampling. If the observed number of patches is significantly less or greater than the expected value (at the assigned level of significance), then the field may be treated as non-random.

182. The present method supplies a convenient and easy test of the randomness or otherwise of a linear or time series for which the statistical variate is recorded in alternate categories. Instead of binomial fields consideration may be given to trinomial or multinomial fields in which the statistical variate $z$ is sorted or labelled into three or more different

TABLE 3. OBSERVED NUMBER OF PATCHES OF THREE DIFFERENT COLOURS IN TRINOMIAL FIELDS: NON-FREE SAMPLING. $N_0 = 400$

| proportions | $M$ | mean | standard deviation | coefficient of variation |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 0·200 | 100 | 16·13 | 0·86 | 5·36 |
| 0·300 | 100 | 21·33 | 0·94 | 4·42 |
| 0·500 | 100 | 25·56 | 1·30 | 5·07 |
| 0·250 | 50 | 18·91 | 0·91 | 4·81 |
| 0·350 | 50 | 22·97 | 1·40 | 6·07 |
| 0·400 | 50 | 23·87 | 1·18 | 4·93 |
| 0·100 | 89 | 8·93 | 0·41 | 4·57 |
| 0·200 | 89 | 16·30 | 0·92 | 5·64 |
| 0·700 | 89 | 21·17 | 1·12 | 5·28 |

categories. Some work has already been done on this problem; further work is proceeding, and the results will be published in a subsequent paper. I am giving in table 3 an illustrative example of model sampling experiments with a trinomial field.

*Two-dimensional patches*

183. Now consider a two-dimensional field, which for simplicity is supposed to be rectangular, consisting of $N_{00} = N_{01} N_{02}$ cells, there being $N_{01}$ rows each containing $N_{02}$ cells, and $N_{02}$ columns each containing $N_{01}$ cells. Each interior cell of the field has lateral contact with four cells (i.e. contact along a side) and has diagonal contact with four other cells (i.e. contact along a corner only).

184. In the simplest case the values of the statistical variate $z$ may be considered to be given in alternate categories, and each cell may be considered to be filled with a black or a white counter according as $z$ belongs to the first or the second of these categories. Let $p$ be the proportion of black counters (or successes), then $pN_{00}$ of the cells are filled with black counters and the rest are white. Out of this field may be picked out a random sample consisting of a rectangular block of $N_0 = N_1 N_2$ cells, consisting of $N_1$ rows and $N_2$ columns. The proportion of black cells in a sample of this type may be considered to be an estimate of $p$. This proportion of course varies from sample to sample. When $N_{00}$ is large in comparison with $N_0$ then this method of sampling may as before be called 'free sampling'.

185. On the other hand, any rectangular block of $N_0 = N_1 N_2$ cells may be filled by $pN_0$ black counters and $qN_0 = (1-p) N_0$ white counters in a purely random manner. A number of samples may be prepared in this manner. This process may be called 'non-free sampling'. In 'free sampling' the proportion of cells with black counters varies from sample to sample. Only its expectation is $p$. On the other hand, in 'non-free sampling' this proportion remains constant and equal to $p$.

186. Consider any sample of size $N_0 = N_1 N_2$ drawn either by 'free sampling' or 'non-free sampling'. Any method of sampling, viz. 'free' and 'non-free', can go with any method of counting the black patches, viz. 'diagonal contact recognized' and 'diagonal contact not recognized'. There are thus four possible cases. The theoretical values of the expectation and variance of the number of black patches are being investigated and results will be published later. Meanwhile a certain amount of experimental work has already been done, and the results of this work are described in the following paragraphs.

187. *Binomial field.* For comparing the results of 'free-sampling' with 'non-free sampling' the following procedure was adopted. $N_{00}$ was taken to be 10,000. From it $M = 25$ samples of size $N_0 = N_1 N_2 = 20 \times 20$ were prepared. These samples may be regarded as free samples. On the other hand, 20 non-free samples of size $20 \times 20$ were also prepared. In each case the number of black patches was counted. The mean and standard deviation of the number of black patches (multiplied by $\frac{1}{4}$, i.e. reduced to 'per hundred' basis) by both methods of counting are given in table 4.

188. In the experiment whose results are given in table 4, the sample size was kept constant, viz. $N_0 = 20 \times 20$. To find how the size of the samples affect our results, another experiment was performed in the case of free sampling only. Here $N_{00}$ was taken as 10,000, whereas the following values of $N_0$ were taken: $5 \times 5$, $8 \times 8$, $10 \times 10$, $12 \times 12$, $15 \times 15$, $20 \times 20$, 25

samples being taken for each size. The results of this experiment are given in table 5. Instead of giving the mean and the standard deviation of the number of patches actually observed, these have been multiplied by $100/N_0$. Thus what has been given is the mean and the standard deviation of the 'number of patches per hundred cells', and the results are directly comparable with those given in table 4.

TABLE 4. OBSERVED NUMBER OF BLACK PATCHES (PER HUNDRED CELLS) IN TWO-DIMENSIONAL FIELD: 'FREE' AND 'NON-FREE' SAMPLING

| | diagonal contact recognized | | | | | | diagonal contact not recognized | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | free sampling $N_{00}=10{,}000$, $N_0=400$, $M=25$ | | | non-free sampling $N_0=400$, $M=20$ | | | free sampling $N_{00}=10{,}000$, $N_0=400$, $M=25$ | | | non-free sampling $N_0=400$, $M=25$ | | |
| abstract proportion | mean | s.d. | c.v. | mean | s.d. | c.v. | mean | s.d. | c.v. | mean | s.d. | c.v. |
| (0) | (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) | (4·1) | (4·2) | (4·3) |
| 0·025 | 2·40 | 0·19 | 7·95 | 2·39 | 0·21 | 8·65 | 2·47 | 0·11 | 4·46 | 2·46 | 0·12 | 4·96 |
| 0·075 | 5·55 | 0·91 | 16·45 | 5·36 | 0·34 | 6·32 | 6·28 | 1·02 | 16·27 | 6·41 | 0·32 | 4·95 |
| 0·125 | 7·52 | 0·76 | 10·15 | 7·22 | 0·69 | 9·53 | 9·73 | 1·10 | 11·31 | 9·60 | 0·84 | 8·72 |
| 0·175 | 7·97 | 0·94 | 11·82 | 7·98 | 0·85 | 10·96 | 11·75 | 0·96 | 8·12 | 11·98 | 0·63 | 5·24 |
| 0·225 | 7·89 | 0·94 | 11·93 | 7·55 | 1·06 | 14·03 | 13·04 | 1·08 | 8·26 | 12·80 | 1·30 | 10·17 |
| 0·275 | 7·01 | 1·10 | 15·66 | 6·89 | 0·70 | 10·16 | 13·84 | 1·00 | 7·19 | 13·89 | 1·17 | 8·41 |
| 0·325 | 5·68 | 1·33 | 23·58 | 5·26 | 1·11 | 21·05 | 13·53 | 1·46 | 10·82 | 12·65 | 1·02 | 8·04 |
| 0·375 | 4·10 | 1·00 | 24·40 | 3·66 | 0·80 | 21·84 | 12·92 | 1·53 | 11·84 | 12·48 | 1·78 | 14·28 |
| 0·425 | 2·85 | 0·78 | 27·29 | 2·12 | 0·69 | 32·49 | 11·09 | 1·34 | 12·10 | 10·32 | 1·24 | 12·04 |
| 0·475 | 1·49 | 0·75 | 50·21 | 1·50 | 0·51 | 34·20 | 9·71 | 1·67 | 17·19 | 8·78 | 1·52 | 17·37 |
| 0·525 | 0·92 | 0·42 | 45·54 | 1·12 | 0·53 | 47·32 | 7·12 | 1·58 | 22·21 | 7·36 | 1·09 | 14·80 |
| 0·575 | 0·63 | 0·41 | 64·96 | 0·66 | 0·27 | 40·91 | 4·96 | 1·38 | 27·90 | 5·21 | 1·07 | 20·48 |
| 0·625 | 0·48 | 0·32 | 67·08 | 0·52 | 0·26 | 50·00 | 3·26 | 1·04 | 32·04 | 3·15 | 0·76 | 24·07 |
| 0·675 | 0·33 | 0·14 | 42·26 | 0·32 | 0·12 | 37·50 | 1·83 | 0·70 | 38·18 | 1·85 | 0·79 | 42·60 |
| 0·725 | 0·31 | 0·11 | 35·23 | 0·28 | 0·08 | 28·57 | 1·26 | 0·55 | 43·42 | 1·26 | 0·50 | 39·84 |
| 0·775 | 0·25 | 0·00 | 0·00 | 0·28 | 0·08 | 28·57 | 0·65 | 0·25 | 38·47 | 0·69 | 0·30 | 43·96 |
| 0·825 | 0·25 | 0·00 | 0·00 | 0·26 | 0·06 | 23·08 | 0·41 | 0·29 | 70·20 | 0·35 | 0·18 | 52·60 |
| 0·875 | 0·25 | 0·00 | 0·00 | 0·26 | 0·04 | 15·38 | 0·29 | 0·15 | 53·84 | 0·35 | 0·14 | 41·12 |
| 0·925 | 0·25 | 0·00 | 0·00 | 0·25 | 0·00 | 0·00 | 0·25 | 0·00 | 0·00 | 0·28 | 0·08 | 27·94 |
| 0·975 | 0·25 | 0·00 | 0·00 | 0·25 | 0·00 | 0·00 | 0·25 | 0·00 | 0·00 | 0·25 | 0·00 | 0·00 |

189. Instead of considering the results for the proportion of one colour it is more useful to consider the total number of patches for both black and white colours. This is shown in table 6 for the case in which diagonal contact is not recognized. This table suggests that as in the case of a linear field the mean value of the patches is probably roughly of the same order in both free and non-free sampling in two dimensions. This mean number rises quite steeply up to something like $p = 0{\cdot}275$, $q = 0{\cdot}725$ and then more slowly. This table also suggests that standard deviations and hence coefficients of variation do not differ very much in the case of free and non-free sampling, at least not so widely as in the case of linear fields. I am giving the above results for two-dimensional fields simply to indicate the type of studies on which work is proceeding at present on both theoretical and experimental lines in the Calcutta Statistical Laboratory.

190. *An observational field.* I may also give an example of experimental studies based on records for an actual field. For this purpose I shall consider the results of crop-cutting experiments on paddy carried out in 1940–1 in the Burdwan-Hooghly-Howrah irrigation area of about 800 sq. miles in Bengal. In this area a large number of sample cuts were located

TABLE 5. OBSERVED NUMBER OF BLACK PATCHES (PER HUNDRED CELLS) FOR DIFFERENT SIZES OF SAMPLES AND PROPORTIONS IN TWO-DIMENSIONAL FIELD: FREE SAMPLING

$N_{00} = 10{,}000$

| size $N_0$ | $M$ | proportion 'p' 0·075 mean | 0·075 s.d. | 0·075 c.v. | 0·175 mean | 0·175 s.d. | 0·175 c.v. | 0·275 mean | 0·275 s.d. | 0·275 c.v. | 0·375 mean | 0·375 s.d. | 0·375 c.v. | 0·475 mean | 0·475 s.d. | 0·475 c.v. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3·1) | (3·2) | (3·3) | (4·1) | (4·2) | (4·3) | (5·1) | (5·2) | (5·3) | (6·1) | (6·2) | (6·3) | (7·1) | (7·2) | (7·3) |
| diagonal contact recognized | | | | | | | | | | | | | | | | |
| 5× 5 | 25 | 7·04 | 3·91 | 55·57 | 9·92 | 3·12 | 31·45 | 10·40 | 4·33 | 41·61 | 10·56 | 4·94 | 46·75 | 8·80 | 3·56 | 40·45 |
| 8× 8 | 25 | 6·12 | 2·42 | 39·57 | 9·06 | 2·85 | 31·50 | 8·62 | 3·26 | 37·76 | 7·88 | 3·64 | 46·13 | 4·38 | 2·51 | 57·21 |
| 10×10 | 25 | 5·84 | 1·83 | 31·32 | 9·08 | 2·60 | 28·65 | 8·48 | 2·58 | 30·47 | 6·44 | 2·31 | 35·90 | 3·20 | 2·06 | 64·43 |
| 12×12 | 25 | 5·69 | 1·62 | 28·38 | 9·06 | 2·35 | 25·95 | 8·47 | 2·30 | 27·19 | 5·31 | 1·88 | 35·42 | 2·72 | 1·40 | 51·50 |
| 15×15 | 25 | 5·58 | 0·86 | 15·44 | 8·71 | 2·11 | 24·25 | 7·45 | 1·56 | 20·90 | 4·50 | 1·73 | 38·63 | 1·99 | 1·05 | 52·81 |
| 20×20 | 25 | 5·51 | 0·91 | 16·45 | 8·18 | 0·97 | 11·82 | 6·78 | 1·06 | 15·66 | 3·76 | 0·92 | 24·40 | 1·55 | 0·78 | 50·21 |
| diagonal contact not recognized | | | | | | | | | | | | | | | | |
| 5× 5 | 25 | 7·52 | 3·90 | 51·88 | 13·60 | 5·04 | 37·04 | 15·20 | 5·78 | 38·03 | 15·68 | 4·78 | 30·46 | 15·20 | 4·62 | 30·40 |
| 8× 8 | 25 | 7·31 | 3·21 | 43·92 | 12·88 | 3·86 | 30·00 | 15·31 | 3·83 | 24·99 | 15·44 | 3·76 | 24·33 | 12·25 | 2·33 | 19·03 |
| 10×10 | 25 | 7·00 | 2·57 | 36·65 | 12·72 | 3·27 | 25·72 | 15·16 | 2·54 | 16·78 | 14·12 | 3·14 | 22·21 | 11·00 | 2·95 | 26·78 |
| 12×12 | 25 | 6·94 | 2·20 | 31·76 | 12·94 | 2·79 | 21·56 | 14·89 | 2·64 | 17·72 | 13·81 | 2·28 | 16·48 | 11·03 | 2·07 | 18·75 |
| 15×15 | 25 | 6·60 | 1·37 | 20·74 | 12·34 | 1·99 | 16·13 | 14·49 | 1·82 | 12·58 | 13·16 | 1·62 | 12·29 | 10·06 | 1·19 | 11·78 |
| 20×20 | 25 | 6·53 | 1·06 | 16·27 | 11·85 | 0·96 | 8·12 | 13·71 | 0·99 | 7·19 | 12·75 | 1·51 | 11·84 | 9·33 | 1·60 | 17·19 |

| size $N_0$ | $M$ | proportion 'p' 0·525 mean | 0·525 s.d. | 0·525 c.v. | 0·625 mean | 0·625 s.d. | 0·625 c.v. | 0·725 mean | 0·725 s.d. | 0·725 c.v. | 0·825 mean | 0·825 s.d. | 0·825 c.v. | 0·925 mean | 0·925 s.d. | 0·925 c.v. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (8·1) | (8·2) | (8·3) | (9·1) | (9·2) | (9·3) | (10·1) | (10·2) | (10·3) | (11·1) | (11·2) | (11·3) | (12·1) | (12·2) | (12·3) |
| diagonal contact recognized | | | | | | | | | | | | | | | | |
| 5× 5 | 25 | 6·55 | 2·80 | 42·69 | 5·12 | 2·11 | 41·26 | 4·48 | 1·76 | 39·19 | 4·32 | 0·83 | 19·23 | 4·00 | — | — |
| 8× 8 | 25 | 3·88 | 1·64 | 42·23 | 2·25 | 1·09 | 48·27 | 1·69 | 0·43 | 25·70 | 1·69 | 0·43 | 25·70 | 1·56 | — | — |
| 10×10 | 25 | 2·56 | 1·26 | 49·28 | 1·52 | 0·84 | 55·58 | 1·08 | 0·28 | 25·70 | 1·04 | 0·20 | 19·61 | 1·00 | — | — |
| 12×12 | 25 | 2·92 | 0·79 | 40·90 | 1·06 | 0·38 | 36·13 | 0·78 | 0·23 | 29·71 | 0·78 | 0·23 | 29·71 | 0·69 | — | — |
| 15×15 | 25 | 1·46 | 0·73 | 50·22 | 0·76 | 0·50 | 65·98 | 0·48 | 0·12 | 25·70 | 0·46 | 0·09 | 19·61 | 0·44 | — | — |
| 20×20 | 25 | 0·94 | 0·43 | 45·54 | 0·53 | 0·36 | 67·08 | 0·26 | 0·09 | 35·23 | 0·26 | — | — | 0·25 | — | — |
| diagonal contact not recognized | | | | | | | | | | | | | | | | |
| 5× 5 | 25 | 12·00 | 5·66 | 47·15 | 9·28 | 5·92 | 63·91 | 6·72 | 3·78 | 56·26 | 5·76 | 2·70 | 46·83 | 4·32 | 1·11 | 25·70 |
| 8× 8 | 25 | 10·00 | 3·88 | 38·82 | 6·38 | 2·31 | 36·14 | 3·69 | 2·51 | 67·98 | 2·31 | 1·43 | 62·07 | 1·69 | 0·43 | 25·70 |
| 10×10 | 25 | 9·48 | 2·50 | 26·40 | 4·32 | 1·48 | 34·15 | 2·48 | 1·43 | 57·57 | 1·40 | 0·64 | 46·07 | 1·12 | 0·33 | 29·53 |
| 12×12 | 25 | 8·61 | 2·10 | 24·43 | 4·25 | 1·56 | 36·78 | 1·97 | 1·02 | 51·54 | 1·06 | 0·61 | 57·34 | 0·75 | 0·19 | 25·70 |
| 15×15 | 25 | 8·44 | 2·08 | 24·69 | 3·86 | 1·25 | 32·25 | 1·65 | 0·73 | 43·96 | 0·69 | 0·32 | 45·67 | 0·48 | 0·12 | 25·70 |
| 20×20 | 25 | 6·97 | 1·55 | 22·21 | 3·42 | 1·10 | 32·04 | 1·23 | 0·53 | 43·44 | 0·43 | 0·30 | 70·20 | 0·28 | — | — |

TABLE 6. OBSERVED NUMBER OF PATCHES OF BOTH COLOURS COMBINED IN TWO-DIMENSIONAL FIELD: DIAGONAL CONTACT NOT RECOGNIZED

| proportions | | free sampling $N_{00}=10{,}000$, $N_0=400$, $M=20$ | | | non-free sampling $N_0=400$, $M=25$ | | |
|---|---|---|---|---|---|---|---|
| black | white | mean | S.D. | C.V. | mean | S.D. | C.V. |
| (1·1) | (1·2) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) |
| 0·025 | 0·975 | 10·88 | 0·44 | 4·05 | 10·85 | 0·49 | 4·51 |
| 0·075 | 0·925 | 26·72 | 4·09 | 15·31 | 26·75 | 1·21 | 4·52 |
| 0·125 | 0·875 | 40·08 | 5·40 | 13·47 | 39·80 | 3·53 | 8·88 |
| 0·175 | 0·825 | 48·64 | 4·61 | 9·48 | 49·30 | 2·47 | 5·02 |
| 0·225 | 0·775 | 54·76 | 5·16 | 9·43 | 53·95 | 5·00 | 9·27 |
| 0·275 | 0·725 | 60·40 | 5·06 | 8·38 | 60·60 | 6·24 | 10·29 |
| 0·325 | 0·675 | 61·44 | 6·28 | 10·21 | 58·00 | 5·93 | 10·22 |
| 0·375 | 0·625 | 64·72 | 7·65 | 11·83 | 62·50 | 9·33 | 14·93 |
| 0·425 | 0·575 | 64·20 | 9·02 | 14·05 | 62·15 | 7·54 | 12·14 |
| 0·475 | 0·525 | 67·32 | 9·23 | 13·71 | 64·55 | 9·56 | 14·80 |

at random, and the crop within each sample cut was harvested, threshed and weighed in the usual manner. The whole area was divided into 800 square cells each of size 1 sq. mile, and for each cell the mean yield in lb./acre was calculated on the basis of the sample cuts falling within the cell. These mean values were then classified into five groups or five levels of yield in such a way that the histogram roughly resembled a histogram for a sample drawn from a normal population. (The proportion of mean values included in each of these yield levels were, beginning from the bottom, 4·75, 24·62, 41·88, 24·25 and 4·50 %.) These classes were then conventionally labelled 1, 2, 3, 4 and 5 respectively. The appropriate yield level was then entered in each of the square cells of size 1 sq. mile. This furnished the basic physical field or micro-state. Patches of cells (ignoring diagonal contact) were then demarcated, and the number of patches was counted for each yield level separately; the total for all levels combined was also obtained. The number of patches obtained in this way may be called the observed number for the particular real field considered here.

191. Using the same set of values of yield levels it is now possible to allot them in a random manner to the different cells with the help of a set of random numbers. The patches can again be demarcated, and the number counted for each yield level separately as well as for all levels combined for the micro-state generated by such a random process. In this way twenty-five different micro-states or space distributions were prepared and the patch number of different levels as well as for all levels combined were counted for each of these twenty-five space distributions. In counting the number of patches a note was kept of the size of the patch, that is, whether the patch was made up of one single basic cell or of 2, 3, 4, 5, 6 or more than 6 basic cells. It was then easy to calculate the mean number of patches (for each yield level and each patch size separately as well as for all levels pooled together) for the twenty-five different space distributions generated by the random process described above. These mean numbers are shown in table 7. The standard deviations as well as the coefficients of variation were also calculated and are given in the same table.

192. It was found that the mean number of patches for all yield levels taken together was 400·92 with a standard deviation of 17·52 and a coefficient of variation of 4·40 %. The actual number of patches for the observed field was 309. This was lower than the observed

TABLE 7. NUMBER OF PATCHES IN SPACE-DISTRIBUTIONS GENERATED FROM AN OBSERVATIONAL FIELD: NON-FREE SAMPLING

$N_0 = 800$; $M = 25$

| level number | pro-portion | patch size: 1 | 2 | 3 | 4 | 5 | 6 | above 6 | mean | s.d. | c.v. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| 1 | 0·0475 | 32·84 | 1·96 | 0·28 | 0·04 | — | — | — | 35·12 | 2·63 | 7·50 |
| 2 | 0·2462 | 75·60 | 19·28 | 8·28 | 5·04 | 2·58 | 1·40 | 2·04 | 114·16 | 6·78 | 5·94 |
| 3 | 0·4188 | 55·92 | 16·08 | 8·64 | 5·24 | 3·80 | 3·52 | 11·48 | 104·68 | 11·17 | 10·64 |
| 4 | 0·2425 | 75·48 | 19·32 | 8·72 | 5·12 | 2·36 | 1·52 | 1·44 | 113·96 | 7·15 | 6·27 |
| 5 | 0·0450 | 30·44 | 2·32 | 0·12 | 0·04 | 0·08 | — | — | 33·00 | 1·61 | 4·87 |
| mean | | 270·28 | 58·96 | 26·04 | 15·48 | 8·76 | 6·44 | 14·96 | 400·92 | — | — |
| standard deviation | | 21·50 | 8·27 | 5·33 | 4·03 | 3·18 | 2·53 | 2·57 | — | 7·52 | — |
| coefficient of variation | | 7·95 | 14·03 | 20·49 | 26·05 | 36·30 | 39·34 | 17·20 | — | — | 4·40 |

average based on twenty-five sets of micro-states generated by a random process. by (400·92 − 309·00 =) 91·92. The observed difference of 91·92 is actually 5·3 times the standard deviation 17·52. As a large number of basic cells ($N_0 = 800$) is being dealt with, it is probably best to treat the distribution as roughly of the normal type. On this assumption, the observed deviation is not likely to have occurred by pure chance. In other words, it is reasonable to consider the observed field to be of a non-random type, that is, to be showing systematic changes in yield levels from one part of the field to another.

193. I am giving the above numerical example merely for purposes of illustration. This method has been found convenient and not unduly laborious in practice for a rapid classification of observed fields into random and non-random types. Observed fields, studied so far, were found to be of the non-random type as judged by the patch number. In certain instances this was followed up by a study of the variance and the correlation function. In each such case it was found that the variance function differed quite appreciably from the normal type and the correlation function was also not zero.

194. In a broad way the results of classification by the method of contours have been thus experimentally found to be consistent with the classification by the method of the variance and/or the correlation function. It has been seen that there exists a one-to-one correspondence between the variance function and the correlation function (for two adjoining basic cells) which furnishes a consistent basis of classification into random and non-random types by either the variance or the correlation function. It is not likely that any such one-to-one correspondence would exist between the method of classification by contour levels and the method of the variance and/or the correlation functions. But the experimental studies carried out by us as well as general logical considerations suggest that there is some kind of broad correspondence. The matter is under investigation.

195. As already mentioned, a number of appendices are being attached to Part II of this paper. These appendices deal with various detailed calculations and numerical examples relating to the variance and correlation functions discussed in the present part. Work is proceeding in many directions and a more systematic account will be given in due course. In the meantime, the examples given here will give some idea of the nature of the problem.

APPENDIX 1. NOTE ON VARIANCE FUNCTION OF LINEAR FIELDS UNDER GRID SAMPLING

It is easy to see in a general way that the variance function in a field of random type is of the normal form. I indicated this in the paper on 'Sample Surveys' presented before the Baroda session of the Indian Science Congress in January 1942. Consider a linear field mapped by contour lines which are labelled with say $m$ consecutive natural numbers; the variance of these $m$ natural numbers arranged in any order is easily seen to be $(m^2-1)/12$, which is a constant, say $V_0$, for any given value of $m$. If $n_0$ numbers are now selected at random out of such $m$ numbers then the variance of the sum of the $n_0$ numbers is easily seen to be $n_0 V_0$; and hence the variance of the mean value of the $n_0$ numbers selected at random which is written $V[R(n_0)] = V_0/n_0$, when $n_0$ is small in comparison with $m$. When the $m$ contour levels are labelled not by $m$ consecutive numbers but by any set of $m$ numbers (some of which may be repetitions), even then the variance of the given set of $m$ numbers would be constant and may be again denoted by $V_0$. If $n_0$ contour levels are chosen at random, or what comes to the same thing, any set of $n_0$ numbers at random, out of the given set of $m$ numbers, then the variance of the sum of $n_0$ numbers would be equal to $n_0 V_0$, and hence the variance of the mean value of the $n_0$ numbers selected at random would be equal to $V_0/n_0$, that is, would be of the normal form when $n_0$ is small in comparison with $m$. Next consider grids of size say $a$. In a field of a random type the number of contour lines $n_0$ likely to fall within the grid will be simply proportional to $a$, and hence $V(a)$ will be proportional to $V[R(n_0)]$ and will vary as $1/n_0$ or as $1/a$. (This result can be immediately extended to two-dimensional fields which are statistically isotropic.) It is thus found that in a field of a random type the variance function is of the normal form. A rigorous proof has been recently worked out for an endless linear field by Mr R. C. Bose, which is given below.

*(a) Variance function for an abstract distribution under random sampling*

For a finite population of $N_0$ individuals $z(1)$, $z(2)$, ..., $z(N_0)$ with variance $V(1)$, it is assumed (without any loss of generality) that

$$\sum_{i=1}^{N_0} [z(i)] = 0. \qquad \text{A–1(1)}$$

Denote, as before, the mean of the $t$th random sample of size $n_0$ by $z_t\{R(n_0)\}$ ($t = 1, 2, ..., N''$, where $N'' = {}_{N_0}C_{n_0}$); and the variance of such a mean (over all values of $t$ from 1 to $N''$) by $V[z_t\{R(n_0)\}]$. Thus

$$z_t\{R(n_0)\} = [z(i_1)+z(i_2)+\ldots+z(i_{n_0})]/n_0, \qquad \text{A–1(2)}$$

where $(i_1, i_2, ..., i_{n_0})$ refer to individuals constituting the particular sample.

$$V = [z_t\{R(n_0)\}] = \frac{1}{N''}\sum_{t=1}^{N''}[z_t\{R(n_0)\}]^2$$

$$= \frac{1}{n_0^2}\left[\frac{n_0}{N_0}\sum_{i=1}^{N_0}\{z(i)\}^2 + \frac{2n_0(n_0-1)}{N_0(N_0-1)}\sum_{i,j=1}^{N_0} z(i)\,z(j)\right].$$

Since $N_0 V(1) = \sum_{i=1}^{N_0}(z_i^2) = -2\sum_{i,j=1}^{N_0}{}'(z_i\,z_j)$ from A–1(1), and writing $z_i$ for $z(i)$, then

$$V[z_t\{R(n_0)\}] = \frac{V(1)}{n_0}\left[1-\frac{n_0-1}{N_0-1}\right]. \qquad \text{A–1(3)}$$

When $N_0$ is large compared to $n_0$

$$V[z_t\{R(n_0)\}] = V(1)/n_0 \qquad \text{A-1(3·1)}$$

can be written, provided that $V(1)$ exists for large values of $N_0$.

*(b) Variance function for endless linear space distributions under grid sampling*

As observed earlier, corresponding to the abstract distribution there are $N$ space distributions (where $N$ is of the same order as $N_0!$). Let one such space distribution (say $s$th) be $z(s,1), z(s,2), \ldots, z(s,N_0)$, where $z(s,i)$ set is the same as the $z(i)$ set but arranged in a different sequence. Thus $z(s,i) = z(j)$, where $\begin{pmatrix} 1, & 2, & \ldots, & i, & \ldots, & N_0 \\ (s,1), & (s,2), & \ldots, & (s,i), & \ldots, & (s,N_0) \end{pmatrix}$ is some permutation of 1, 2, ..., $N_0$. Thus for the mean value of the $t$th grid in the $s$th micro-state ($t = 1, 2, \ldots, N'$, where $N' = N_0$ for an endless linear field), it follows that

$$z_{st}\{Gr(n_0)\} = [z(s,t)+z(s,t+1)+\ldots+z(s,t+n_0-1)]/n_0. \qquad \text{A-1(4·1)}$$

For this particular field the variance function would be given by

$$V[z_{st}\{Gr(n_0)\}] = \frac{1}{n_0}\sum_{t=1}^{N_0}[z_{st}\{Gr(n_0)\}]^2 = \frac{1}{n_0^2 N_0}\Big[n_0\sum_{t=1}^{N_0}\{z(s,t)\}^2+2(n_0-1)\sum_{t=1}^{N_0}\{z(s,t)\,z(s,t+1)\}$$
$$+2(n_0-2)\sum_{t=1}^{N_0}\{z(s,t)\,z(s,t+2)\}+\ldots+2\sum_{t=1}^{N_0}\{z(s,t)\,z(s,t+n_0-1)\}\Big]$$
$$= \frac{V(1)}{n_0}+\frac{1}{n_0^2 N_0}\Big[2(n_0-1)\sum_{t=1}^{N_0}\{z(s,t)\,z(s,t+1)\}+\ldots+2\sum_{t=1}^{N_0}\{z(s,t)\,z(s,t+n_0-1)\}\Big].$$

Setting $$f_{s,k} = \sum_{t=1}^{N_0}\{z(s,t)\,z(s,i+k)\}, \quad \text{and} \quad f_s = \sum_{t=1}^{N_0}\{z(s,i)\}^2,$$

then $$V[z_{st}\{Gr(n_0)\}] = \frac{V(1)}{n_0}+\frac{1}{n_0^2 N_0}[2(n_0-1)f_{s,1}+2(n_0-2)f_{s,2}+\ldots+2f_{s,n_0-1}]. \qquad \text{A-1(5·1)}$$

Now $V[z_{st}\{Gr(n_0)\}]$ $(t = 1, 2, \ldots, N_0)$ may be regarded as stochastic variables each with $N$ values corresponding to the $N$ (of the same order as $N_0!$) possible fields or micro-states. If $E_s$ denotes the mathematical expectation over the different micro-states, then it can be easily shown that

$$E_s V[z_{st}\{Gr(n_0)\}] = V[z_t\{R(n_0)\}] = \frac{V(1)}{n_0}\Big[1-\frac{n_0-1}{N_0-1}\Big] = V(N_0, n_0) \text{ say.} \qquad \text{A-1(5·2)}$$

If $B(N_0, n_0)$ denotes the dispersion of $N_0 V[z_{st}\{Gr(n_0)\}]$ over the different micro-states, that is, if

$$B(N_0, n_0) = E_s\{N_0 V[z_{st}\{Gr(n_0)\}]-N_0 V(N_0, n_0)\}^2, \qquad \text{A-1(5·3)}$$

then from the law of large numbers

$$P[|V[z_{st}\{Gr(n_0)\}]-V(N_0, n_0)| \leqq \epsilon] > 1-\frac{B(N_0, n_0)}{\epsilon^2 N_0^2} \qquad \text{A-1(5·4)}$$

for any arbitrary $\epsilon$ and for a fairly large $N_0$. Now calculate $B(N_0, n_0)$

$$B(N_0, n_0) = E_s[N_0 V[z_{st}\{Gr(n_0)\}]-N_0 V(N_0, n_0)]^2$$
$$= \frac{N_0^2}{(N_0-1)^2}\frac{n_0-1}{n_0^2}[V(1)]^2+\frac{4N_0(n_0-1)}{(N_0-1)n_0^3}V(1)\,E_s[(n_0-1)f_{s,1}+(n_0-2)f_{s,2}+\ldots+2f_{s,n_0-1}]$$
$$+\frac{1}{n_0^4}E_s[2(n_0-1)f_{s,1}+2(n_0-2)f_{s,2}+\ldots+2f_{s,n_0-1}]^2. \qquad \text{A-1(6·1)}$$

Now it is evident that for any value of $s$,

$$\sum_{k=1}^{N_0}(f_{s,k}) = \sum_{i,j=1}^{N_0}(z_i z_j) = -N_0 V(1), \qquad \text{A-1(7·1)}$$

when one passes over to the abstract distribution, a procedure which can be easily justified. Also

$$E_s[f_{s,1}] = E_s[f_{s,2}] = \ldots = E_s[f_{s,n_0-1}]. \qquad \text{A-1(7·2)}$$

Hence, for any value of $k$,

$$E_s[f_{s,k}] = -\frac{N_0}{N_0-1}V(1). \qquad \text{A-1(7·3)}$$

Therefore

$$B(N_0, n_0) = -\frac{N_0^2(n_0-1)^2}{(N_0-1)^2 n_0}[V(1)]^2 + \frac{1}{n_0^4}E_s[2(n_0-1)f_{s,1}+2(n_0-2)f_{s,2}+\ldots+2f_{s,n_0-1}]^2. \qquad \text{A-1(7·4)}$$

Now

$$E_p[f_{s,k}]^2 = \frac{1}{(N_0)!}\Big[2N_0\{(N_0-2)!\}\sum_{i,j=1}^{N_0}(z_i^2 z_j^2) + 4N_0\{(N_0-3)!\}\sum_{i,j,u=1}^{N_0}(z_i^2 z_j^2 z_u^2)$$
$$+24(N_0)(N_0-3)\{(N_0-4)!\}\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v)\Big], \quad \text{where } i\neq j\neq u\neq v,$$
$$= \frac{2}{N_0-1}\sum_{i,j=1}^{N_0}(z_i^2 z_j^2) + \frac{4}{(N_0-1)(N_0-2)}\sum_{i,j,u=1}^{N_0}(z_i^2 z_j z_u) + \frac{24}{(N_0-1)(N_0-2)}\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v).$$

But $$\sum_{i=1}^{N_0}(z_i)\sum_{i,j,u=1}^{N_0}(z_i z_j z_u) + 4\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v) = 0 \quad \text{for} \quad i\neq j\neq u\neq v.$$

Also $$\sum_{i,j=1}^{N_0}(z_i z_j)^2 = \sum_{i,j=1}^{N_0}(z_i^2 z_j^2) = -2\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v).$$

Hence $$E_s[f_{s,k}^2] = \frac{2}{N_0-1}\sum_{i,j=1}^{N_0}(z_i z_j)^2 + \frac{4N_0}{(N_0-1)(N_0-2)}\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v),$$

$$E_s[f_{s,k}f_{s,l}] = \frac{8}{(N_0-1)(N_0-2)}\sum_{i,j,u=1}^{N_0}(z_i^2 z_j z_u) + \frac{24}{(N_0-1)(N_0-2)(N_0-3)}\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v)$$
$$= \frac{-8N_0\sum_{i,j,u,v=1}^{N_0}(z_i z_j z_u z_v)}{(N_0-1)(N_0-2)(N_0-3)}.$$

Remembering that $\sum_{i=1}^{N_0}(z_i) = 0$, then $2\sum_{i=1}^{N_0}(z_i z_j) = -\sum_{i=1}^{N_0}(z_i^2) = -N_0 V(1)$, and

$$4\sum_{i=1}^{N_0}(z_i z_j z_u z_v) = 2\Big[\sum_{i,j=1}^{N_0} z_i z_j\Big]^2 - 2\sum_{i=1}^{N_0}(z_i^4) = \tfrac{1}{2}N_0^2[V(1)]^2 - N_0\mu_4, \qquad \text{A-1(8·1)}$$

where $\mu_4$ is the fourth moment of $z(i)$'s with $\Sigma z(i) = 0$, that is, of the abstract distribution. Hence remembering that $1^2+2^2+3^2+\ldots+(n_0-1)^2 = n_0(n_0-1)(2n_0-1)/6$, and also that

$$\sum_{i,j=1}^{N_0}(i\times j) = n_0(n_0-1)(n_0-2)(3n_0-1)/12,$$

then

$$B(N_0, n_0) = [V(1)]^2\Big[-\frac{(n_0-1)^2N_0^2}{n_0^2(N_0-1)^2} + \frac{2(n_0-1)(2n_0-1)N_0^2}{3n_0^3(N_0-2)} - \frac{(n_0-1)(n_0-2)(n_0-3)N_0^3}{3n_0^3(N_0-1)(N_0-2)(N_0-3)}\Big]$$
$$-\mu_4\Big[\frac{2(n_0-1)(2n_0-1)N_0^2}{3n_0(N_0-1)(N_0-2)} - \frac{2(n_0-1)(n_0-2)(3n_0-1)N_0^2}{3n_0^3(N_0-1)(N_0-2)(N_0-3)}\Big]. \qquad \text{A-1(9·1)}$$

If (large for values of $N_0$) $\mu_4$ remains finite, then $B(N_0, n_0)/N_0^2$ is of the order of $1/N_0$, and hence from A–(5·4),

$$P[|V[z_{st}\{Gr(n_0)\}] - V(N_0, n_0)| \leqq \epsilon] > 1 - \frac{B(N_0, n_0)}{\epsilon^2 N_0^2},$$

by sufficiently increasing $N_0$ it is seen that $V[z_{st}\{Gr(n_0)\}]$ stochastically converges to $V(N_0, n_0)$.

## Appendix 2. Approximate formula for correlation function

Now write:

$z(x, y)$ = statistical variate (density) at the point $(x, y)$,

$(i, j)$ = gap vector for the correlation function,

$z(x+i, y+j)$ = statistical variate at $(x+i, y+j)$,

$A$ = area over which the function is being studied,

$A(i, j)$ = effective area over which the first set of members of the two correlation series is being taken,

$\square$ = fundamental space cell for the particular field.

A bar over any quantity will denote the mean value either over the total area or the effective area as indicated by the context.

The corrected product moment of the correlation between $z(x, y)$ and $z(x+i, y+j)$ may be written as

$$\square \iint_{A(i,j)} z(x,y)\, z(x+i,y+j)\, dx\, dy - \square\, \overline{z(x,y)} \iint_{A(i,j)} z(x+i,y+j)\, dx\, dy$$

$$= \square \iint_{A(i,j)} z\left[z + \left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right) + \frac{1}{2}\left(i^2\frac{\partial^2 z}{\partial x^2} + j^2\frac{\partial^2 z}{\partial y^2} + 2ij\frac{\partial^2 z}{\partial x\, \partial y}\right) + R_3\right] dx\, dy$$

$$- \square\, \overline{z(x,y)} \iint_{A(i,j)} \left[z + \left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right) + \frac{1}{2}\left(i^2\frac{\partial^2 z}{\partial x^2} + j^2\frac{\partial^2 z}{\partial y^2} + 2ij\frac{\partial^3 z}{\partial x\, \partial y}\right) + R_3\right] dx\, dy, \quad \text{A–2(1)}$$

where $R_3$ means the remainder after the third term in Taylor's expansion. Under certain circumstances, which will be stated at the end of this note, $R_3$ can be neglected in comparison with the previous terms. By Green's theorem the raw product moment or the first term on the right-hand side of equation A–2(1) becomes

$$= \square A(i,j)\, \overline{z^2} + \tfrac{1}{2}\square \int z^2(il + jm)\, ds$$

$$+ \tfrac{1}{2}\square \int z(il+jm)\left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right) ds - \tfrac{1}{2}\square \iint_{A(i,j)} \left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right)^2 dx\, dy, \quad \text{A–2(1·1)}$$

where $ds$ is an element of arc on the boundaries of $A(i, j)$ and $l$ and $m$ are the direction cosines of the normal to $ds$ which we shall call the unit normal $\vec{n}$. Let the gap vector $(i, j)$ be called $\vec{t}$ with magnitude $t$. The raw product moment can be written (with the usual notation for scalar product) as

$$\square A(i,j)\, \overline{z^2} + \tfrac{1}{2}\square \int z^2(\vec{t}\ \vec{n})\, ds + \tfrac{1}{4}\square \int (\vec{t}\ \vec{n})\, (\vec{t}\,\mathrm{grad}\, z^2)\, ds - \tfrac{1}{2}\square \iint_{A(i,j)} (\vec{t}\,\mathrm{grad}\, z)^2\, dx\, dy. \quad \text{A–2(1·2)}$$

The correction term (for the mean) can be written in the same way as

$$\square A(i,j)\,\bar{z}^2+\square\bar{z}\int z(\vec{t}\,\vec{n})\,ds+\tfrac{1}{2}\square\bar{z}\int(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds. \qquad \text{A-2(2)}$$

Using A-2(1·2) and A-2(2) the corrected product moment reduces to

$$\square A(i,j)\,(\overline{z^2}-\bar{z}^2)+\tfrac{1}{2}\square\int(z^2-2z\bar{z})\,(\vec{t}\,\vec{n})\,ds$$
$$+\tfrac{1}{2}\square\int(z-\bar{z})\,(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds-\tfrac{1}{2}\square\iint_{A(i,j)}(\vec{t}\,\mathrm{grad}\,z)^2\,dx\,dy. \qquad \text{A-2(3)}$$

Again, the corrected sums of squares of $z(i,j)$

$$=A(i,j)\,(\overline{z^2}-\bar{z}^2). \qquad \text{A-2(4)}$$

Now consider the corrected sum of squares of $z(x+i,y+j)$. Neglecting terms coming from $R_3$ the raw sum of squares of $z(x+i,y+j)$

$$=\square^2\iint_{A(i,j)}\left[z+\left(i\frac{\partial z}{\partial x}+j\frac{\partial z}{\partial y}\right)+\frac{1}{2}\left(i^2\frac{\partial^2 z}{\partial x^2}+j^2\frac{\partial^2 z}{\partial y^2}+2ij\frac{\partial^2 z}{\partial x\,\partial y}\right)\right]^2 dx\,dy. \qquad \text{A-2(5)}$$

Using again Green's theorem and also neglecting higher order terms this reduces to

$$\square^2 A(ij)\,\bar{z}^2+\square^2\int z^2(\vec{t}\,\vec{n})\,ds+\square^2\int z(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds. \qquad \text{A-2(5·1)}$$

The correcting term for the mean neglecting higher order terms

$$=-\frac{\square^2}{A(i,j)}\left[\iint\left\{z+\left(i\frac{\partial z}{\partial x}+j\frac{\partial z}{\partial y}\right)+\frac{1}{2}\left(i^2\frac{\partial^2 z}{\partial x^2}+j^2\frac{\partial^2 z}{\partial y^2}+2ij\frac{\partial^2 z}{\partial x\,\partial y}\right)\right\}dx\,dy\right]^2$$
$$=-\frac{\square^2}{A(i,j)}\left[A^2(i,j)\,\bar{z}^2+\left\{\int z(\vec{t}\,\vec{n})\,ds\right\}^2+2A(i,j)\,\bar{z}\int z(\vec{t}\,\vec{n})\,ds+A(i,j)\,\bar{z}\int(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds\right]. \qquad \text{A-2(6)}$$

Using A-2(5·1) and A-2(6) the corrected sum of squares of $z(x+i,y+j)$ becomes

$$\square^2 A(i,j)\,(\overline{z^2}-\bar{z}^2)+\square^2\int(z^2-2z\bar{z})\,(\vec{t}\,\vec{n})\,ds$$
$$-\frac{\square^2}{A(i,j)}\left[z(\vec{t}\,\vec{n})\,ds\right]^2+\square^2\int(z-\bar{z})\,(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds. \qquad \text{A-2(7)}$$

It is clear that $\frac{1}{A(i,j)}\int z(\vec{t}\,\vec{n})\,ds$ is negligible compared to unity for moderate values of $t$, which alone are of practical interest. Hence the corrected sum of squares ultimately reduces to

$$\square^2 A(i,j)\,(\overline{z^2}-\bar{z}^2)+\square^2\int(z^2-2z\bar{z})\,(\vec{t}\,\vec{n})\,ds+\square^2\int(z-\bar{z})\,(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds. \qquad \text{A-2(7·1)}$$

Thus from A-2(3), A-2(4) and A-2(7·1) the correlation function reduces to $P/Q$, where

$$P\equiv\square A(i,j)\,(\overline{z^2}-\bar{z}^2)+\tfrac{1}{2}\square\int(z^2-2z\bar{z})\,(\vec{t}\,\vec{n})\,ds$$
$$+\tfrac{1}{2}\square\int(z-\bar{z})\,(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds-\tfrac{1}{2}\square\iint_{A(i,j)}(\vec{t}\,\mathrm{grad}\,z)^2\,dx\,dy$$

and $Q\equiv[A(i,j)\,(\overline{z^2}-\bar{z}^2)]^{\frac{1}{2}}\Big[\square^2 A(i,j)\,(\overline{z^2}-\bar{z}^2)$

$$+\square^2\int(\overline{z^2}-2z\bar{z})\,(\vec{t}\,\vec{n})\,ds+\square^2\int(z-\bar{z})\,(\vec{t}\,\vec{n})\,(\vec{t}\,\mathrm{grad}\,z)\,ds\Big]^{\frac{1}{2}}. \qquad \text{A-2(7·2)}$$

Then, after some simplification, the correlation function $\rho(i,j)$

$$= 1 - \frac{1}{2A(i,j)\,(\bar{z}^2 - \bar{z}^2)} \iint_{A(i,j)} (\vec{t}\,\mathrm{grad}\,z)^2\,dx\,dy, \qquad \text{A-2(8)}$$

$$= 1 - \frac{1}{2A(i,j)\,(\bar{z}^2 - \bar{z}^2)} \left[ i^2 \iint_{A(i,j)} \left(\frac{\partial z}{\partial x}\right)^2 dx\,dy + j^2 \iint_{A(i,j)} \left(\frac{\partial z}{\partial y}\right)^2 dx\,dy + 2ij \iint_{A(i,j)} \left(\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\right) dx\,dy \right]. \qquad \text{A-2(8·1)}$$

Neglecting higher order terms $A(i,j)$ can be replaced by $A$, and then formula A-2(8) becomes in vector notation

$$1 - \rho(i,j) = \frac{1}{2A(\bar{z}^2 - \bar{z}^2)} \iint_A (\vec{t}\,\mathrm{grad}\,z)^2\,dx\,dy, \qquad \text{A-2(9)}$$

with similar changes in A-2(8·1).

*Basic assumptions behind the approximation.* The fundamental assumption is that $R_3$ (the remainder after the third term in Taylor's expansion) can be neglected in comparison with the sum of the first three terms. This would approximately hold either when (i) the gap $t$ is small compared to what is called unity, or when (ii) the function $z(x,y)$ is on the whole so slowly changing that $R_3$ is negligible compared to the sum of the first three terms.

There are natural statistical situations (which can obviously be conceived) where (i) the gap $t$ is small. Natural situations (and of course diverse mathematical functions) are possible where the rate of variation of $z(x,y)$ is such that $z(x+i, y+j)$ can be adequately represented by

$$z(x,y) + \left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right) + \frac{1}{2!}\left(i\frac{\partial z}{\partial x} + j\frac{\partial z}{\partial y}\right)^2.$$

But in case (i) or (ii) does not hold the foregoing calculation and the formula A-2(8) based thereon would not be valid. If, for instance—as happens in some natural fields with a particular pattern or network of survey—there is (on the scale of the mesh system) a rapidly fluctuating field, then the above formula would not be valid nor would it do if the calculus of finite differences (without other suitable modifications) was applied in place of ordinary infinitesimal calculus. For such situations another formula is being worked out and will be given in a later paper.

## Appendix 3. Variance function for certain artificial fields

In this Note numerical examples will be given of the variance function obtained by using compact grids which were shifted all over the field, that is, which were made to occupy all possible locations.

(1) *Pyramidal field.* These fields are generated by taking a central cell, surrounding it by a ring of $(3^2-1^2)$ cells, that ring again by another ring of $(5^2-3^2)$ cells and so on. To the outermost ring the value zero is attached, to the next inner ring attach (1), the next inner ring (2) and so on till the innermost cell is reached. One portion of such a field (which may be called a pyramidal field) is shown in figure A-3(1); the field being symmetrical the lower portion will be simply a mirror image. Taking the origin at the central cell and attaching the value $M$ to it, it is noticed that the value for the cell $(i,j)$ is $M-|i|$ or $M-|j|$ according as $|i|>|j|$ or vice versa. This means that if any pair of random numbers $(x,y)$ is taken, the smallest of the four numbers $x$, $y$, $|2M+1-x|$ and $|2M+1-y|$ would give the value of $z$

for the cell located by $(x, y)$. Any set of random numbers may therefore be treated directly (without recourse to the map or figure) as a random sample from the field.

FIGURE A-3(1). Square pyramidal field of side 21 units

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 8 | 8 | 8 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

Now write $z_t$ for the mean value (per unit cell) of the $t$th grid, and $f(z_t)$ for the frequency values of $z_t$. For a square field of side $(2M+1)$ under a square grid of side $n$ the following expressions can then easily be written down by summing over for all possible grid values for the whole field:

$$\Sigma[f(z_t)] = (2M-n+2)^2, \qquad \text{A-3(1·1)}$$

$$n^2\Sigma[z_t f(z_t)] = \tfrac{4}{3}n^2M^3 - n^2M(\tfrac{4}{3}n^2 - 4n + 3) + (\tfrac{1}{3}n^5 - \tfrac{11}{6}n^4 + \tfrac{16}{3}n^3 - \tfrac{11}{3}n^2 - \tfrac{17}{3}n), \qquad \text{A-3(1·2)}$$

$$\begin{aligned} n^4\Sigma[z_t^2 f(z_t)] = {} & \tfrac{2}{3}n^4M^4 + M^2(-\tfrac{1}{3}n^6 + \tfrac{2}{3}n^4) + M(-\tfrac{2}{3}n^7 + \tfrac{28}{9}n^6 - \tfrac{20}{9}n^4 - \tfrac{28}{3}n^3 + \tfrac{1}{9}n^2) \\ & + (\tfrac{11}{24}n^8 - \tfrac{41}{18}n^7 + \tfrac{43}{16}n^6 + \tfrac{296}{90}n^5 + \tfrac{28}{9}n^4 - \tfrac{119}{18}n^3 - \tfrac{8}{9}n^2 - \tfrac{2}{5}n) \\ & + \tfrac{2}{9}(M-n+1)\sum_{r=1}^{n-2}[(n-r-1)^2(n-r)^2(n-r+1)^2]. \end{aligned} \qquad \text{A-3(1·3)}$$

From the above expressions the mean value and variance of $z_t$ in terms of $M$ and $n$ can easily be written down:

$$\text{Mean value of } z_t = \bar{z}_t = \frac{\Sigma[z_t f(z_t)]}{\Sigma[f(z_t)]}, \qquad \text{A-3(2)}$$

$$V(z_t) = \frac{\Sigma[z_t^2 f(z_t)]}{\Sigma[f(z_t)]} - (\bar{z}_t)^2. \qquad \text{A-3(3)}$$

For a particular value of $M$ and assigned values of $n$, that is, for fields of a particular size and for certain specific grid sizes, the mean and variance were calculated from the above formulae, and also (as an alternative) directly from the field itself by shifting the grids and noting at each position of the grid. Values obtained from the formulae and directly from the field itself (to be called respectively the expected and observed values) are given in table A-3(1). Observed and expected values of the mean agree exactly. For the variance, if all powers of $1/M$ are retained in the calculation of the expected value the agreement between

TABLE A-3(1). SQUARE PYRAMIDAL FIELD OF SIZE 21 UNITS

| size of grid | mean | | variance | | coefficient of variation | |
|---|---|---|---|---|---|---|
| | observed | expected | observed | expected | observed | expected |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1×1 | 3·02 | 3·02 | 6·09 | 5·96 | 81·8 | 81·2 |
| 2×2 | 3·32 | 3·32 | 5·45 | 5·45 | 70·2 | 70·2 |
| 3×3 | 3·62 | 3·62 | 4·80 | 4·85 | 60·6 | 60·8 |
| 4×4 | 3·88 | 3·88 | 4·15 | 4·13 | 52·5 | 52·3 |

observed and expected values is excellent for all values of $n$; but if the highest power of $1/M$ is omitted in the calculation of the expected variance, the agreement is tolerably good up to $n = 4$ but not beyond that. This, of course, is simply due to what has been called the border effect.

(2) *Repeated lattice field.* These fields are generated by taking a lattice of $(s \times s)$ cells filled serially with numbers from 1 to $s^2$ (the last cell of any row and the first cell of the next lower row being regarded as contiguous), and then repeating this lattice $p$ times both row-wise and column-wise; of course both the lattice and the field could have been taken to be rectangular, i.e. $s \times s'$ and $M \times M'$; but for purposes of illustration this is scarcely necessary. In particular, take $s = 4$ and $p = 5$, as shown in figure A-3(2).

FIGURE A-3(2). Repeated lattice field ($s=4$, $p=5$)

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 | 13 | 14 | 15 | 16 |

The algebraic expressions for mean and variance of $z_t$ (where $z_t$ as before denotes the mean per unit cell of the $t$th grid) for the general case of a square field of side $M$ (where $M = p \times s$), and a square grid of side $n$ are given below.

$$\text{Mean } z \text{ over the whole field} = 8{\cdot}5. \qquad \text{A-3(4)}$$

Variance over the whole field for different types of grids according as $n$ is of the form $(4i-3)$, $(4i-2)$ or $(4i-1)$:

Type (1): $(4i-3) \times (4i-3)$ $$\frac{85}{4}\frac{1}{n^2}. \qquad \text{A-3(5·1)}$$

Type (2): $(4i-2) \times (4i-2)$ $$\frac{34(M-n+2)}{n^2(M-n+1)}. \qquad \text{A-3(5·2)}$$

Type (3): $(4i-1) \times (4i-1)$ $$\frac{17(5M-5n+13)}{4n^2(M-n+1)}. \qquad \text{A-3(5·3)}$$

Observed values of the mean and variance (obtained by shifting the grid all over the field and noting $z$ for each position of the grid) and expected values calculated from the above formulae are given in table A-3(2); the agreement is, of course, exact.

TABLE A-3(2). REPEATED LATTICE FIELD ($s=4$, $p=5$)

| size of grid | mean | | variance | | coefficient of variation | |
|---|---|---|---|---|---|---|
| | observed | expected | observed | expected | observed | expected |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1 × 1 | 8·50 | 8·50 | 21·25 | 21·25 | 54·23 | 54·23 |
| 2 × 2 | 8·50 | 8·50 | 8·95 | 8·95 | 35·21 | 35·21 |
| 3 × 3 | 8·50 | 8·50 | 2·57 | 2·57 | 18·89 | 18·89 |
| 5 × 5 | 8·50 | 8·50 | 0·85 | 0·85 | 11·20 | 11·20 |
| 6 × 6 | 8·50 | 8·50 | 1·01 | 1·01 | 11·80 | 11·80 |

(3) *Point-charge field.* This field is constructed by taking $z = 1/(i'^2+j'^2)^{\frac{1}{2}}$, $(i', j')$ being the co-ordinates of any cell with the nuclear cell as origin. (It will be recognized that this is the potential field due to unit charge at the nucleus.) For experimental purposes use was made of square layers of cells surrounding but excluding the nucleus, taking in particular $i'$ from $-7$ to $+7$ and $j'$ from $-7$ to $+7$. The field is symmetrical in $i'$ and $j'$; and one quadrant (for values of $i'$ and $j'$ from 0 to $+7$) is shown in figure A–3(3).

FIGURE A–3(3). Point-charge field

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0·143 | 0·141 | 0·137 | 0·131 | 0·124 | 0·116 | 0·108 | 0·101 |
| 0·167 | 0·164 | 0·158 | 0·149 | 0·139 | 0·128 | 0·118 | 0·108 |
| 0·200 | 0·196 | 0·186 | 0·172 | 0·156 | 0·141 | 0·128 | 0·116 |
| 0·250 | 0·242 | 0·224 | 0·200 | 0·177 | 0·156 | 0·139 | 0·124 |
| 0·333 | 0·316 | 0·277 | 0·236 | 0·200 | 0·172 | 0·149 | 0·131 |
| 0·500 | 0·447 | 0·354 | 0·277 | 0·224 | 0·186 | 0·158 | 0·137 |
| 1·000 | 0·707 | 0·447 | 0·316 | 0·242 | 0·196 | 0·164 | 0·141 |
| | 1·000 | 0·500 | 0·333 | 0·250 | 0·200 | 0·167 | 0·143 |

Values of $z_t$ for grids of different sizes were directly obtained for all possible positions of the grid, and mean values and variances of $z_t$ are shown in table A–3(3). The observed values of variance were also graduated by an equation of the form $b/(a)^g$, where $a$ is the area of the grid; and the graduated values are shown in col. 8 of table A–3(3).

TABLE A–3(3). POINT-CHARGE FIELD

| serial number | size of grid | number of grid | mean | variance | |
|---|---|---|---|---|---|
| | | | | observed | graduated |
| (1) | (2) | (3) | (4) | (5) | (6) |
| 1 | 1×1 | 224 | 0·2187 | 0·0234 | 0·0243 |
| 2 | 2×2 | 192 | 0·2258 | 0·0136 | 0·0132 |
| 3 | 3×3 | 160 | 0·2258 | 0·0102 | 0·0092 |
| 4 | 4×4 | 128 | 0·2298 | 0·0065 | 0·0072 |
| | | | | | $\log b = 2·3852$ |
| | | | | | $g = 0·4407$ |

The average value of $g$ (over the range used here) is 0·4407, which is of the same order as values of $g$ actually observed in the case of crop surveys. It is scarcely necessary to point out that the graduating equation adopted here is not an exact specification of the variance function for the point-charge field. The same thing may also be true of observational fields —in fact, the particular graduating formula used here has a purely empirical basis and its validity rests entirely on pragmatic grounds.

## APPENDIX 4. EXAMPLES OF CORRELATION FUNCTION

(1) *Linear point-charge field.* Consider $m$ cells arranged in a straight line on either side of unit charge, then the value of the variate (here the potential) is given by $z(i') = \frac{1}{|i'|}$, where $|i'|$ denotes the absolute value of $i'$, and $i' = m, m+1, \ldots, -1, +1, +2, \ldots, +m$. The raw product moment of the two series entering in the correlation function (excluding $i' = -i$) is given by

$$\sum_{i'=-m}^{m-2}\left[\frac{1}{|i'|}\cdot\frac{1}{|i'+i|}\right] = \frac{2}{i^2}+\frac{4}{i}\sum_{i'=1}^{i-1}\left[\frac{1}{i'}\right]-\frac{2}{i}\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'}\right]. \qquad \text{A–4(1)}$$

The sum of the first series entering into $\rho(i)$ is equal to the sum of the second series and is given by

$$2\sum_{i'=1}^{m}\left[\frac{1}{i'}\right]-\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'}\right]-\frac{1}{i}. \qquad \text{A-4(2)}$$

Hence the corrected product moment may be written down as

$$P\equiv\frac{2}{i^2}+\frac{4}{i}\sum_{i'=1}^{i-1}\left[\frac{1}{i'}\right]-\frac{2}{i}\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'}\right]-\frac{1}{(2m-i-1)}\left[2\sum_{i'=1}^{m}\left[\frac{1}{i'}\right]-\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'}\right]-\frac{1}{i}\right]^2. \qquad \text{A-4(3)}$$

The corrected sum of squares for the first series is also equal to the corrected sum of squares of the second series and is equal to

$$Q\equiv 2\sum_{i'=1}^{m}\left[\frac{1}{i'^2}\right]-\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'^2}\right]-\frac{1}{(2m-i-1)}\left[2\sum_{i'=1}^{m}\left[\frac{1}{i'}\right]-\sum_{i'=m-i+1}^{m}\left[\frac{1}{i'}\right]-\frac{1}{i}\right]^2-\frac{1}{i^2}. \qquad \text{A-4(4)}$$

Hence $\rho(i)=P/Q$, where $P$ is given by A–4(3) and $Q$ by A–4(4). For $m=100$, and $i=1,2,3,4,5$ and 6, then $\rho(i)=0{\cdot}7602$, $0{\cdot}7300$, $0{\cdot}5948$, $0{\cdot}4906$, $0{\cdot}4127$, $0{\cdot}3522$ respectively.

(2) *Two-dimensional point-charge field.* In this case the value of the variate is given by $z(i',j')=(i'^2+j'^2)^{-1}$ (excluding the origin or nucleus containing the charge). Two fields have been used: (*a*) one with $i'$ varying from $-7$ to $+7$, and $j'$ from 5 to 19, and (*b*) with $i'$ ranging from $-25$ to $+25$ and $j'=25$; and the value of $\rho(i,0)$ has been calculated for various values of $i$.

In such cases if the border effect is taken into consideration then

$$1-\rho(i,0)=\tfrac{1}{2}i^2\frac{\sum_{A(i,0)}(\Delta_x z)^2}{V[z(x,y)]\,V[z(x+i,y+j)]}.$$

In numerical work $\sum_{A}(\Delta_x z)^2$ and $V(z)$ are calculated for the whole area $A$, and then for each gap is calculated simply the relevant border terms and combining with the above the other terms are obtained. Observed values of the correlation directly calculated from the field and expected values calculated from equation A–4(5) are shown in table A–4(1).

TABLE A–4(1). POINT-CHARGE FIELD: VALUES OF $\rho(i)$

| | $i'=-7$ to $+7$; $j'=5$ to 19 | | | $i'=-25$ to $+25$; $j'=25$ | | |
|---|---|---|---|---|---|---|
| gap $i$ | number of grids | observed | expected | number of grids | observed | expected |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1 | 210 | 0·9904 | 0·9904 | 50 | 0·9909 | 0·9909 |
| 2 | 195 | 0·9644 | 0·9639 | 49 | 0·9624 | 0·9623 |
| 3 | 180 | 0·9269 | 0·9247 | 48 | 0·9127 | 0·9118 |

(3) *Paddy-yield field.* In the section on contour levels I have described an observed field relating to the yield of paddy in a certain tract of about 800 sq. miles in Bengal. The whole area was divided into 800 square cells each of area 1 sq. mile. A map was then prepared by entering the mean yield of paddy in each such basic cell. This furnished an observed or physical field or space distribution. Coefficients of correlation were calculated between the yields of paired cells ($i\times 1$ sq. mile) at different distances apart along (1) rows, (2) columns, (3) diagonals, and are given in table A–4(2). The values given are all averaged out values

calculated respectively from all rows, columns, and diagonals, by first changing over to Fisher's $z$-transformation for correlation, weighting by the reciprocal of variance, averaging out and finally transforming back to $\rho$.

TABLE A–4(2). PADDY-YIELD FIELD

| distances in miles | rows | | columns | | north-east | | south-east | |
|---|---|---|---|---|---|---|---|---|
| | $N'$ | $r$ | $N'$ | $r$ | $N'$ | $r$ | $N'$ | $r$ |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 1 | 970 | 0·332 | 937 | 0·310 | — | — | — | — |
| $\sqrt{2}$ | — | — | — | — | 916 | 0·2636 | 929 | 0·2636 |
| 2 | 944 | 0·149 | 878 | 0·090 | — | — | — | — |
| $2\sqrt{2}$ | — | — | — | — | 831 | 0·0500 | 869 | 0·1293 |
| 3 | 918 | 0·009 | 819 | −0·029 | — | — | — | — |
| $3\sqrt{2}$ | — | — | — | — | 756 | −0·0500 | 812 | 0·0699 |
| 4 | 892 | 0·034 | 760 | −0·041 | — | — | — | — |
| $4\sqrt{2}$ | — | — | — | — | 688 | −0·0200 | 754 | 0·0599 |

## APPENDIX 5. CALCULATION OF VARIANCE FUNCTION FROM CORRELATION FUNCTION

Let $Gr\{(i, 1)\}$ denote a compact grid of area $(i \times 1)$ with $i$ cells row-wise and 1 cell column-wise, and $Gr\{(1, i)\}$ a compact grid of area $(1 \times i)$ with 1 cell row-wise and $i$ cells column-wise. The corresponding variance functions will be denoted by $V[Gr(i, 1)]$, and $V[Gr(1, i)]$. Further, let $\rho(i, 0)$ and $\rho(0, j)$ denote correlation functions for gaps $(i, 0)$ and $(0, j)$ respectively. Then from equation (106·1) the following values are obtained:

$$4V[Gr(2,1)] = V(1)\,[2+2\rho(1,0)],$$
$$9V[Gr(3,1)] = V(1)\,[3+4\rho(1,0)+2\rho(2,0)],$$
$$16V[Gr(4,1)] = V(1)\,[4+6\rho(1,0)+4\rho(2,0)+2\rho(3,0)],$$
$$25V[Gr(5,1)] = V(1)\,[5+8\rho(1,0)+6\rho(2,0)+4\rho(3,0)+2\rho(4,0)].$$

Four similar equations are also obtained by interchanging 1 with 2, 3, 4, 5 respectively. From the observed values of $\rho$'s the $V$'s can be calculated by the above formulae. These will be called the expected values. The observed value of $V$'s were also obtained which were directly calculated by shifting the grids of appropriate size over the whole field. The two sets of values are given in table A–5(1) in cols. (2) and (3) respectively.

TABLE A–5(1). VARIANCE FOR GRIDS OF SIZE

| $n_0$ | calculated | observed | graduated |
|---|---|---|---|
| (1) | (2) | (3) | (4) |
| | rows $(n_0 \times 1)$ | | |
| 1 | 97·2 | 97·2 | 100·0 |
| 2 | 64·7 | 64·9 | 62·5 |
| 3 | 50·0 | 49·3 | 47·3 |
| 4 | 40·1 | 39·1 | 39·0 |
| 5 | 33·5 | 32·1 | 33·6 |
| | columns $(1 \times n_0)$ | | |
| 1 | 91·4 | 91·4 | 93·0 |
| 2 | 60·8 | 60·5 | 59·4 |
| 3 | 44·8 | 46·8 | 45·7 |
| 4 | 35·2 | 37·9 | 38·0 |
| 5 | 28·6 | 32·1 | 32·9 |

The observed values of the variance have been also graduated by a formula of the type $b/(n_0)^g$ (where $n_0$ is the number of cells constituting the compact grid). The values of $g$ were 0·680 and 0·646 for row-wise and column-wise grids respectively. The graduated values are shown in col. (4) of the same table. The agreement between expected, observed and graduated values is sufficiently close for all practical purposes.

### Appendix 6. Border effect

Consider a field of rectangular shape consisting of $N_0 = (N_i \times N_j)$ basic cells arranged in the form of $N_i$ columns each consisting of $N_j$ cells, and $N_j$ rows each consisting of $N_i$ cells. Consider a compact grid of rectangular shape made up of $i \times j$ cells. In the case of overlapping grids the total number of cells which will be repeated full $(i \times j)$ times in the sampling will be $(N_i - 2i + 2)(N_j - 2j + 2)$. Such cells will be called interior cells.

In the border there will be $2N_i(j-1) + 2N_j(i-1) - 4(i-1)(j-1)$ cells which will be repeated less than $(i \times j)$ times. The actual number will depend on the distance from the boundary. For example, the four cells at the four corners of the field will be included in the grid only once each.

If it is assumed that the average value of $z$ over border cells is not greater than the average value of $z$ over interior cells, then the bias introduced by the border effect will not exceed

$$[2N_i(j-1) + 2N_j(i-1) - 4(i-1)(j-1)]/N_i N_j. \qquad \text{A-6(1)}$$

When $N_i = N_j = N$ say, and $N_0 = N^2$, then (3) will reduce to

$$[2N(i+j-2) - 4(i-1)(j-1)]/N^2. \qquad \text{A-6(2)}$$

In addition, if $i = j$, the order of the bias is

$$\frac{1}{N^2}[4N(i-1) - 4(i-1)^2]' = \frac{4(i-1)}{N}\left[1 - \frac{i-1}{N}\right]. \qquad \text{A-6(3)}$$

In the case of the jute survey in Bengal the size of the field covered by the grid sampling is at least $N^2 = 40{,}000$ sq. miles; also the maximum size of the grid is say

$$40 \text{ acres} = 1/16 \text{ sq. mile} = i^2.$$

Thus $(i^2/N^2) = 1/16 \times 40{,}000$, and $(i/N) = 1/800$. The border effect is of the order of 1/200. For a single zone of size say 144 sq. miles and a grid of size 40 acres, $i/N = 1/48$, and the bias may be so large as 1/12. But even here with the number of grids of the order of 100 the sampling fluctuation is many times larger. It is thus seen that with grids of the size actually used in practice the border effect is not of importance.

## PART III. APPLICATION TO ESTIMATION OF AREA UNDER CROPS

196. In this section I shall give an account of recent work on the jute crop in Bengal to illustrate the application of the abstract theory to the concrete problem of estimating the area under crops. It will be convenient if I first explain the general procedure of the sample survey of crop acreage as it is being done at present in Bengal.

## 1. General description of sample survey of area under jute

197. *Planning of the survey.* The first thing necessary is to prepare a detailed plan of the survey. On the statistical side the most important things are: (*a*) demarcating the whole area under a suitable number of zones; (*b*) deciding the form of the survey for purposes of controlling recording mistakes, that is, deciding whether the survey is to be done in the form of two or more interpenetrating or partially or completely overlapping subsamples; and (*c*) deciding the optimum size and density of grids in each zone subject to the given conditions. That is, if the total amount of money which can be spent on the work is fixed, to settle the distribution in such a way as to secure a final estimate with the lowest possible margin of error. On the other hand, if the margin of error is fixed, to prepare the plan in order that the work might be done at a minimum cost. I have already discussed the abstract solution of the problem in general terms, and I shall give numerical illustrations in a later section. At this stage we take it for granted that the zones have been suitably demarcated, and the form of the survey and the optimum size-density distribution of grids have been settled in a satisfactory manner.

198. *Location of sample units or grids.* The preparatory stage of the work can now begin with the actual location of the allotted number of grids at random over each zone. For this purpose large-scale cadastral survey maps of villages are used (usually on the scale of 16 in. = 1 mile, and printed on sheets of about $30 \times 22$ in.), showing the boundary and the revenue serial number of each plot.

199. The procedure for locating the grids strictly at random on these maps was studied experimentally, and it was found that this could be done most conveniently with the help of a simple apparatus which has been called a co-ordinatograph. This is essentially a frame of rectangular or square shape made of four pieces of wood joined at right angles at four corners. Two steel or wooden scales (which are called 'frame scales') are mounted on two opposite sides of the co-ordinatograph, and a bridge (which is simply a piece of wood of about the same width as the framework) rests and slides freely on the two sides containing the scales. A third scale (which may be called the 'bridge scale') is mounted on the bridge on which slides freely a cursor carrying a pencil. The movement of the bridge is thus parallel to the two sides of the frame, while the movement of the pencil on the bridge is in a perpendicular direction. The procedure is simple. The co-ordinatograph is placed on a map. A pair of random numbers, say $x$ and $y$, are now taken, each consisting of two or three digits. The sliding bridge (which has an arrow notched at either end) is adjusted at the reading corresponding to the random number $x$ on the frame scales. The sliding pencil on the bridge is adjusted at the point $y$ on the bridge scale, and the pencil is lowered to mark the point on the map. The co-ordinates of this point (with an arbitrary origin) are thus clearly $x$ and $y$, and as this is a pair of random numbers, the position of the marked point is also purely random. The co-ordinatograph can be easily constructed and is inexpensive. There is nothing to go out of order, and even untrained workers can learn to use it in the course of a day or two. The output is good; in the Calcutta Statistical Laboratory it is possible to mark roughly from 40 to 50 points per hour.

200. The required number of points having been located at random, grids of the appropriate size are then stamped on the maps with the help of rubber stamps or engraved brass

plates. (Each grid is stamped in a specified manner, for example, using the sample point as the south-west corner point of the square.) The grids are serially numbered as they are stamped on the village maps. A list is then prepared of the revenue serial numbers of plots which lie either entirely within, or partly within and partly outside, each grid. This is called the 'field list of plots'; usually a single slip of paper is used for each grid for convenience of subsequent tabulation work. Each such sheet contains the serial number of the grid, its size in acres, the name of the village and other geographical particulars as well as the serial number of the zone for convenience of identification; and necessary blank space is left for field entries including date and signature of investigator. A duplicate list is also prepared at the same time for use in the Statistical Laboratory in the manner explained a little later.

201. *Field survey.* The field workers are supplied with cadastral village maps and field lists of grids, and are sent out to the field in units consisting of from five to eight investigators under the charge of one field inspector; within each zone the work is distributed among the field investigators by the inspector in charge. Each investigator carries his own bundle of maps, takes up his residence in a village and goes out to the field with the village map, identifies the plots (of which the revenue serial numbers are given in the field lists), makes an actual physical examination of the crops on each of these plots, and enters in these lists the estimated proportion of land under jute (and/or other crops) in these plots. As soon as the work in the neighbourhood is completed the investigator shifts his camp to another village. Usually two or three days are spent at each village. But halts of even six or seven days or more are made on occasions, while camps are sometimes shifted even after a single night. The inspector in charge of each field unit keeps in touch with the investigators by moving from one camp to another. He is responsible for the accuracy of the work done by his unit, and is required to check a certain proportion of the sample grids from time to time. Printed instructions in detail are supplied to investigators and inspectors who are also given some preliminary training.

202. *Area extraction.* In the meantime the area of each individual plot included in the field lists is compiled in the Statistical Laboratory. At one time these used to be copied from records in the local revenue offices, but this was found troublesome and unreliable. At the present time photographic scales or graticules are used which are simply lantern plates on which a graph paper is photographed on a scale such that each small square is 0·01 acre. This photographic scale is placed over the map, and the area of that portion of each plot which is included within the boundaries of the grid is directly read off and entered in a column provided for this purpose in the Laboratory copy of the field list.

203. *Tabulation.* The investigators hand over the field lists completed by them during the week to the inspector in charge, who then despatches the whole lot directly to the Statistical Laboratory. In these field lists the estimates of the proportion of land under jute (or other crops) are given in the customary Indian unit of *annas* which is based on the standard Indian coin of 1 rupee consisting of 16 annas. (One *anna* is thus $\frac{1}{16}$ of the whole unit, which in this case is the total area of each individual plot.) As soon as the field lists are received in the Laboratory these *anna* estimates are converted into decimal figures and entered in a column provided for this purpose in the Laboratory copy of the field lists. The area of that portion of the plot which is included within the grid is already entered in an adjoining column;

multiplying these two figures the actual area in acres under jute (or other crops) in each plot is entered in a third column provided for this purpose. Adding the figures in this column, the total area in acres under jute for each sample unit is obtained. The subsequent work of tabulation is straightforward. Subtotals are built up for suitable administrative units and by zones, and regional and provincial figures are finally compiled from these subtotals.

204. *Statistical analysis.* This so far as the direct estimate of the area under jute is concerned. A good deal of other statistical work has to be done, of course, for comparison of different subsamples, calculation of errors, application of appropriate statistical tests, study of cost and variance functions, etc. But these are not included in the present study.

205. *An optical method.* The procedure is thus quite simple and straightforward. The only difficulty lies in the large scale of operations. In the preparatory stage, for example, something like 102,000 different sheets of maps have to be handled for the marking of grids, preparation of field lists and the measurement of area of individual plots. The possibilities of using an optical method were therefore explored, and after some experimentation it was found that this could be done with the help of comparatively simple apparatus. The basic idea is simple. The maps would be photographed on positive cinema films through a glass plate carrying a photographic scale of square cells each of, say, size 0·01 acre. When this photograph is projected on a screen not only the boundary and revenue serial numbers of plots are shown, but also a superimposed mesh showing squares of size, say, 0·01 acre. As the projection of the map would be on a fairly large scale it would be possible to measure the area of individual plots directly by visual examination. Another picture would be superimposed on the screen which would show simply a number of sample grids of the required size. It would be thus possible, first of all, to note down the serial number of plots included wholly or partly within each sample unit or grid, and secondly also to note down the area of each individual plot falling within the grid by visual examination. In this method the location of grids at random, the listing of plots falling within the grids, and the measurement of the area of such plots would be all done practically at one single operation. The method worked satisfactorily in the Laboratory but could not be adopted on a large scale owing to a shortage of cinema films in India under war conditions.

## 2. Recording mistakes

206. I shall next describe certain studies of what I have called recording mistakes in a previous section. These have their origin in the human observers (as distinguished from sampling fluctuations), and may be broadly divided into two groups. The first group consists of 'chance' errors of classical theory amenable to statistical treatment. The second group consists of inaccuracies arising from false entries or deliberately negligent work which cannot in any way be brought under a probabilistic scheme. I shall give a few typical results for certain kinds of mistakes which were studied experimentally.

207. In 1937, 1938 and 1939, extensive material in the form of duplicate or triplicate records of the crop grown on individual plots in the same geographical area surveyed by different sets of investigators was collected for this purpose. The nature and volume of the material is shown in table 8. Grids were, of course, of various sizes, while the area of individual plots also varied widely.

TABLE 8. VOLUME OF FIELD DATA COLLECTED BETWEEN 1937 AND 1941

| year | random sample unit | number | acres | plots |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 1937 | police stations | 2 | 79,498 | 283,565 |
| | villages | 20 | 8,345 | 28,417 |
| | grids | 1,488 | 7,440 | 27,238 |
| | plots | 14,159 | 3,974 | 14,159 |
| | total | | 99,257 | 353,379 |
| 1938 | police stations | 8 | 264,835 | 478,639 |
| | grids | 7,888 | 16,471 | 78,711 |
| | plots | 2,540 | 1,406 | 2,540 |
| | total | | 282,712 | 559,890 |
| 1939 | police stations | 7 | 429,431 | 912,067 |
| | grids | 12,311 | 50,184 | 215,676 |
| | plots | 45,530 | 17,757 | 45,530 |
| | total | | 497,372 | 1,173,273 |

208. *Field survey.* Mistakes were committed in locating or identifying individual plots and in making field observations relating to the proportion of land under jute or other crops in individual plots. A detailed comparison of the entries was made plot by plot for all plots for which more than one set of records were available. Records of complete enumeration where available, and in other cases records based on the larger unit (such as villages in preference to grids, grids in preference to random plots), were adopted as the standard. This is, of course, purely a matter of convention, as there is no reason for believing that the complete enumeration of any other particular set was more reliable than other records. For studying discrepancies it is, however, immaterial which set of records is used as the standard.

209. If a plot is shown under jute in the standard set of records but under some other head in the duplicate set under comparison, then this was considered to be a positive error (requiring negative correction) for jute. On the other hand, if a plot is shown under jute in a duplicate set but under some other head in the standard set, then this was considered to be a negative error. In this way discrepancies were classified as positive or negative, and were studied separately for jute and a number of other crops.

210. Several things became clear in the course of these studies. The number of discrepancies at the stage of the field survey was very high. The absolute sum of both positive and negative discrepancies gives a convenient picture of the accuracy of the field work. In 1937, for example, it was found that for a group of villages taken together the absolute discrepancy was as high as 58 % of the actual area under jute. The positive and negative discrepancies, however, occurred to a large extent in equal proportions, so that they tended to cancel out. The algebraic sum of discrepancies was thus much smaller; and in the case of the same group of villages considered above the total algebraic discrepancy was of the order of only 5 %. A good proportion of the recording mistakes at this stage were thus amenable to statistical treatment.

211. This is satisfactory, but clear evidence was also found of inaccuracies which could not have arisen excepting from false entries or gross negligence. The magnitude of the discrepancy, both algebraic and absolute, also varied widely from one investigator to another. A part of this no doubt may be ascribed to differences in the 'personal equation

of the individual workers, but detailed comparison and scrutiny of the material left little doubt that some of the investigators were dishonest in their field work.

212. *Crop record.* In the next stage of the work, namely, preparation of crop records, a similar detailed comparison was carried out. Here the absolute discrepancy was something of the order of 9 or 10 %, while the algebraic discrepancy was less than 2 %. On the whole inaccuracies at this stage were far smaller in magnitude than the mistakes which occurred at the stage of field survey. This is, of course, just what may be expected in view of the fact that the field survey has had to be carried out under far more difficult conditions.

213. *Area measurement.* A detailed study was also made of errors occurring at the stage of copying the area of individual plots from revenue records which were kept in the district headquarters and were thus scattered all over the province. This arrangement was difficult to supervise, and large mistakes were detected. From 1940, therefore, the practice was adopted of measuring the area of individual plots directly in the Laboratory with the help of photographic scales. The absolute discrepancy by this method is of the order of 2 %, and the algebraic discrepancy appreciably below 1 %.

214. *Border effect.* In a sample survey on a large scale there were naturally many other sources of error, some of which were studied experimentally. For example, there was the question of the border effect. It was found that there was persistent over-estimating in working with units of very small size. In the case of field survey the obvious explanation is that the investigator has a tendency to include rather than to exclude plants or land which stand near the boundary line or perimeter of the grid. This boundary effect naturally becomes less and less important as the size of the grid is increased. In crop-cutting work on jute it was found, for example, that mean values for all the characters studied (such as number of plants per acre, weight of green plants, weight of dry fibre) were much higher for sample units of small size, so that it was not at all safe to work with cuts of a size less than say 25 sq. ft. In the case of the area survey it was generally found inadvisable to work with grids of size less than about 1 acre.

215. The above studies revealed the great importance of controlling and eliminating as far as possible the mistakes which occurred at the stage of the field survey. This is why from the very beginning special attention was given to the need of building up a reliable human agency. In 1937 there was not a single trained field worker, and only about half a dozen computers. Whatever training was possible was given in the very short time at the disposition of the Laboratory, and this had to be repeated every year, as the scheme was sanctioned from year to year. The whole of the field staff was recruited for only three of four months, and continuity of employment could not be guaranteed. A large number, especially the abler men, left after one season and did not come back, so that work had to be carried on with a large proportion of untrained men each year. On the statistical side, however, it became possible to train up and give more or less continuous employment to a good proportion of computers by employing them on other projects.

216. Various attempts were made to improve the efficiency of the survey by proper selection of workers. With this purpose in view, a study was made of variations in output (and in certain instances also of mistakes) of individual workers. Without entering into details I may mention one or two typical results. The average output of all workers for any

particular type of work was adopted as the basis for comparison, and the index of output of each individual worker was found by dividing his actual output by the adopted standard and multiplying by 100. Individual variations were enormous. For example, in the field survey the index number in 1937 varied from 48 for a particular worker to 146 for another investigator; the output of the quickest worker was thus three times as large as the output of the slowest. The coefficient of variation fluctuated roughly between 25 and 40 %, depending on the particular type ofwork in the case of the field survey.

217. The position was much the same in the statistical portion of the work. The coefficient of variation in output among individual workers was roughly about 20 or 25 % in the case of simple operations like listing and comparison of entries, and of the order of 30 or 35 % in the case of work involving computations. The question of accuracy was also studied to some extent by comparing the proportion of mistakes made by different workers for different types of work. Here also large variations were found.

218. In 1940 and 1941 arrangements were made from the Indian Statistical Institute to hold examinations for the award of certificates for computing work and field survey. I am making a passing reference to these things to indicate the kind of methods which were adopted from time to time for selecting suitable workers with a view to improving the general efficiency of the survey.

### 3. Variance function

219. I shall now give a brief account of the experimental studies of the variance function with special reference to the work on the jute crop in Bengal. Two different methods were adopted for this purpose. One was to use sampling units or grids of various sizes in the field survey itself, and to compare the variances for different sizes of grids. This method is direct, but expensive and subject to errors of field observations which vary widely from investigator to investigator.

220. The second method is to prepare in the first instance a complete inventory or enumeration of the crop grown on each plot for a suitable area. On the basis of this material it is then possible to carry out model sampling experiments in the Laboratory with various sizes of grids. Apart from errors of the original complete inventory all subsequent results are necessarily free from errors of field observations. Also, as large tracts of land are surveyed by the same investigator, results for grids of various sizes within each such area is strictly comparable. This method is much cheaper than the direct method; once a complete enumeration is made it is possible to carry out practically an unlimited series of model sampling experiments in the laboratory at a comparatively small cost.

221. Both direct and model sampling methods were used to the fullest extent possible within budgetary limits. Most of the material was collected during 1937–41. Use was also made of the results of a sample survey carried out in 1935 with grids of size 40 acres located at random with uniform density of one grid per 16 sq. miles; the total number of grids collected in 1935 was 2844, and the total area covered about 45,000 sq. miles.

222. As already mentioned in Part II it was found that satisfactory graduation could be obtained with a variance function of the form $V(a) = b/(a)^{\varepsilon}$, where $V(a)$ is the variance of the mean value of $z$ based on grids each of size $a$ measured in conventional units, say acres,

and $b$ and $g$ are constants which may vary from zone to zone. (As there was no direct method of demarcating what was called the basic cells of the field, it was necessary to use some conventional unit of measurement like the acre.) If $n_0$ is the number of basic cells, and $\square$ the area of each basic cell measured in conventional units, say, acres, then $a = n_0\square$ or $n_0 = a/\square$. Now for a truly binomial or random field the variance would be given by $pq/n_0$, where $p$ is the proportion of basic cells under jute, and $q = (1-p)$. For a non-random field the above formula may be modified to $pq/(n_0)^g = pq(\square)^g/(a)^g$. Comparing with $b/(a)^g$, then $b = pq(\square)^g$, or $\square = (b/pq)^{1/g}$.

223. It will be noticed that the form of the variance function adopted here involves two constants, $b$ and $g$. Using material for two or more different sizes of grids it is thus possible to estimate the values of both $b$ and $g$ by usual methods of least square fitting. From a knowledge of $b$ and $g$ it is then possible to make an estimate of the value of the quad $\square$, that is, of the size of the ultimate basic cells.

224. For the actual purpose of obtaining the optimum size and density of grids it is, however, not necessary to determine the size of the basic cell, as numerical solutions can all be worked out in terms of the fitted constants $b$ and $g$. I shall not therefore enter into a detailed discussion of the different estimates of the size of the basic cells which were obtained in the course of the work. I may, however, mention that the actual value of the quad $\square$ varies from zone to zone. The average value over the whole area of about 20,000 sq. miles surveyed in 1940 was 0·034 acre, and over the whole area of about 48,000 sq. miles under grid survey in 1941 was 0·057 acre. This variation from year to year may be partly due to the difference in the coverage of the two surveys, and also partly to variations in the proportion of land sown with jute and changes in condition of cultivation. On the whole it was found that the average value of the quad or size of the basic cells was something of the order of say 0·04 or 0·05 acre. However, in Bengal the average size of individual plots is about 0·4 acre; the size of the basic cells thus appears to be something of the order of a tenth of the average size of plots. In revenue work the smallest fraction of which it is considered possible to take cognizance is '1 anna' (that is, 1 anna out of a rupee of 16 annas) or $\frac{1}{16}$th part of a plot; this supplies a rough unit of $\frac{1}{16}$th of 0·4 acre, or 0·025 acre. Estimate of the quad or the ultimate unit of jute cultivation is about 0·04 or 0·05, or say twice the lowest unit recognized in revenue or agricultural practice. This is physically plausible and indicates that the graduating equation has a real basis in fact.

225. However, to return to the more important quantity, namely, the numerical value of the constant $g$. As already mentioned, extensive material relating to the proportion of land under jute based on sampling units of various sizes ranging from individual plots to 40-acre grids was collected in 1935 and from 1937 to 1941 by both field survey and model sampling experiments. A detailed discussion of this material will be out of place here; I shall merely give a brief summary.

226. The results of certain model sampling experiments carried out in 1938 and 1939 are given in table 9, in which col. (0) shows the size of grids in acres. The name of each zone is given at the head of each section, together with the year in which complete inventory was made; below the name of each zone is given the average value of $p$ (the proportion of land under jute), and the estimated value of $g$ obtained by a logarithmic fit by least squares.

Under each zone is shown the number of grids, the observed variances for different sizes of grids ($a$) and corresponding expected values calculated from graduating equations of the form $V(a) = b/(a)^g$, where $a$ is the size of the grid in acres.

TABLE 9. VARIANCE FUNCTION FOR AREA UNDER JUTE IN BENGAL BASED ON MODEL SAMPLING EXPERIMENTS

| zone | Iswarganj (1938) $p=0{\cdot}3712$, $g=0{\cdot}3377$ | | | Tejgaon (1938) $p=0{\cdot}0930$, $g=0{\cdot}2924$ | | | Laksam (1938) $p=0{\cdot}0242$, $g=0{\cdot}4006$ | | | Nandail (1939) $p=0{\cdot}3634$, $g=0{\cdot}2812$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| size of grids in acres ($a$) | number of grids | variance observed | variance expected | number of grids | variance observed | variance expected | number of grids | variance observed | variance expected | number of grids | variance observed | variance expected |
| (0) | (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) | (4·1) | (4·2) | (4·3) |
| 1·00 | 1756 | 1120 | 1085 | 399 | 394 | 379 | 266 | 81 | 81 | 266 | 903 | 916 |
| 2·25 | 1377 | 813 | 825 | 300 | 299 | 299 | 200 | 74 | 59 | 200 | 748 | 730 |
| 4·00 | 1161 | 659 | 680 | 249 | 257 | 252 | 166 | 34 | 47 | 166 | 605 | 621 |
| 6·25 | 951 | 577 | 585 | 201 | 220 | 222 | 134 | 33 | 39 | 134 | 602 | 547 |
| 9·00 | 832 | 505 | 517 | 174 | 170 | 199 | 116 | 40 | 34 | 116 | 469 | 494 |
| 12·25 | 732 | 454 | 466 | 150 | 140 | 182 | 100 | 31 | 30 | 100 | 409 | 453 |
| 16·00 | 618 | 419 | 426 | 126 | 180 | 168 | 84 | 25 | 27 | 84 | 466 | 420 |
| 25·00 | 476 | 398 | 366 | 99 | 184 | 148 | 66 | 36 | 22 | 66 | 366 | 371 |
| 36·00 | 365 | 342 | 324 | 75 | 166 | 133 | 50 | 14 | 19 | 50 | 328 | 335 |
| $\chi^2$ | | 4·40 | | | 12·21 | | | 26·53 | | | 1·90 | |
| probability | | 0·73 | | | 0·09 | | | 0·00 | | | 0·96 | |

| zone | Pirgacha (1939) $p=0{\cdot}2778$, $g=0{\cdot}4076$ | | | Palashbari (1939) $p=0{\cdot}3344$, $g=0{\cdot}3485$ | | | Belkuchi (1939) $p=0{\cdot}1650$, $g=0{\cdot}2918$ | | |
|---|---|---|---|---|---|---|---|---|---|
| size of grids in acres ($a$) | number of grids | variance observed | variance expected | number of grids | variance observed | variance expected | number of grids | variance observed | variance expected |
| (0) | (5·1) | (5·2) | (5·3) | (6·1) | (6·2) | (6·3) | (7·1) | (7·2) | (7·3) |
| 1·00 | 266 | 751 | 677 | 266 | 638 | 627 | 266 | 277 | 287 |
| 2·25 | 200 | 407 | 486 | 200 | 448 | 473 | 200 | 229 | 226 |
| 4·00 | 166 | 368 | 385 | 166 | 420 | 387 | 166 | 206 | 191 |
| 6·25 | 134 | 376 | 321 | 134 | 309 | 331 | 134 | 163 | 168 |
| 9·00 | 116 | 216 | 276 | 116 | 272 | 292 | 116 | 186 | 150 |
| 12·25 | 100 | 257 | 244 | 100 | 301 | 262 | 100 | 117 | 138 |
| 16·00 | 84 | 258 | 219 | 84 | 243 | 239 | 84 | 105 | 128 |
| 25·00 | 66 | 177 | 182 | 66 | 177 | 204 | 66 | 127 | 112 |
| 36·00 | 50 | 172 | 157 | 50 | 198 | 180 | 50 | 99 | 101 |
| $\chi^2$ | | 10·32 | | | 3·35 | | | 6·74 | |
| probability | | 0·17 | | | 0·85 | | | 0·46 | |

N.B. Variance figures have been multiplied by 10,000.

227. Using the large sample expression for the sampling error of the expected variance it is possible to compare the goodness of fit by the above graduation. The observed values of $\chi^2$ together with the probability of occurrence (the degrees of freedom are seven in each case) are given at the bottom of the table. With one exception (Laksam 1938), the probability is quite high. The above results show that a graduating equation of the form $V(a) = b/(a)^g$ is quite satisfactory; and secondly, that the value of $g$ is much smaller than unity.

228. The same form of graduating equation was found to give fairly satisfactory results in the case of material collected directly on the field. It is not necessary to enter into details, but, as one would expect, the agreement between observed and graduated values is somewhat less close than in the case of model sampling experiments.

229. I shall also briefly state without detailed discussion certain results relating to the $g$-parameter which emerged in the course of the present studies. It was found that the value

of $g$ was always less than unity, but even in the same season it varied from zone to zone. This is in keeping with the view adopted in the present paper, namely, that the value of $g$ is determined by the degree of correlation existing between the proportion of land under jute in neighbouring plots. This correlation would be naturally determined by local conditions or habits of jute cultivation. Variations in such local conditions may, therefore, be expected to cause variations in the correlation and hence in the values of $g$. It may be noted that $\rho$ the intra-grid correlation and the $g$-parameter are connected by the relation

$$\rho(n_0-1) = (n_0)^{1-g}-1.$$

230. Secondly, it was found that, speaking generally, the value of $g$ decreased with increasing values of $p$ or the proportion of land under jute. This is also just what is to be expected. When the proportion of land under jute is small, plots sown with jute would be widely scattered; in this case the correlation between adjoining or neighbouring plots would be necessarily small and hence the value of $g$ would be high. On the other hand, when the proportion of land under jute is high, plots under jute would lie close together; the value of the correlation between neighbouring plots would be, therefore, high, and the value of $g$ would be small. This is what would happen under usual conditions of cultivation but not in all circumstances.

231. For the material as a whole the linear regression of $g$ on $p$ is given by

$$g = 0{\cdot}4748-0{\cdot}6686(p). \qquad (190{\cdot}1)$$

The parabolic fit is given by

$$g = 0{\cdot}4954-1{\cdot}0389(p)+0{\cdot}9642(p^2). \qquad (190{\cdot}2)$$

232. The actual fitting was made on twenty-seven different values of $g$. I am showing grouped values in table 10 in which col. (1) gives the years of survey, col. (2) the number of individual grids and col. (3) the number of values of $g$ on which group values are based, col. (4) the mid-point of the range of $p$ (proportion of land under jute), col. (5) the observed average value of $g$, col. (6) the expected value of $g$ calculated from the linear equation, and col. (7) the corresponding expected values from the parabolic equation.

TABLE 10. OBSERVED AND GRADUATED VALUES OF '$g$'

| graduations based on: years of survey | number of: individual grids | number of: values of '$g$' total = 27 | proportion of land under jute ($p$) mid-point of range | values of '$g$': observed | values of '$g$': graduated by: linear regression | values of '$g$': graduated by: quadratic regression |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1935, 1937–41 | 23,683 | 6 | 0·025 | 0·517 | 0·470 | 0·458 |
| 1935, 1938–41 | 23,300 | 5 | 0·075 | 0·366 | 0·423 | 0·425 |
| 1935, 1939–41 | 20,527 | 4 | 0·125 | 0·365 | 0·381 | 0·391 |
| 1939–41 | 9,901 | 3 | 0·175 | 0·300 | 0·343 | 0·358 |
| 1939–40 | 8,789 | 2 | 0·225 | 0·319 | 0·310 | 0·324 |
| 1939–40 | 6,611 | 2 | 0·275 | 0·317 | 0·283 | 0·291 |
| 1939–40 | 6,886 | 2 | 0·325 | 0·346 | 0·260 | 0·258 |
| 1940 | 12,660 | 2 | 0·375 | 0·249 | 0·241 | 0·224 |
| 1940 | 2,034 | 1 | 0·425 | 0·120 | 0·228 | 0·191 |

233. In examining the values of $g$ given in table 10 it is necessary to keep several things in mind. In 1935 and 1941 the values were based on practically the whole province; in other years the survey area differed widely. Secondly, the regression equations are based on material for different zones in the same year as well as for the same zone in different years. Finally, the field material was collected by sometimes entirely and sometimes widely different sets of investigators. In spite of such wide sources of variation one thing is certain, namely, that the value of $g$ is substantially less than unity. Although the agreement between observed and expected values cannot be considered satisfactory the general tendency is also clear, namely, for $g$ to decrease with increasing values of $p$. (As no work was performed with very low values of $p$ or intensities of cultivation of less than say 2 %, the above graduation cannot be expected to supply any information in the neighbourhood of $p = 0$ where the theoretical value of $g$ should be unity under usual conditions of cultivation.)

234. It has been seen that in the same year the value of $g$ fluctuates from one region to another. This may be ascribed partly to differences in the proportion of land under jute, and partly to what may be broadly called differences in local conditions of jute cultivation. The variations in the value of $g$ which occur in the same jute season may be called static fluctuations under stationary conditions of cultivations.

235. Now consider the variation from year to year or dynamic fluctuations. In the same region this is mainly due to changes in $p$, the proportion of land under jute. When the proportion under jute increases the value of $g$ would naturally fall; and when the proportion of land under jute decreases the value of $g$ would rise. This was fully corroborated in 1940 and 1941. Sowings were exceptionally heavy in 1940, and the average value of $g$ was so low as 0·21. In 1941 sowings under jute were restricted by the Government of Bengal to something like one-third of the area under jute in 1940, and the average value of $g$ increased from 0·21 in 1940 to 0·46 in 1941.

236. It is conceivable or even likely that the dynamic and static fluctuations in $g$ should be different, especially under any scheme of regulation. When the proportion of land under jute is reduced to say half by the operation of economic and other natural cases such reduction would take place in accordance with the physical habits of jute cultivation. On the other hand, when such reduction is made by Government decree the shrinkage would occur practically uniformly over the whole area. The reduction in the correlation function (and hence the increase in the value of $g$) may therefore differ in these two cases. Until knowledge of the subject is expanded further, however, it is not possible to say anything more in this connexion. It is only when reliable data for the same area becomes available for a number of years that it will be possible to study the question of dynamic fluctuations or time shifts in the value of $g$ in an adequate manner.

237. It will not serve any useful purpose to go into further details about the variance function. I may mention, however, that good a deal of material has been secured relating to various crops like jute, paddy, wheat and sugar cane. I am giving one illustrative set of figures in table 11 for the yield of jute based on crop-cutting experiments carried out in a number of districts in Bengal in 1940–1. The variates studied were: (i) the number of plants per acre, (ii) weight of green plants per acre, and (iii) weight of dry fibre per acre; and eight different sizes of grids (i.e. sample cuts) were used, namely, 1, 4, 9, 16, 48, 64, 144 and 256

sq. ft. In each case the graduating equation was of the form $b/(a)^g$; and the values of the $g$-parameter are shown here. It will be noticed that the value of $g$ is appreciably less than unity in every case, so that these fields are of a non-random type in the sense defined in the

TABLE 11. VALUES OF $g$-PARAMETER; JUTE IN BENGAL, 1940–1

| district | number of plants per acre | weight of green plants per acre | weight of dry fibre per acre |
|---|---|---|---|
| (1) | (2) | (3) | (4) |
| Mymensingh | 0·403 | 0·407 | 0·380 |
| Rangpur | 0·304 | 0·352 | 0·518 |
| Tipperah | 0·391 | 0·384 | 0·538 |

present paper. From an abstract point of view this immediately establishes the need of using the theory of configurational sampling. For this purpose it has been found that an equation of the form $b/(a)^g$ is convenient for graduating the variance function; but this is purely a matter of empirical observation. It is quite possible that field material relating to other variates would require different graduating equations. The approach adopted in the present paper does not, however, depend upon any particular form of the variance function, and it is not necessary to attach any special importance to the particular form used. The real point is one of convenience and usefulness in practice.

## 4. COST FUNCTION

238. One of the objects of the exploratory surveys was to study the relative cost of operations in working with grids of different size and density. Arrangements were therefore made for all field workers to keep a diary in which entries were made every day of the time spent in different kinds of work from which the whole day of 24 hr. was split up into a number of components. Along with the time records, entries were also made showing the size and number of grids surveyed every day and also of other special type of work, if any.

239. A large amount of material was collected every year and was subjected to various types of graduation. It was finally decided to work with four components:

(i) Journey time required for moving from one camp to another, from camp to field and back, and from grid to grid.

(ii) Time spent in identifying and examining the plots comprising each grid, and in making necessary entries in the field book. This is collectively called 'enumeration time'.

(iii) Time spent on miscellaneous work such as receiving instructions, meeting cultivators and village officials, preparing copies of records, interviews with inspectors, etc.

(iv) Time spent on all other indirect purposes, that is, the balance of the time which added to the above three items would make up to 24 hr. for each day.

240. The above components were settled after a good deal of experimental work. For example, in 1939 the time in moving from camp to field and from grid to grid, that is, the time spent in the neighbourhood of each camp (which we later called the time for 'small journey'), had been included under 'enumeration time'. In 1940 the time for 'small journey' was treated as a separate component. But after some further experimentation we decided to pool together the time required for both big and small journey under one head as

'journey time'. In the diary separate entries were kept of all time lost on account of leave, sickness or enforced stoppage of work owing to bad weather conditions or other causes. This was convenient for administrative purposes, but the time lost in this way was finally pooled with the time required for taking food, sleep, rest, etc., that is, for all indirect purposes.

241. After much experimentation with the graduation of the primary material it was found that a comparatively simple specification was adequate for all practical purposes. Without entering into a detailed discussion of the material, which will be out of place here, I shall describe the broad tendencies in the case of each of the four components.

242. *Time spent for indirect purposes.* It was found that the time spent for indirect purposes was practically a constant fraction (nearly two-thirds) of the whole day and was independent of both size and density of grids. Relevant material is shown for the three years 1939–40–41 in tables 12·1 and 12·2. Under each year the first column shows the size of grids in acres or the density in number of grids per square mile; the second column the number of grids on which the time records are based; and the third column the time spent on indirect work expressed as a percentage of total time, that is, 24 hr.

TABLE 12·1. PERCENTAGE TIME REQUIRED FOR INDIRECT PURPOSES BY SIZE OF GRIDS IN ACRES

| 1939 | | | 1940 | | | 1941 | | |
|---|---|---|---|---|---|---|---|---|
| size in acre | number of grids | percentage time | size in acre | number of grids | percentage time | size in acre | number of grids | percentage time |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) |
| 1·0 | 7,652 | 67·4 | 1·00 | 882 | 67·1 | 1·0 | 10,435 | 65·9 |
| 4·0 | 2,238 | 71·0 | 2·25 | 1,223 | 65·8 | 2·0 | 7,908 | 66·9 |
| 9·0 | 1,432 | 67·6 | 4·00 | 16,499 | 66·0 | 2·5 | 7,201 | 67·7 |
| 16·0 | 475 | 67·8 | 6·25 | 972 | 66·2 | 3·0 | 13,378 | 66·3 |
| | | | 9·00 | 1,096 | 61·9 | 4·0 | 11,537 | 68·6 |
| | | | | | | 6·0 | 2,157 | 68·9 |
| | | | | | | 9·0 | 595 | 68·6 |
| all sizes | 11,797 | 68·0 | | 20,672 | 65·3 | | 53,211 | 67·3 |

TABLE 12·2. PERCENTAGE TIME REQUIRED FOR INDIRECT PURPOSES BY DENSITY OF GRIDS IN NUMBER PER SQUARE MILE

| 1939 | | | 1940 | | | 1941 | | |
|---|---|---|---|---|---|---|---|---|
| density per sq. mile | number of grids | percentage time | density per sq. mile | number of grids | percentage time | density per sq. mile | number of grids | percentage time |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) |
| 0·5 | 324 | 69·8 | 0·548 | 747 | 76·7 | 0·11 | 595 | 68·6 |
| 1·0 | 1,048 | 68·9 | 0·759 | 1,256 | 69·2 | 0·33 | 2,157 | 68·9 |
| 2·0 | 2,476 | 69·6 | 0·920 | 1,025 | 65·0 | 0·66 | 11,537 | 68·6 |
| 3·0 | 1,263 | 66·6 | 1·224 | 1,799 | 58·2 | 0·88 | 13,378 | 66·3 |
| 4·0 | 1,904 | 65·7 | 1·306 | 3,098 | 64·3 | 1·10 | 7,201 | 67·7 |
| 6·0 | 2,478 | 66·2 | 1·314 | 1,373 | 62·4 | 1·32 | 7,908 | 66·9 |
| 8·0 | 2,304 | 66·5 | 1·464 | 8,139 | 64·0 | 1·65 | 10,435 | 65·9 |
| | | | 1·513 | 3,235 | 62·9 | | | |
| all densities | 11,797 | 68·0 | | 20,672 | 65·3 | | 53,211 | 67·3 |

243. From tables 12·1 and 12·2 it is quite clear that the proportion of time spent in indirect purposes is independent of both size and density of grids. The actual percentage varied to some extent from year to year and was 68·0 % in 1939, 65·3 % in 1940 and 67·3 % in 1941.

A part of this variation may be ascribed to sampling fluctuations; it is also likely that the field investigators differed to some extent in the amount of time they spent on indirect purposes in different years. On the whole, about two-thirds of the day is spent on indirect purposes; that is, roughly 8 hr. per day are available on an average for direct field work.

244. *Time required for miscellaneous work.* The time required for miscellaneous work also appears to be fairly steady, i.e. independent of either size or density of grids and of the order of about 7 % or a little over an hour and a half per day. Relevant material for 1939 and 1940 is shown in tables 13·1 and 13·2, in which the arrangement is similar to that in tables 12·1 and 12·2.

TABLE 13·1. PERCENTAGE TIME REQUIRED FOR MISCELLANEOUS WORK BY SIZE OF GRIDS IN ACRE

| 1939 | | | 1940 | | | 1941 | | |
|---|---|---|---|---|---|---|---|---|
| size in acre | number of grids | percentage time | size in acre | number of grids | percentage time | size in acre | number of grids | percentage time |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) |
| 1·0 | 7,652 | 8·7 | 1·00 | 882 | 5·9 | 1·0 | 10,435 | 6·8 |
| 4·0 | 2,238 | 7·5 | 2·25 | 1,223 | 7·5 | 2·0 | 7,908 | 6·2 |
| 9·0 | 1,432 | 7·3 | 4·00 | 16,499 | 6·6 | 2·5 | 7,201 | 7·2 |
| 16·0 | 475 | 7·7 | 6·25 | 972 | 7·7 | 3·0 | 13,378 | 7·5 |
| | | | 9·00 | 1,096 | 8·8 | 4·0 | 11,537 | 6·2 |
| | | | | | | 6·0 | 2,157 | 7·6 |
| | | | | | | 9·0 | 595 | 7·7 |
| all sizes | 11,797 | 7·7 | | 20,672 | 7·2 | | 53,211 | 6·9 |

TABLE 13·2. PERCENTAGE TIME REQUIRED FOR MISCELLANEOUS WORK BY DENSITY OF GRIDS IN NUMBER PER SQUARE MILE

| 1939 | | | 1940 | | | 1941 | | |
|---|---|---|---|---|---|---|---|---|
| density per sq. mile | number of grids | percentage time | density per sq. mile | number of grids | percentage time | density per sq. mile | number of grids | percentage time |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) | (3·1) | (3·2) | (3·3) |
| 0·5 | 324 | 9·4 | 0·548 | 747 | 5·3 | 0·11 | 595 | 7·7 |
| 1·0 | 1,048 | 8·6 | 0·759 | 1,256 | 6·0 | 0·33 | 2,157 | 7·6 |
| 2·0 | 2,476 | 6·8 | 0·920 | 1,025 | 7·7 | 0·66 | 11,537 | 6·2 |
| 3·0 | 1,263 | 5·7 | 1·224 | 1,799 | 7·3 | 0·88 | 13,378 | 7·5 |
| 4·0 | 1,904 | 9·4 | 1·306 | 3,098 | 7·6 | 1·10 | 7,201 | 7·2 |
| 6·0 | 2,478 | 7·1 | 1·314 | 1,373 | 9·8 | 1·32 | 7,908 | 6·2 |
| 8·0 | 2,304 | 8·6 | 1·464 | 8,139 | 6·5 | 1·65 | 10,435 | 6·8 |
| | | | 1·513 | 3,225 | 7·9 | | | |
| all densities | 11,797 | 7·7 | | 20,672 | 7·2 | | 53,211 | 6·9 |

245. *Enumeration time.* The enumeration time covers the work of identifying the plots constituting the grid, examining the grids and making necessary field entries. Relevant material is given in table 14. Under each year the first column shows the density of grids, the second column the total number of grids, and the third column the time required for enumeration work in hours per grid. In 1939 this time decreased appreciably with increasing density of grids. This was due to the fact that one portion which we have called the time for 'small journey', namely, the time required for moving from one grid to another, had been included under enumeration time. Naturally this time decreased with increasing density of grids which led to the enumeration time in hours per grid to fall appreciably with increasing

densities in 1939. In 1940 the whole of the time required for 'small journey' was, however, separated from enumeration time and was included in journey time; and the enumeration time became independent of density of grids.

TABLE 14. ENUMERATION TIME IN HOURS PER GRID BY DENSITY OF GRIDS

| 1939 | | | 1940 | | |
|---|---|---|---|---|---|
| density of grids | number of grids | time in hours per grid | density of grids | number of grids | time in hours per grid |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) |
| 0·5 | 324 | 1·52 | 0·548 | 747 | 1·04 |
| 1·0 | 1,048 | 1·25 | 0·759 | 1,256 | 0·98 |
| 2·0 | 2,476 | 0·87 | 0·920 | 1,025 | 0·98 |
| 3·0 | 1,263 | 0·87 | 1·224 | 1,799 | 1·10 |
| 4·0 | 1,904 | 0·55 | 1·306 | 3,098 | 0·96 |
| 6·0 | 2,478 | 0·54 | 1·314 | 1,373 | 0·68 |
| 8·0 | 2,304 | 0·48 | 1·464 | 8,139 | 1·00 |
| | | | 1·513 | 3,235 | 1·18 |
| all densities | 11,797 | 0·76 | | 20,672 | 1·01 |

246. *Journey time.* Time required for moving from one camp to another, from camp to field, and from one grid to another or back from field to camp should be independent of the size of grids. This is corroborated by the field material given in table 15, in which under each year are shown figures relating to the size of grids in acres, total number of grids used, and the observed average number of hours per square mile required for journey purposes.

TABLE 15. JOURNEY TIME IN HOURS PER SQUARE MILE BY SIZE OF GRIDS

| 1939 | | | 1940 | | |
|---|---|---|---|---|---|
| size of grids | number of grids | time in hours per sq. mile | size of grids | number of grids | time in hours per sq. mile |
| (1·1) | (1·2) | (1·3) | (2·1) | (2·2) | (2·3) |
| 1·0 | 7,652 | 2·73 | 1·00 | 882 | 2·02 |
| 4·0 | 2,238 | 2·00 | 2·25 | 1,223 | 1·92 |
| 9·0 | 1,432 | 3·02 | 4·00 | 16,499 | 2·00 |
| 16·0 | 475 | 2·33 | 6·25 | 972 | 1·86 |
| 36·0 | 344 | 2·06 | 9·00 | 1,096 | 2·32 |
| all sizes | 12,141 | 2·60 | | 20,672 | 2·01 |

247. *Enumeration and journey time.* We may now consider the relative importance of journey and enumeration time. Relevant material is shown in table 16 in which cols. (1) and (2) give the year and the extent of the survey, and col. (3) the total number of hours spent

TABLE 16. TIME FOR 'JOURNEY' and 'ENUMERATION'

| year of survey | area surveyed in sq. mile | total number of man-hours for all purposes | number of man-hours spent for | | | percentage of time spent for | |
|---|---|---|---|---|---|---|---|
| | | | journey | enumeration | total | journey | enumeration |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 1938 | 414 | 152,640 | 12,974 | 27,475 | 40,449 | 31·9 | 68·1 |
| 1939 | 2,563 | 382,320 | 53,525 | 45,878 | 99,403 | 53·8 | 46·2 |
| 1940 | 20,533 | 459,360 | 78,091 | 50,530 | 128,621 | 60·7 | 39·3 |
| 1941 | 59,199 | 677,520 | 128,051 | 46,749 | 174,800 | 73·4 | 26·6 |
| all years | 82,709 | 1,671,840 | 272,641 | 170,632 | 443,273 | | |

on field survey by investigators for all (including indirect) purposes. The number of man-hours spent in each year on 'journey' and 'enumeration' and for both combined are shown in cols. (4), (5) and (6) respectively. The respective percentages are given in cols (7) and (8).

248. In large-scale sample surveys the time required for moving from grid to grid is much larger than the time required for actual enumeration work. In the full-scale survey in 1941 the journey time was nearly three times larger than the enumeration time. In earlier years in the exploratory stage the proportion was naturally smaller, as care was taken to carry out various types of enumeration work on comparatively small tracts of land for reasons of economy. As the extent or coverage of the survey increased the proportion of 'journey time' also increased. It is scarcely necessary to stress the need for taking journey time into consideration in the planning of large-scale sample surveys.

249. *Graduation of enumeration time.* Although enumeration time is independent of the density of grids it should of course increase with their size. This is fully corroborated by the field records. After a good deal of experimentation it was found that a simple linear fit is fairly satisfactory. The constant term was permitted to vary from year to year, but the coefficient of the linear term was kept the same for all three years 1939, 1940 and 1941. In this way a graduating equation of the form

$$t(e) = c_0' + 0{\cdot}0793(a), \qquad (249{\cdot}1)$$

was obtained, where $t(e)$ is the enumeration time in hours per grid of size $(a)$ in acres, and the value of the constant $c_0' = 0{\cdot}4620$, $0{\cdot}6826$ and $0{\cdot}5199$ for the material collected in 1939, 1940 and 1941 respectively. From the comparatively concordant values of the constant term it may reasonably be inferred that the graduation is fairly satisfactory.

250. The goodness of fit was tested by analysis of variance shown in table 17. In fact, linear equations were fitted separately for the material collected in 1939, 1940 and 1941, and obtained respectively the values $0{\cdot}0661$, $0{\cdot}0672$ and $0{\cdot}1009$ for the linear coefficient. The improvement due to such separate fitting was tested, but as can be seen from table 17 was not significant. The use of a common value, namely, $0{\cdot}0793$, is thus fully justified.

TABLE 17. GOODNESS OF FIT: ANALYSIS OF VARIANCE DUE TO DIFFERENT FITTINGS

| | enumeration | | | | journey | | | |
|---|---|---|---|---|---|---|---|---|
| sources of variation | D.F. | sum of squares | variance | ratio | D.F. | sum of squares | variance | ratio |
| (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| between years | 2 | 1192 | — | — | 2 | 143 | — | — |
| due to joint fit | 1 | 1949 | 1949·0 | 116·2 | 2 | 7,775 | 3887·5 | 341·0 |
| improvement | 2 | 35 | 17·5 | 1·0 | 4 | 1,794 | 448·5 | 39·2 |
| due to yearly fit | 3 | 1984 | 661·3 | 39·4 | 6 | 9,569 | 1594·8 | 139·8 |
| residual | 52 | 872 | 16·7 | — | 47 | 536 | 11·4 | — |
| within years | 55 | 2856 | — | — | 53 | 10,105 | — | — |
| total | 57 | 4048 | | | 55 | 10,248 | | |

251. Observed values of the enumeration time together with corresponding values graduated by equation (249·1) are shown in table 18·1, in which under each year the size (in acres) of grids and the total number used are given in the first two columns, and the observed and expected values of the enumeration time in hours per grid in the third and fourth columns respectively.

TABLE 18·1. ENUMERATION HOURS PER GRID BY SIZE OF GRIDS IN ACRES

| 1939 | | | | 1940 | | | | 1941 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size in acres | number | hours per grid | | size in acres | number | hours per grid | | size in acres | number | hours per grid | |
| | of grids | observed | graduated | | of grids | observed | graduated | | of grids | observed | graduated |
| (1·1) | (1·2) | (1·3) | (1·4) | (2·1) | (2·2) | (2·3) | (2·4) | (3·1) | (3·2) | (3·3) | (3·4) |
| 1 | 7,652 | 0·55 | 0·54 | 1·00 | 882 | 0·82 | 0·76 | 1·0 | 10,435 | 0·59 | 0·60 |
| 4 | 2,238 | 0·82 | 0·78 | 2·25 | 1,223 | 0·90 | 0·86 | 2·0 | 7,908 | 0·63 | 0·68 |
| 9 | 1,432 | 1·18 | 1·18 | 4·00 | 16,499 | 1·00 | 1·00 | 2·5 | 7,201 | 0·67 | 0·70 |
| 16 | 475 | 1·46 | 1·73 | 6·25 | 972 | 1·06 | 1·18 | 3·0 | 13,378 | 0·83 | 0·76 |
| | | | | 9·00 | 1,096 | 1·40 | 1·40 | 4·0 | 11,537 | 0·80 | 0·84 |
| | | | | | | | | 6·0 | 2,157 | 0·98 | 1·00 |
| | | | | | | | | 9·0 | 595 | 1·79 | 1·25 |

252. I may note here that in certain years a quadratic term in $(a)$ was obtained, but the coefficient was small. It has, therefore, been thought advisable to neglect it in the present discussion, although for a finer analysis it may be necessary to retain this term.

253. *Graduation of the journey time.* Again, although journey time is independent of the size of grids, it depends of course on their density. Here also, after a good deal of experimentation, it was decided to use a quadratic expression in $w$ (the density or number of grids per square mile) for a joint graduation of the material for the three years 1939, 1940 and 1941. The constant term was allowed to vary from year to year as in the case of enumeration time, but the coefficients of the linear and quadratic terms were kept the same for all three years. In this way

$$t(j) = c_0'' + 2{\cdot}1120(w) - 0{\cdot}5906(w^2), \qquad (253{\cdot}1)$$

was obtained, where $t(j)$ is the time required for journey in hours per square mile, and $c_0''$ has the value 0·4437, 0·2857 and 0·4531 in 1939, 1940 and 1941 respectively.

254. Here separate parabolic fits for the three years were tried; but the improvement shown in table 17, although statistically significant, was not large, so that there is not much harm in working with the joint fit.

255. The observed and graduated value of journey time for various densities of grids are shown in table 18·2, in which the density (or number per square mile) of grids and their total number are shown in the first two columns, and observed and graduated values of 'journey time' in hours per square mile in the third and fourth columns respectively under each year. The graduation, although not very close, clearly brings out the general features of the function and their broad concordance in different years.

TABLE 18·2. JOURNEY HOURS PER SQUARE MILE BY DENSITY OF GRIDS PER SQUARE MILE

| 1939 | | | | 1940 | | | | 1941 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| density per sq. mile | number | hours per sq. mile | | density per sq. mile | number | hours per sq. mile | | density per sq. mile | number | hours per sq. mile | |
| | of grids | observed | graduated | | of grids | observed | graduated | | of grids | observed | graduated |
| (1·1) | (1·2) | (1·3) | (1·4) | (2·1) | (2·2) | (2·3) | (2·4) | (3·1) | (3·2) | (3·3) | (3·4) |
| 0·25 | 84 | 1·23 | 0·93 | 0·548 | 747 | 1·58 | 1·27 | 0·11 | 595 | 0·74 | 0·68 |
| 0·50 | 324 | 1·72 | 1·35 | 0·759 | 1,256 | 1·75 | 1·55 | 0·33 | 2,157 | 1·12 | 1·09 |
| 1·00 | 1,048 | 1·98 | 1·97 | 0·920 | 1,025 | 1·65 | 1·73 | 0·66 | 11,537 | 1·49 | 1·59 |
| 2·00 | 2,476 | 2·20 | 2·30 | 1·224 | 1,799 | 1·83 | 1·99 | 0·88 | 13,378 | 1·95 | 1·85 |
| | | | | 1·306 | 3,098 | 1·90 | 2·04 | 1·10 | 7,201 | 2·06 | 2·06 |
| | | | | 1·314 | 1,373 | 1·80 | 2·04 | 1·32 | 7,908 | 2·17 | 2·21 |
| | | | | 1·464 | 8,139 | 1·87 | 2·11 | 1·65 | 10,435 | 2·34 | 2·35 |
| | | | | 1·513 | 3,235 | 2·16 | 2·13 | | | | |

256. It is now possible to form the joint-cost function for enumeration plus journey time. Multiplying the former (which is given on a grid basis), that is, equation (249·1), by $w$, the density of grids, and adding to equation (253·1),

$$t(e+j) = c_0 + c_1(w) + 0{\cdot}0793(wa) - 0{\cdot}5906(w^2) \qquad (256{\cdot}1)$$

is obtained, where $t(e+j)$ is the total time for enumeration plus journey and

$$c_0 = 0{\cdot}4437,\ 0{\cdot}2857,\ 0{\cdot}4531, \qquad c_1 = 2{\cdot}5740,\ 2{\cdot}7946,\ 2{\cdot}6319,$$

for three years 1939, 1940 and 1941 respectively.

257. From the above material it has thus been possible to obtain a fairly good idea of the general form of the cost function. The values of the parameters had changed from year to year—a point which will be considered presently—but the data collected in the course of the field survey in different years are in broad agreement, and are also in keeping with what one would expect from general considerations. The present material thus illustrates the possibility of determining the cost function by actual experimental studies on the field.

258. Now consider the question of fluctuations in the parameters from year to year. There are many factors at work, such as regional differences in transport facilities, camping arrangements, attitude of local cultivators, etc., all of which affect the rate of output and hence the cost of operations. Weather conditions change not only from season to season, but also from one part of the season to another, and from region to region in the same season, especially during the pre-monsoon and early monsoon periods.

259. In the present survey, as already pointed out, there were also large differences in training and experience of individual workers in different years. Training exerts a two-fold influence on the cost of operations. First of all a certain amount of time has to be spent in giving training to raw recruits; and as this time has to be provided out of the total quantum of time spent for field work the cost is naturally increased. A large proportion of trained workers thus means lower costs owing to less time having to be spent in giving them training. Secondly, trained workers with previous experience on the whole would be able to do the work more quickly which would reduce costs of operations. A large proportion of trained workers would thus reduce costs, first by consuming less time for training, and secondly by getting the work done more quickly and with greater efficiency.

260. Then there is the inherent variability of the human factor. This is surprisingly high, as was found by a detailed examination of individual records. I have already mentioned, for example, that for a quick and efficient worker the output may be three times as great as that of an indifferent investigator. Finally, owing again to the human element, there are the recording errors which act as a powerful disturbing factor on graduation.

261. The material for the field survey was thus extremely heterogeneous. In this situation it is quite gratifying to find that the general features of the cost function have come out fairly well, and also that graduations based on material collected in different years are in broad agreement.

*Cost of statistical work*

262. Like the field investigators the laboratory workers were also required to maintain diaries in which records were kept of time spent in different types of work. The total number

of hours in this case is of course not 24, but the actual number of hours of work in the Statistical Laboratory; the time spent on indirect purposes thus does not come into the picture.

263. The components in this case are fairly simple. Certain items, such as arranging the village maps and classifying them by zones, would depend only on the number of maps, that is, on the extent of area surveyed, but would be independent of either the density or the size of grid. Items like the locating and marking of grids, calculating averages, etc., would depend on the number (that is, density of grids) but would be independent of their size. Other items again, like the measurement of plot areas or listing of plots, would depend on both size and number (or density) of grids.

264. After some experimentation it was decided to classify the statistical work under the following heads:

(i) Initial work which is independent of either number or size of grid but which depends on the extent of the area (that is, the number of square miles) covered by the survey.

(ii) Work which depends on number (i.e. density) of grids but is independent of their size.

(iii) Work which depends on both size and number (i.e. density) of grids.

265. On the basis of the material collected in successive years it was found that the number of hours of statistical work required for initial arrangements (independently of the size and density of grids) was 0·34 hr./sq. mile. It was also found that the number of hours per sq. mile for the variable portion of the work (depending upon the size and density of grids) could be expressed as a simple function of the following type:

$$t(s) = 0{\cdot}420(w) + 0{\cdot}084(aw), \qquad (265{\cdot}1)$$

where $(a)$ gives the area in acres and $(w)$ the density per square mile of sample grids.

*Total cost per square mile*

266. In order to get the total cost per square mile, the cost for the field and statistical work must be added. But so far these costs have been given in terms of investigator-hours and computer-hours respectively. These must now be converted into money values. For this the cost is required in money per hour of field and statistical work which can be directly calculated from a knowledge of the total number of man-hours and the corresponding total cost in money. The cost for each man-hour is of course itself based on a large number of items. For example, for each computer-hour there must be included not only the pay of the computer but also the pay of the computing inspector and statistical staff, cost of calculating machines, furniture, stationery and other contingent expenses, etc. For organizational purposes these components were studied separately, but these details are not relevant here. The cost in rupees per hour for field and statistical work in different years is shown in

TABLE 19. RELATIVE COST OF FIELD AND STATISTICAL WORK

| year of survey | working hours | | cost in rupees per hour | | ratio of field:statistical |
|---|---|---|---|---|---|
| | field | statistical | field | statistical | |
| (1) | (2) | (3) | (4) | (5) | (6) |
| 1938 | 45,685 | 26,352 | 0·3983 | 0·6261 | 0·6361 |
| 1939 | 125,669 | 54,432 | 0·3763 | 0·6687 | 0·5627 |
| 1940 | 159,398 | 57,024 | 0·3739 | 0·6242 | 0·5990 |
| 1941 | 221,414 | 67,680 | 0·3604 | 0·6870 | 0·5245 |

table 19, in which the total number of man-months spent for both types of work is also shown to give an idea of the volume of material. The cost in man-hours can thus easily be converted into cost in money per square mile by multiplying by the appropriate factor, and money cost per square mile can then be obtained by addition.

*Justification of the present approach*

267. The fact that the numerical value of $g$ is much less than unity shows that the physical field in the case of the jute crop in Bengal is definitely of a non-random type. From an abstract point of view this immediately justifies the use of the theory of grid sampling. In order to appreciate the magnitude of the difference in costs made in practice it is, however, necessary to consider numerical examples. Suppose that it is desired to attain the final estimate with a percentage error of 1 % over a particular region of 5507 sq. miles. Adopting the 1941 cost figures and the actual distribution of size and density used in 1941 on a proportionate basis, it is found that the total cost would be 90,652 investigator-hours for the field work. If, instead of using the above size-density distribution, one uniform size of 20-acre grids had been adopted, then the cost for attaining the same precision would have been 239,926 investigator-hours. If 40-acre grids had been used the cost would have risen still higher to 281,168 investigator-hours. This is a typical example. It is clear that under actual conditions of work in Bengal there is no doubt that the approach adopted in the present work has been definitely more efficient.

268. I may mention here in passing that in 1938 at a meeting of the Jute Census Committee an eminent agricultural expert gave it as his considered opinion that grids of size less than 36 acres must not be used, as otherwise results would be unreliable. The above numerical example shows that had his advice been accepted then the cost for the field portion of the work alone would have been easily three times greater. To obtain a final estimate with the same precision in 1941 the actual cost in working with grids of 36 acres would have been something like Rs. 2,47,400 (£18,555) against Rs. 79,800 (£5925) in the method actually adopted, which means a saving of Rs. 1,67,060 (£12,630) in field work alone in one single year. Further remarks are scarcely necessary.

## 5. Numerical solution of optimum distribution with discussion of errors, etc.

### A. *The graphical-numerical method of solution*

269. Explanation will now be given of the graphical-numerical method of obtaining in particular cases the optimum size and density of grids at a given cost level, using the special forms which were adopted for the area survey work for acreage under jute in Bengal in 1940. The following graduation equations were used for the total cost $T$ and the variance $V$ of the estimated total area under jute:

$$V = \sum_{k=1}^{5} [A_k b_k / w_k (a_k)^{\varepsilon}], \tag{269·1}$$

$$T = \sum_{k=1}^{5} [A_k (c_0 + c_1 w_k + c_2 w_k a_k + c_3 w_k^{\gamma}], \tag{269·2}$$

where in this particular case the summation $k$ was over five zones depending on five different ranges into which $p$ (the proportion under jute) was divided; $c_0$, $c_1$, $c_2$ and $c_3$ are cost para-

meters supposed to be constant over the different zones; $A_k$ is the area of the $k$th zone; $a_k$ and $w_k$ are the size and density of the grids in the $k$th zone in any conventional units, say acres and number per square miles respectively, $b_k$ is a zonal constant, and $g$ is a pooled value over all zones and is obtained in the following way.

270. The variance function in the $k$th zone is graduated by the formula

$$V_k = A_k b_k / w_k (a_k)^{gk} \quad (k = 1, 2, 3, \ldots, 5). \tag{270·1}$$

(In any zone, for purposes of graduation a number of size-density combinations in the exploratory state were used, though in the final design only one combination was used, namely, the optimum.) Having obtained $b_k$'s ($k = 1, 2, \ldots, 5$), the whole variance material is graduated by the equation

$$\frac{V_k}{b_k} = \frac{A_k}{w_k (a_k)^g}. \tag{270·2}$$

It is to be noted here that for any size and density there may be a number of values of observed variances; each such value is to be divided by the $b_k$ of the zone from which the variance has been observed; having got the values of $V_k/b_k$ from the whole area, the value of $g$ was fitted by the formula (270·2). This is, as already mentioned, a pooled '$g$', the pooling being as explained above. For optimum then

$$\delta V + \frac{1}{\lambda} \delta T = 0, \tag{270·3}$$

which in the present case leads to the equations

$$(c_1 + c_2 a_k + 2 c_3 w_k)\,(a_k)^g\, w_k^2 / b_k = \lambda, \quad c_2 w_k^2 (a_k)^{g+1} / b_k g = \lambda. \tag{270·4}$$

From (270·4), it follows that

$$a_k = (c_1 + 2 c_3 w_k)\, g / (1 - g)\, c_2, \quad (k = 1, 2, \ldots, 5). \tag{270·5}$$

Substituting this in any of the equations (270·4) it is found that*

$$c_2 w_k^2 \left\{ \frac{(c_1 + 2 c_3 w_k)\, g}{(1 - g)\, c_2} \right\}^{g+1} \frac{1}{g} = b_k \lambda \equiv \mu_k, \quad (k = 1, 2, \ldots, 5). \tag{270·6}$$

271. There are eleven unknowns, $\lambda$, $a_k$, $w_k$ ($k = 1, 2, \ldots, 5$). There are also ten equations (270·4), while the total cost $T$ given by (269·2) must be kept fixed at a given value, which furnishes a further equation. It can easily be seen that these equations are compatible and independent. The equations (269·2) and (270·4) cannot, however, be algebraically solved. Recourse is had to a graphical numerical procedure for solving these equations on the following lines. In (270·6) dropping $k$, a sufficient number of values of $w$ are taken and the corresponding $\mu$'s calculated by formula (270·6); this material is arranged in the form of a table or a large-scale $(w, \mu)$-graph is drawn. From the graph or from the $(w, \mu)$-table (by interpolation if necessary), $w$ can be read off or calculated for any given value of $\mu$, there being only one real positive value of $w$ for any given value of $\mu$. For any value of $\lambda$ there are, from (270·6), five values of $\mu_k$ depending on five values of $b_k$ for the different zones. For each

* The auxiliary parameters $\lambda$ and $\mu$ introduced in equations (270·3) and (270·6) respectively are of course different from the parameters $\lambda$ and $\mu$ of Part II. In fact, $\lambda$ here is the reciprocal of the $\lambda$ used in Part II.

of these values of $\mu_k$ there is a real positive value of $w_k$ from which is obtained the corresponding value (real and positive) of $a_k$ by (270·5). Substituting these values of $a_k$ and $w_k$ (five pairs altogether obtained from five values of $b_k$ for five different zones) in the cost equation (269·2), a value of $T$ is thereby obtained. Starting from any value of $\lambda$ and proceeding along a particular chain a certain value for the total cost $T$ is reached. (Also, by substitution of these $a_k$, $w_k$ the variance $V$ is obtained from (269·1), which is, of course, the minimum variance at the corresponding cost level; but this is not of immediate interest.) For each value of $\lambda$ there is then a corresponding value of $T$. After obtaining a sufficient number of pairs of values of $\lambda$ and $T$, a $(\lambda, T)$-table and/or a large-scale $(\lambda, T)$-graph can then be prepared.

272. To get the optimum size-density distribution at any given cost level $T$ it is necessary to go back along the chain by which $T$ was obtained from $\lambda$, to be more explicit, for any value of $T$ the corresponding value of $\lambda$ must be read off from the $(\lambda, T)$-graph, or calculated by interpolation from the $(\lambda, T)$-table. From that value of $\lambda$, five different values of $\mu_k$ are obtained from (270·6); for each of these $\mu_k$'s ($k = 1, 2, \ldots, 5$) the corresponding $w_k$ is acquired from the $(w, \mu)$-table; for each $w_k$, $a_k$ is obtained from (270·5). These values of $a_k$, $w_k$ ($k = 1, 2, \ldots, 5$) constitute the optimum size-density distribution at the given cost level $T$. Finally, substituting these $a_k$'s and $w_k$'s in (269·1), the error for the optimum solution is obtained. This will be the minimum error; it can be shown that a solution is reached which gives the true minimum and not merely a stationary value for the variance $V$.

273. There is one point of practical importance to be noted in the above procedure. It has been observed that from any $\lambda$ it is possible to reach by a certain chain the corresponding $T$ and vice versa. Unless optimum solutions for a large number of cost levels are called for it is usually a wasteful process to draw a $(\lambda, T)$-graph or table which is very close over a wide range of values of $T$. A close graph for the region here described would be more worth while and hence more economical. This can be achieved if in the first instance a rough $(\lambda, T)$-graph or table is prepared from which is obtained a first approximation to the value of $\lambda$ for a given $T$. Near about this value of $\lambda$ on both sides, a large number of values of $\lambda$ can be taken and the corresponding values of $T$ calculated, thus preparing a $(\lambda, T)$-graph and table which are sufficiently close in the region of interest.

274. It may be mentioned in conclusion that over all the three years of area census of jute 1939–41 the form of the variance function was the same as (269·1), but the cost was of the general form

$$T = \sum_{k=1}^{l} [A_k(c_0 + c_1 w_k + c_2 a_k w_k + c_3 w_k^2 + c_4 a_k^2 w_k)], \qquad (274\cdot1)$$

and the coefficient $c_4$ was zero in particular years; (269·2) was thus a special case of (274·1). The numbers of zones were different in different years, and so also the numerical values of all the parameters $c_0$, $c_1$, $c_2$, $c_3$ and $c_4$ as also $A_k$'s; this is true also of $g$ and $b_k$ of (269·1). Such variation for purposes of experimentation was essential in the exploratory stage. However, it is obvious that the graphical-numerical method explained in the foregoing pages can be immediately applied to (269·1) and (274·1), i.e. to find out the size-density distribution which will minimize the variance given by (269·1) at a given fixed value for the cost $T$ given by (274·1).

275. A numerical illustration of the above method giving for the 1941 material the $(w, \mu)$ and $(\lambda, T)$-tables, and also the optimum distributions at a few cost levels are given in tables 20·1–20·4.

Sample survey of area under jute in Bengal, 1941: pooled $g = 0{\cdot}4707$

$$T = \sum_{k=1}^{5} [A_k(0{\cdot}4881 + 3{\cdot}3519w_k + 0{\cdot}1536a_k w_k - 0{\cdot}7481w_k^2)], \qquad (275{\cdot}1)$$

$$V = \sum_{k=1}^{5} [A_k b_k / w_k (a_k)^{0{\cdot}4707}]. \qquad (275{\cdot}2)$$

TABLE 20·1. VALUES OF PARAMETER FOR DIFFERENT ZONES

| numbers ($k$) ... | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| range of ($p$) ... | 0·00–0·05 | 0·05–0·10 | 0·10–0·15 | 0·15–0·20 | 0·20–0·25 |
| zone | | | | | |
| $b_k$ | 0·0076 | 0·0167 | 0·0285 | 0·0360 | 0·0719 |
| $A_k$ (sq. miles) | 14,186 | 12,677 | 5,937 | 1,418 | 203 |

TABLE 20·2. TOTAL COST IN RUPEES AGAINST DIFFERENT VALUES OF $\lambda$

| serial numbers ... | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| values of $\lambda$ ... | 2·5 | 3 | 10 | 12 | 14 | 20 | 25 | 43 | 75 | 150 |
| total cost in rupees | 25,347 | 26,121 | 33,469 | 33,417 | 36,747 | 41,545 | 42,480 | 50,206 | 58,725 | 78,541 |

TABLE 20·3. VALUES OF $\mu$ AND $w$

| $\mu$ | $w$ | $\mu$ | $w$ | $\mu$ | $w$ | $\mu$ | $w$ | $\mu$ | $w$ |
|---|---|---|---|---|---|---|---|---|---|
| 0·019 | 0·030 | 0·360 | 0·120 | 0·863 | 0·180 | 1·256 | 0·220 | 2·669 | 0·350 |
| 0·076 | 0·055 | 0·430 | 0·130 | 0·900 | 0·190 | 1·439 | 0·240 | 3·093 | 0·380 |
| 0·152 | 0·090 | 0·502 | 0·150 | 1·007 | 0·200 | 1·541 | 0·250 | 4·218 | 0·470 |
| 0·234 | 0·099 | 0·571 | 0·160 | 1·137 | 0·210 | 1·798 | 0·280 | 5·395 | 0·560 |
| 0·285 | 0·110 | 0·714 | 0·170 | 1·227 | 0·220 | 2·141 | 0·500 | 10·789 | 1·040 |

TABLE 20·4. SIZE AND DENSITY OF GRIDS, AND PERCENTAGE ERROR AT DIFFERENT LEVELS OF COST

| zones | | levels of cost: Rs. 25,347 | | | Rs. 36,747 | | | Rs. 50,206 | | | Rs. 78,541 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | range of '$p$' | $w$ | $a$ | % error | $w$ | $a$ | % error | $w$ | $a$ | % error | $w$ | $a$ | % error |
| (0·1) | (0·2) | (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 | 0·00–0·05 | 0·030 | 19·1 | 8·43 | 0·065 | 18·8 | 5·75 | 0·120 | 18·4 | 4·25 | 0·210 | 17·6 | 3·25 |
| 2 | 0·05–0·10 | 0·041 | 19·1 | 3·78 | 0·099 | 18·5 | 2·45 | 0·170 | 17·9 | 1·88 | 0·340 | 16·5 | 1·36 |
| 3 | 0·10–0·15 | 0·054 | 18·9 | 3·78 | 0·140 | 18·2 | 2·37 | 0·220 | 17·5 | 1·91 | 0·470 | 15·3 | 1·35 |
| 4 | 0·15–0·20 | 0·060 | 18·9 | 5·89 | 0·150 | 18·1 | 3·75 | 0·250 | 17·2 | 2·94 | 0·560 | 14·6 | 2·05 |
| 5 | 0·20–0·25 | 0·086 | 18·7 | 14·32 | 0·200 | 17·7 | 9·52 | 0·380 | 16·1 | 7·05 | 1·040 | 10·4 | 4·73 |
| all | 0·00–0·25 | | | 2·31 | | | 1·51 | | | 1·16 | | | 0·89 |

### B. *Uncertainty in the estimated optimum size, density and variance due to errors in the parameters of graduation*

276. In the method followed here particular forms are assumed for the cost and variance functions, and then the observed material is graduated by estimating the parameters of cost and variance by the method of least squares applied either directly to the functions themselves or to their logarithms. It is plausible to assume that the material in use (on which the cost and variance functions have been fitted) forms but one possible sample drawn out of

a population conforming to the usual well-known postulates behind the fitting of single or multi-variate linear regressions or the fitting of a polynomial regression of one variate on another. On this view there will be an error in each of the fitted or estimated values of the parameters of the cost and variance functions, and the error can be derived from the corrected second moment of the fitted parameter in question over the totality of all possible samples of which the present material happens to be one. In this particular context the given material will sometimes be called a sample, the population behind it being assumed to conform to the well-known pattern of regression theory. A parameter obtained from large samples of bulky material may also be assumed under certain mild and well-known restrictions to be approximately normally distributed, and in such a situation the 'error' of the parameter in question can be calculated approximately with sufficient accuracy for all practical purposes. If the sample be small or if otherwise the properties of large sample theory cannot be used, the exact sampling distribution of the fitted parameters and some characteristic of this distribution (which may be other than the corrected second moment) may be needed for purposes of determining the confidence interval at any assigned level. In any case, whether it be the 'error' or some other characteristic of the exact sampling distribution of the parameters intended to supply the confidence interval, for simplicity this will be called the 'uncertainty' of the fitted parameter. This uncertainty will lead to a corresponding uncertainty in the estimated optimum distribution of size and density of grids over different zones at a given level of total cost. Given the 'uncertainty' in the parameters of the fitted cost and variance equations the uncertainty in the estimated optimum size-density distribution is always theoretically determinate, although the actual calculation of the numerical values might in practice offer considerable difficulties. Having given an outline of the problem in an abstract form, numerical examples will now be given illustrating the procedure in a special case which rests upon certain assumptions, simple but general enough for present purposes or for most practical purposes so far as large-scale sample surveys are concerned.

277. Consider the 1941 material for the sample survey of area under jute. This material is large enough to justify making the usual assumptions of large-sample theory. In such a case it will do if only the 'errors' (or corrected second moments of the large sample distribution) of the parameters of the fitted equations are calculated by the usual well-known process which need not be discussed. The 'errors' of these parameters having been calculated—the corresponding confidence interval at any assigned level when the sampling distribution of the parameters follows the normal law will be then given in terms of the 'errors'. These being given, the procedure whereby the errors in the estimated optimum size-density distributions are obtained will be illustrated.

278. For this material the cost $T$ and the variance $V$, as observed earlier, were graduated by

$$V = \sum_{k=1}^{5} [A_k b_k / w_k (a_k)^g], \tag{278·1}$$

$$T = \sum_{k=1}^{5} [A_k (c_0 + c_1 w_k + c_2 a_k w_k + c_3 w_k^2)], \tag{278·2}$$

where, as already explained earlier, the whole field was divided into five zones depending on the proportion of area under jute in each. The value of $g$ in (278·1), and of $c_0$, $c_1$, $c_2$, $c_3$ in

(278·2) are parameters supposed to be constant for all zones; $b_k$ in (278·1) is a zonal constant (depending on $p$, the proportion under jute and also on other physical features of the zone); $A_k$ is the area of $k$th zone; and $w_k$ and $a_k$ denote the density and size of grids in any conventional units ($k = 1, 2, 3, 4, 5$).

279. The optimum equations for minimum variance at an assigned cost level would be, as noted earlier,

$$\delta V + \frac{1}{\lambda}\delta T = 0, \tag{279·1}$$

which leads in the present case to

$$\left.\begin{aligned} \frac{b_k g}{(a_k)^{g+1}} - \frac{1}{\lambda}(c_2 w_k^2) &= 0, \\ \frac{b_k}{(a_k)^g} - \frac{1}{\lambda}\{w_k^2(c_1 + c_2 a_k + 2c_3 w_k)\} &= 0, \end{aligned}\right\} \quad (k = 1, 2, \ldots, 3) \tag{279·2}$$

with $T$ = constant. Solving these equations we can obtain the optimum $w_k$ and $a_k$ with ($k = 1, 2, \ldots, 5$), as also the undetermined multiplier $\lambda$. Denoting now by $\delta$ the error in any parameter or in $w_k$ or $a_k$, and assuming that this is small compared to the quantities themselves, and ignoring squares and higher powers of $\delta$, it follows

$$\frac{g}{(a_k)^{g+1}}\delta b_k - \frac{b_k g(g+1)}{(a_k)^{g+2}}\delta a_k + \left\{\frac{b_k}{(a_k)^{g+1}} - \frac{b_k g \log a_k}{(a_k)^{g+1}}\right\}\delta g - \frac{1}{\lambda}(2w_k c_2)\,\delta w_k - \frac{1}{\lambda}w_k^2 \delta c_2 + \frac{1}{\lambda^2}c_2 w_k^2 \delta\lambda = 0, \tag{279·3}$$

$$\frac{\delta b_k}{(a_k)^g} - \frac{b_k g}{(a_k)^{g+1}}\delta a_k - \frac{b_k \log a_k}{(a_k)^g}\delta g - \frac{1}{\lambda}\{2w_k(c_1 + c_2 a_k + 3c_3 w_k)\}\delta w_k - \frac{1}{\lambda}w_k^2(\delta c_1 + a_k \delta c_2 + c_2 \delta a_k + 2w_k \delta c_3) + \frac{1}{\lambda^2}w_k^2(c_1 + c_2 a_k + 2c_3 w_k)\,\delta\lambda = 0, \tag{279·4}$$

with $k = 1, 2, \ldots, 5$. The total cost being fixed the further equation then follows:

$$\delta T = \sum_{k=1}^{5}[A_k(\delta c_0 + w_k \delta c_1 + c_1 \delta w_k + w_k^2 \delta c_3 + 2c_3 w_k \delta w_k + a_k w_k \delta c_2 + c_2 w_k \delta a_k + c_2 a_k \delta w_k)] = 0. \tag{279·5}$$

Given $\delta c_0$, $\delta c_1$, $\delta c_2$, $\delta c_3$, $\delta g$ and $\delta b_k$ ($k = 1, 2, \ldots, 5$) from the usual processes of linear regression theory, and $\lambda$, $w_k$, $a_k$ ($k = 1, 2, \ldots, 5$) from the eleven equations (279·2) and $T$ = constant, it is possible to obtain in a fairly simple manner the errors $\delta\lambda$, $\delta w_k$ and $\delta a_k$ ($k = 1, 2, \ldots, 5$) from the eleven equations (279·3), (279·4) and (279·5); $\delta V$ from the differential of (278·1); and finally the percentage errors

$$100\frac{\delta w_k}{w_k}, \quad 100\frac{\delta a_k}{a_k}, \quad \text{and} \quad 100\frac{\delta V}{V}.$$

280. In table 21 are given the relevant numerical values for the 1941 material and graduation for a cost level of Rs. 50,000. It will be noticed that the percentage errors for $w_k$ and $V_k$, i.e. for the optimum density of grids and the zonal variance, fluctuate roughly between 6 and 12 %, while the percentage error of the optimum size of grids, $a_k$, fluctuates between 11 and 24 %. These numerical values are given here, of course, for purposes of

illustration, but they indicate broadly the order of precision which was attained in practice. On the whole the results are not unsatisfactory.

*SAMPLE CENSUS OF AREA UNDER JUTE*, 1941

TABLE 21. ERRORS OF OPTIMUM SIZE AND DENSITY OF GRIDS AND VARIANCE DUE TO ERRORS IN THE PARAMETERS OF GRADUATION

| $k$ | $p$-level | $b_k$ | $w_k$ | $a_k$ | $V_k$ |
|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) |
| 1 | 0·000–0·050 | 0·007582 | 0·12 | 18·36 | 227·78 |
| 2 | 0·050–0·100 | 0·016740 | 0·17 | 17·93 | 320·78 |
| 3 | 0·100–0·150 | 0·028543 | 0·22 | 17·50 | 200·24 |
| 4 | 0·150–0·200 | 0·035992 | 0·25 | 17·24 | 53·22 |
| 5 | 0·200–0·250 | 0·071929 | 0·38 | 16·11 | 10·38 |

| | maximum error of | | | | maximum percentage error | | |
|---|---|---|---|---|---|---|---|
| $k$ | $b_k$ | $w_k$ | $a_k$ | $V_k$ | $w_k$ | $a_k$ | $V_k$ |
| (1) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| 1 | 0·000337 | 0·006800 | 2·2062 | 15·8425 | 5·67 | 12·02 | 6·96 |
| 2 | 0·000854 | 0·010056 | 1·8990 | 18·8449 | 5·91 | 10·59 | 5·87 |
| 3 | 0·000856 | 0·008575 | 2·0495 | 12·9892 | 3·90 | 11·71 | 6·49 |
| 4 | 0·002360 | 0·017872 | 2·5488 | 4·0756 | 7·15 | 14·78 | 7·63 |
| 5 | 0·007912 | 0·045170 | 3·9208 | 1·2895 | 11·89 | 24·34 | 12·38 |

### C. *Fuller numerical treatment of the single-zone problem*

281. For purposes of illustration of the theory for a single zone, a start is made from a cost function of the form

$$T = A(c_0 + c_1 w + c_2 aw + c_3 w^2) \tag{281·1}$$

and a variance function of the form

$$V = \frac{Apq}{w(n_0)^g} = \frac{Apq(\square)^g}{w(a)^g}, \tag{281·2}$$

where $A$ is the total area under survey in square miles, $w$ is the number of grids per square mile, $n_0$ the number of basic cells in the grid, and $a = n_0\square$ is the size in acres of each grid, $p$ is the proportion of total area under the crop (being surveyed), and $q$, of course, is $(1-p)$; $c_0, c_1, c_2, c_3$ are parameters which depend upon the physical conditions of the zone (and the crop in question); and $\square$ and $g$ may be related to $p$. These forms are suggested by graduation of the material relating to the sample surveys (conducted over 5 years, 1937–41) of area under jute in Bengal. The percentage variability of the estimated proportion will be denoted by $100\sqrt{(V)}/Ap$ $(= 100s/Ap$, where $s = \sqrt{(V)}$ is written as the standard error of the estimate); and as in Part II, § 5–B, the amount of information in respect of the estimated proportion of area under the crop is given by $A^2/V$. For given values of the parameters $A, c_0, c_1, c_2, c_3, p$, $\square$ and $g$, and a pre-assigned cost level $T$, we can by methods indicated in Part II, § 5–B, calculate the optimum $w$ and $a$ for minimum $V$ (or error) and obtain therefrom the maximum information $(\vartheta)$. As indicated in the section just referred to, the rate of change of information can also be obtained with regard to total cost $\partial T/\partial\vartheta$, the proportional rate of change of information with cost $\frac{1}{\vartheta}\frac{\partial\vartheta}{\partial T}$, and finally the proportional rate of change of information with respect to proportional change in cost $\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}$.

282. It has already been observed that $\square$ and $g$ may depend on $p$, the nature of dependence differing for different crops. Independent variations to $g$ and $p$ will not be allowed (with a fixed value of $\square$), giving thus in one scheme solutions for different crops (involving different relationships between $g$, $\square$ and $p$). With this provision it will be evident from § 5–B of Part II that the optimum size-density distribution $(a, w)$ would be independent of the total area $A$ and also of $p$, and would depend wholly on the other cost and variance parameters, and on the cost per square mile; $\left(\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ is independent of $A$ and $p$ but depends on the cost per square mile and other parameters; $\left(\frac{1}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ is independent of $p$ but depends on the other factors; $\left(\frac{\partial\vartheta}{\partial T}\right)$ is independent of $A$; while $V$ and $100s/Ap$ depend on all the factors considered. In particular,

$$T = 10{,}000\{0{\cdot}3559 + 1{\cdot}3237w + 0{\cdot}060(aw) - 0{\cdot}2645w^2\}, \qquad (282{\cdot}1)$$

and $\square = 0{\cdot}04$ (from the 1940 jute material). The nature of the dependence of the solution on $g$, $p$ and $t = T/A$, that is, cost per square mile, is now brought out in table 22. It will be seen that, within our range of values, for fixed $g$ and $p$, with increasing $t$, $(a)$ decreases, $(w)$ increases and $100s/ap$ decreases; for fixed $t$ and $p$ with increasing $g$, $(a)$ increases, $(w)$ decreases and percentage variability decreases; and finally, for fixed $t$ and $g$, $(a)$ is independent of $p$. Coming now to information and its rate of change it will be noticed that $\frac{1}{\vartheta}\frac{\partial\vartheta}{\partial T}$ decreases with increasing total cost $(T)$ or rather with $t$ ($A$ being kept constant), the decrease being slower for larger values of $g$.

TABLE 22. OPTIMUM SINGLE-ZONE SOLUTION FOR DIFFERENT COST LEVELS

| cost in rupees per sq. mile | size | density | $\frac{1}{\vartheta}\frac{\partial\vartheta}{\partial T}$ | $\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}$ | $p=0{\cdot}05$ % error | $p=0{\cdot}05$ $\vartheta$ | $p=0{\cdot}05$ $\frac{\partial\vartheta}{\partial T}$ | $p=0{\cdot}20$ % error | $p=0{\cdot}20$ $\vartheta$ | $p=0{\cdot}20$ $\frac{\partial\vartheta}{\partial T}$ | $p=0{\cdot}40$ % error | $p=0{\cdot}40$ $\vartheta$ | $p=0{\cdot}40$ $\frac{\partial\vartheta}{\partial T}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6·1) | (6·2) | (6·3) | (7·1) | (7·2) | (7·3) | (8·1) | (8·2) | (8·3) |
| | | | | | | $g=0{\cdot}3$ | | | | | | | |
| 0·50 | 9·07 | 0·0773 | $10^{-4}\times 3{\cdot}9384$ | 1·9692 | 6·95 | $10^4\times 8{\cdot}27$ | 32·57 | 3·19 | $10^4\times 2{\cdot}46$ | 9·69 | 1·95 | $10^4\times 1{\cdot}64$ | 6·46 |
| 0·75 | 8·54 | 0·2209 | 1·4775 | 1·1081 | 4·15 | 23·20 | 34·28 | 1·90 | 6·90 | 10·19 | 1·17 | 4·60 | 6·80 |
| 1·00 | 7·95 | 0·3776 | 0·9384 | 0·9384 | 3·20 | 38·91 | 36·51 | 1·47 | 11·55 | 10·83 | 0·90 | 7·69 | 7·22 |
| 1·25 | 7·30 | 0·5517 | 0·7097 | 0·8871 | 2·69 | 55·25 | 39·21 | 1·23 | 16·45 | 11·67 | 0·76 | 10·96 | 7·78 |
| 1·50 | 6·55 | 0·7522 | 0·5921 | 0·8882 | 2·34 | 72·95 | 43·20 | 1·07 | 21·69 | 12·84 | 0·66 | 14·47 | 8·57 |
| 2·00 | 4·40 | 1·3265 | 0·5546 | 1·1093 | 1·87 | 114·94 | 63·75 | 0·86 | 34·01 | 18·86 | 0·42 | 22·62 | 12·55 |
| | | | | | | $g=0{\cdot}6$ | | | | | | | |
| 0·50 | 32·19 | 0·0440 | $10^{-4}\times 1{\cdot}6467$ | 0·8233 | 2·79 | $10^4\times 51{\cdot}28$ | 84·44 | 1·28 | $10^4\times 15{\cdot}22$ | 25·06 | 0·78 | $10^4\times 10{\cdot}15$ | 16·71 |
| 0·75 | 31·14 | 0·1239 | 0·6141 | 0·4606 | 1·65 | 140·85 | 86·50 | 0·77 | 42·02 | 25·81 | 0·47 | 28·01 | 17·20 |
| 1·00 | 30·03 | 0·2085 | 0·3854 | 0·3854 | 1·31 | 232·56 | 89·63 | 0·60 | 69·45 | 26·77 | 0·37 | 46·08 | 17·76 |
| 1·25 | 28·85 | 0·2988 | 0·2860 | 0·3576 | 1·11 | 322·58 | 92·28 | 0·51 | 97·09 | 27·78 | 0·31 | 64·52 | 18·46 |
| 1·50 | 27·58 | 0·3957 | 0·2319 | 0·3479 | 0·98 | 416·61 | 96·64 | 0·45 | 125·00 | 28·99 | 0·27 | 83·33 | 19·33 |
| 2·00 | 24·65 | 0·6797 | 0·1784 | 0·3567 | 0·81 | 625·00 | 111·47 | 0·37 | 181·82 | 32·43 | 0·23 | 121·95 | 21·75 |

283. The same thing happens to $\left(\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ with regard to increasing cost per square mile, but beyond a certain value of $t$, this quantity appears to start increasing with increase of $t$. The cost equation used here would not, however, hold far beyond this level, and hence any conclusions about how $\left(\frac{T}{\vartheta}\frac{\partial\vartheta}{\partial T}\right)$ behaves beyond this level would not be safe. Both $\partial$ and $\frac{\partial\vartheta}{\partial T}$ increase with increasing cost for $g$, but decrease with increasing $p$.

284. For an area of 10,000 sq. miles and for the values already assumed of the cost and variance parameters (except, of course, of $g$ and $p$ and $t$ which we keep flexible) from table 23 it is possible to find out $t$, the minimum cost in rupees per square mile for different levels of percentage error (percentage variability) and for two different values of $g$ and three different values of $p$. Finally, table 24 gives at once, for the same values of $g$ and $p$, the cost in rupees per unit of information for different levels of $t$, the cost in rupees per square mile. It will be seen from table 23 that the minimum cost per square mile decreases with increase in either the pre-assigned percentage error or in $g$ or $p$. Table 24 brings out that the cost in rupees per unit of information decreases with increasing $t$, this decrease being less rapid for larger values of $g$; further, this quantity (for any given value of $t$) decreases with increasing $g$ but increases with increasing $p$.

TABLE 23. MINIMUM COST IN RUPEES PER SQUARE MILE FOR DIFFERENT LEVELS OF PERCENTAGE ERROR

| percentage error | $p=0{\cdot}05$ | | $p=0{\cdot}20$ | | $p=0{\cdot}40$ | |
|---|---|---|---|---|---|---|
| | $g=0{\cdot}3$ | $g=0{\cdot}6$ | $g=0{\cdot}3$ | $g=0{\cdot}6$ | $g=0{\cdot}3$ | $g=0{\cdot}6$ |
| (1) | (2·1) | (2·2) | (3·1) | (3·2) | (4·1) | (4·2) |
| 0·2 | — | — | — | — | — | 2·1 |
| 0·3 | — | — | — | — | — | 1·2 |
| 0·4 | — | — | — | 1·8 | 1·8 | 0·9 |
| 0·5 | — | — | — | 1·2 | 0·9 | 0·7 |
| 1·0 | — | 1·5 | 1·5 | 0·6 | 0·7 | — |
| 1·5 | — | 0·9 | 1·0 | 0·5 | 0·5 | — |
| 2·0 | 1·9 | 0·7 | 0·7 | — | — | — |
| 2·5 | 1·4 | 0·5 | 0·6 | — | — | — |
| 3·0 | 1·1 | — | 0·5 | — | — | — |
| 3·5 | 0·9 | — | — | — | — | — |
| 4·0 | 0·8 | — | — | — | — | — |
| 4·5 | 0·7 | — | — | — | — | — |
| 5·0 | 0·6 | — | — | — | — | — |

TABLE 24. COST IN RUPEES PER UNIT OF INFORMATION FOR DIFFERENT LEVELS OF COST IN RUPEES PER SQUARE MILE

| cost per sq. mile | $p=0{\cdot}05$ | | $p=0{\cdot}20$ | | $p=0{\cdot}40$ | |
|---|---|---|---|---|---|---|
| | $g=0{\cdot}3$ | $g=0{\cdot}6$ | $g=0{\cdot}3$ | $g=0{\cdot}6$ | $g=0{\cdot}3$ | $g=0{\cdot}6$ |
| (1) | (2·1) | (2·2) | (3·1) | (3·2) | (4·1) | (4·2) |
| 0·5 | 604 | 98 | 2032 | 328 | 3051 | 492 |
| 1·0 | 323 | 53 | 1087 | 178 | 1631 | 268 |
| 1·5 | 257 | 43 | 866 | 144 | 1300 | 217 |
| 2·0 | 226 | 39 | 760 | 129 | 1140 | 194 |
| 2·5 | 206 | 36 | 692 | 120 | 1036 | 180 |
| 3·0 | 174 | 32 | 588 | 110 | 884 | 164 |

## 6. GENERAL DESCRIPTION OF JUTE CENSUS, 1941

285. In this section I propose giving a concise account of the sample survey of the area under jute in Bengal as carried out in 1941–2. The total area covered by the survey was 59,199 sq. miles, consisting of 505 *thanas* (a *thana* is an administrative unit of roughly 120 sq. miles on an average, which is under the jurisdiction of a single police station) or 'police stations' as they are called in English. The proportion of land under jute in each police

station, which fluctuated widely, was obtained from official records of jute registration in 1939. This furnished the basis for the planning of the sample survey in 1941. The area under jute was found to be extremely small in each of 75 thanas covering 11,234 sq. miles; and these were left out of the sample survey, but arrangements were made for collection of information through personal inquiry and inspection.

286. *Zoning.* The remaining 430 thanas covering 47,965 sq. miles were included in the sample survey. These thanas were grouped into ten different zonal classes in accordance with the proportion of land under jute (0·01–0·02; 0·02–0·04; 0·04–0·08; 0·08–(0·06)–0·04, above 0·44). The basis for zoning was, however, two-fold, namely, the above abstract zonal classes, and secondly, the thana as the geographical unit. The size and density of grids in each of the ten abstract zonal classes was determined by the intensity of cultivation. But each such abstract zone consisted of a large number of thanas which did not necessarily form a compact block but were often scattered throughout the whole area under survey. Theoretically there would be of course only one optimum size and density for each abstract class of zones and hence for all thanas included in the respective zonal class. In 1941, however, the single optimum size and density were not strictly adhered to but grids of somewhat different sizes were used for each zonal class with a view to collecting information for the study of the variance and cost functions.

287. *Half-sample method.* As already mentioned, the survey was arranged in the form of two interpenetrating subsamples which were called half-samples (A) and (B). The sample units were arranged in pairs, and one grid of each pair was allotted at random to half-sample (A) and the other group to half-sample (B). Information was collected independently for the two half-samples by different sets of investigators, and the time programme was arranged in such a way that the two sets of investigators never worked in the same region at the same time, so that the chance of comparing records was practically nil. As will be seen later, this arrangement proved entirely successful. Each pair of grids was located purely at random. A dumbbell-shaped figure was used for this purpose, the two ends representing the two half-samples. The distance between the two half-samples was thus kept constant, but the principle of randomization was preserved by allowing the dumbbell-shaped figure itself to be located at random, and also to have a random orientation by suitable optical methods. Randomization was completed separately for each zone, that is for each thana or geographical unit within each zonal class.

288. *Preparatory work.* The preliminary work was started in September 1940 and was in full swing by the end of the year. The plan of the survey was settled on the lines already indicated. The sample grids, 28,942 pairs or 57,880 in number altogether, were located and marked at random on village maps for which something like 102,000 cadastral survey sheets had to be handled. The field lists were then prepared, and these were arranged in separate bundles for the different sets of investigators, and were supplied to the field branch from the Calcutta Statistical Laboratory at the end of April 1941.

289. *Organization of field work.* The whole area was divided into ten blocks, each of which was in charge of a chief inspector; and each block was divided into two or three subblocks, each of which consisted of 15–20 thanas covering roughly 2000 sq. miles; in each sub-block there were two inspectors (each in charge of from six to eight investigators), one for

each of the two half-samples (A) and (B). Certain additional field units were also employed for providing comparisons between different subblocks.

290. The total field staff in 1941 consisted of 372 investigators, fifty-three inspectors, ten chief inspectors, sixty-three messenger peons, and an office staff of eleven under one assistant supervisor, and a field supervisor, who was a Government official of the rank of a deputy magistrate; and the whole work was done under my general guidance as Honorary Secretary of the Indian Statistical Institute. The field staff were recruited from the middle of April, and the investigators and inspectors were given training in field work for about a fortnight. The field survey began between 7 and 14 May in different parts of Bengal and was completed on 13 August 1941.

291. *Area extraction and tabulation.* In the meantime the area of individual plots included within the grids were being measured in the Statistical Laboratory. This was laborious work and involved measurements of about 490,000 different plots. The first batch of field records was received in the Laboratory on 13 May 1941, and the work of tabulation and calculations started on the very same day, and involved records for 175,579 individual plots which were reported to be either wholly or partly under jute.

292. The whole work naturally had to be done at high pressure, and forward planning was needed to avoid bottlenecks. The number of computers had to be varied to suit the exigencies of work from day to day. The average number of computers engaged in the jute scheme was about forty in 1941, but this number had sometimes to be increased to eighty or ninety or reduced to only ten or twelve at certain times. This was done by shifting the men from one project to another.

293. As already mentioned, the actual field survey was completed on 13 August, and the last batch of field records was received in the Statistical Laboratory on 20 August. A preliminary estimate of the area under jute was submitted to the Government of Bengal on 27 August or within exactly one week of the receipt of the last batch of field records. A general report dealing with the area under jute in individual districts, and administrative and budgetary matters, was submitted in December 1941, after which the more advanced statistical analysis of the material was taken up, of which an account is being given here.

294. The whole work was done under the general guidance of the Calcutta Statistical Laboratory, and it was necessary to give a good deal of attention to the organization and day to day progress of the field inquiry. The survey thus involved not only statistical but a large amount of administrative and organizational work of a most varied kind. It was possible to undertake this only because of the willing co-operation and efficient team-work of the staff of the Statistical Institute.

## *Precision of the sample estimate*

295. *Agreement between subsamples (A) and (B).* It is not necessary to discuss here the estimates for different regions, but the reliability of the results is a matter of direct statistical interest. As already mentioned, information was available for each of a linked pair of grids (A) and (B) obtained by different sets of investigators. These were scrutinized and compared pair by pair, but the results are too bulky to be reproduced here. For present purposes it would be sufficient if the agreement between pairs of grids for the *thana* (police station or

zone) were considered as a whole, for each of which the procedure of randomization was completed separately. In each thana or zone the difference between the estimates for each pair of grids was tabulated, and from these differences the value of 'Student's' $t$-statistic was obtained. It is not possible to use Bartlett's ($\sin^{-1}\sqrt{p}$) transformation, as the estimates are not true binomial proportions, but degrees of freedom being fairly large mean values may be presumed to conform approximately to large-sample theory. On this assumption, the probability of occurrence of the observed value of the $t$-statistic was obtained in the usual way for each of 379 zones (omitting forty-three zones in which either the number of pairs of grid was only one or two, or where owing to gaps in the field survey information was not available for both grids of the same pair).

296. In 109 out of these 379 zones the probability of occurrence of the observed value of the $t$-statistic was less than 5 %, while the expected number of discrepancies at this level is only nineteen. A scrutiny of the field records showed, however, that in no less than eighty-four cases this could be ascribed to real physical differences. For example, it was found that in fourteen zones the information for the first half-sample was collected at a time when weather was still dry and sowings were seriously retarded, while the information for the second half-sample was collected at a later date after a good deal of rain had fallen and had stimulated heavy sowings; in such cases the later records naturally showed much higher proportions of land under jute. In forty-five other thanas it was found that sufficient rainfall had occurred before the first survey, so that the records showed reasonably high proportions under jute. This was, however, followed by excessive rainfall causing serious damage to the young plants, which materially reduced the proportion under jute in the second set of records. Finally, in the case of twenty-five thanas the first survey had been undertaken before the sowings were completed and the second survey after full sowings, so that the later figures were appreciably higher.

297. The discrepancies in the case of eighty-four thanas may therefore be reasonably ascribed to the influence of real differences in weather conditions during the two periods in which the two surveys took place. Omitting these eighty-four thanas there are 295 left. The actual position is shown in table 25, in which the 295 values of 'Student's' $t$-statistic are grouped in accordance with their respective probabilities of occurrence.

TABLE 25. COMPARISON OF HALF-SAMPLES (A) AND (B): STUDENT'S '$t$' FOR ZONES

| range of probability of occurrence of $t$-values | number of cases | | difference between observed and expected | $\chi^2$ |
|---|---|---|---|---|
| | observed | expected | | |
| (1) | (2) | (3) | (4) | (5) |
| less than 0·01 | 11 | 2·95 | + 8·05 | 21·96 |
| 0·01–0·05 | 14 | 11·80 | + 2·20 | 0·87 |
| 0·05–0·10 | 20 | 14·75 | + 5·25 | 1·87 |
| 0·10–0·90 | 235 | 236·00 | − 1·00 | 0·00 |
| 0·90–0·95 | 12 | 14·75 | − 2·75 | 0·51 |
| 0·95–1·00 | 3 | 14·75 | −11·75 | 9·36 |
| total | 295 | 295·00 | | 34·57 |

298. In only eleven zones the probability is less than 0·01 against the two half-samples being in satisfactory agreement. At this level, however, only three such discrepant values would be expected, which gives us an excess of eight unsatisfactory cases that cannot be

explained by chance causes. It is likely that the work done by either of the two sets of investigators was unsatisfactory in these eight zones.

299. It is worth noting, however, that there were no cases of suspiciously close agreement between the estimates for the two half-samples. In fact, the number of cases with a probability of occurrence of more than 0·95 was only three against an expected number of fifteen. This proves conclusively that the chief object of using the half-sample method was entirely successful, and the two estimates based on the two half-samples were obtained independently.

300. *Precision of the sample survey.* The estimates based on the two half-samples were 1,527,431 and 1,624,706 acres, with a pooled value of 1,576,069 acres. The standard error of the pooled value was estimated to be $\pm 17{,}000$ acres on certain assumptions. The difference between the two half-sample estimates was $97{,}275 \pm 34{,}000$ acres, which means a deviation of 2·86 times the standard error. Judged from the point of view of sampling fluctuations this is not quite satisfactory, and shows that recording errors had not been completely eliminated, so that statistically controlled conditions were not fully established. The divergence between the two half-samples was, however, not large in absolute magnitude.

301. One point is worth noting at this stage. I have already mentioned that each pair of grids belonging to either half-sample was separated by the same constant distance of three-quarters of a mile. This naturally introduced a certain degree of space correlation between two grids forming each pair; and the value of the coefficient of correlation was found to be $+0{\cdot}13$ by direct calculation. By equation (103·1) of Part II, § 3, the estimated variance of the mean value based on both half-samples would be thus equal to $(s_1^2+s_2^2+2s_1s_2\times 0{\cdot}13)/4$, where $s_1^2$ and $s_2^2$ are the observed variances for the two half-samples respectively. When the grids are located independently at random the value of the correlation would be zero, and the estimated variance of the mean value would be $(s_1^2+s_2^2)/4$. The use of linked pairs of grids thus increased the variance by 13 %. This gives a measure of the amount of information sacrificed in order to gain some control over recording mistakes. In view of the satisfactory agreement between half-samples the above policy, however, appears to have been fully justified.

302. *Accuracy of the estimate.* To come back to the question of accuracy, in 1941 an external check, as a complete census of all plots under jute, was carried out by the Government of Bengal for purposes of jute regulation. This official estimate was 1,532,855 acres as published in the *Indian Trade Journal* of 2 April 1942 (**145**, 23). This is less than the sample estimate by 42,214 acres or 2·75 % of the official estimate, which was well within the permitted margin of 5 % as settled by the Jute Census Committee. The discrepancy in terms of the standard error of $\pm 17{,}000$ acres was about 2·45.

303. In assessing the reliability of the sample survey as compared to the complete census it is, however, necessary to remember that the area under jute is not a mathematical quantity which remains constant throughout the crop season. Plots once sown with jute at the beginning of the season may go out of this crop later on owing to drought or excessive rainfall or through damage by pests. The actual area under jute thus fluctuates from day to day. Usually the area would increase up to a maximum fairly early in the crop season and then gradually decrease until the crop is actually harvested. The official estimate refers to the

end of the season, while the sample census being spread over the whole season may be considered to be roughly centred at the middle of the season. On this view the official estimate should be somewhat lower than the sample value, and this is found to be true. In fact the revised official estimate itself showed a decrease of about 90,000 acres as compared to the preliminary official estimate which was published earlier and was based on the first enumeration made during the earlier part of the season. Halving the difference, 45,000 acres or about 3 % of the total area is obtained as a rough estimate of the average physical shrinkage between the middle and the end of the season. This suggests that there exists a residual physical uncertainty of something of the order of say 3 %, so that the accuracy attained in 1941 was fully adequate, and that attempts to go beyond this would serve no useful purpose.

304. To sum up, it was found that, although full statistical control was not established and recording mistakes were not completely eliminated, the accuracy of the sample survey attained in 1941 was sufficient for all practical purposes. The internal precision was high and of the order of 1 %; the margin of error on the basis of comparison with an external and entirely independent estimate was also less than 3 %; while the cost of the sample survey was something of the order of one-fifteenth of the cost of a complete enumeration.

## 7. The exploratory stage

305. The approach adopted in the present paper is especially suited to sample surveys which are carried out each year or at least from time to time. The success of the method depends on securing previous information relating to such quantities as the proportion of land under the crop under survey, the random or non-random nature of the field as characterized by values of the $g$-parameter and relevant information regarding cost of field operations. Where the survey is to be done only once the method of unitary unrestricted or some simple modification thereof would probably be the best course to adopt in practice, but wherever previous information is available or when the work can be spread over a number of years the present approach would be more suitable.

306. The planning of large-scale surveys thus naturally falls into two stages: (*a*) the exploratory, and (*b*) the final stage. The object of the exploratory stage is to study (*a*) the variations of $p$ in different regions for purposes of zoning, (*b*) the variance function, and (*c*) the cost function for deciding the optimum distribution of size and density of grids. If expense is no consideration and a large trained staff is available (an important qualification), then it would be possible to secure all necessary information in the course of a single season along with a general sample survey of the area under question. This would, however, necessarily involve a huge amount of expenditure for experimenting with various sizes and densities of grids, most of which would be of no use in the final stage. In the case of the jute survey the cost of collecting the information in one year was prohibitive and trained workers were completely lacking. Recourse to the exploratory method was thus inescapable.

307. In this method the scale of operations was expanded in the light of the experience gained on the field from year to year. The progress of the survey from year to year can be easily appreciated from tables 26·1–26·3. Beginning with 124 sq. miles in 1937 an area of 59,144 sq. miles was covered in 1941. In the same time the number of grids used increased from 1488 to 57,362. The volume of work increased from 166 man-months in 1937 to 1411

man-months in 1941; and the expenditure grew from Rs. 16,800 to Rs. 1,52,300. While the total expenditure was rising the rate of expenditure per square mile in both time and money was rapidly falling. The total amount of labour required was 1339 man-months in 1937, but decreased to only 24 man-months per thousand sq. miles in 1941; the corresponding rate of expenditure decreased from Rs. 135.48 to Rs. 2.57 per sq. mile in the same period. The decrease in the cost of the statistical portion of the work was relatively more steep.

*Successive stages of the Five-year Scheme*

TABLE 26·1. VOLUME OF WORK

| jute season (year) | area in sq. miles | number of sample units | volume of work in man-months | | |
|---|---|---|---|---|---|
| | | | field | statistical | total |
| (1) | (2) | (3) | (4) | (5) | (6) |
| 1937 | 124 | 1,488 | 64 | 102 | 166 |
| 1938 | 414 | 7,888 | 212 | 183 | 395 |
| 1939 | 2,563 | 12,000 | 531 | 378 | 909 |
| 1940 | 20,553 | 41,345 | 638 | 396 | 1,034 |
| 1941 | 59,199 | 57,362 | 941 | 470 | 1,411 |

TABLE 26·2. VOLUME OF EXPENDITURE

| jute season (year) | expenditure in rupees in round numbers | | | | cost in rupees per man-month | |
|---|---|---|---|---|---|---|
| | field | statistical | overhead and non-recurring | total | field | statistical |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1937 | 6,700 | 10,000 | 100 | 16,800 | 104·69 | 98·04 |
| 1938 | 18,200 | 16,500 | 3,100 | 37,800 | 85·85 | 90·16 |
| 1939 | 47,300 | 36,400 | 6,800 | 90,500 | 89·08 | 96·30 |
| 1940 | 59,600 | 35,600 | 21,700 | 116,900 | 93·42 | 89·90 |
| 1941 | 79,800 | 46,500 | 26,000 | 152,300 | 84·80 | 98·94 |

TABLE 26·3. COST IN TIME AND MONEY PER SQUARE MILE

| jute season (year) | man-months per 1,000 sq. miles | | | expenditure in rupees per sq. mile | | |
|---|---|---|---|---|---|---|
| | field | statistical | total | field | statistical | total |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1937 | 516 | 823 | 1339 | 54·03 | 80·65 | 135·48 |
| 1938 | 512 | 442 | 954 | 43·96 | 39·86 | 91·30 |
| 1939 | 207 | 147 | 254 | 18·45 | 14·20 | 35·31 |
| 1940 | 31 | 19 | 50 | 2·90 | 1·73 | 5·69 |
| 1941 | 16 | 8 | 24 | 1·35 | 0·79 | 2·57 |

308. In planning the work of the exploratory stage it was found that the best method was to try to get quickly some idea, even if very rough, of the physical order of the quantities involved. For example, it matters a great deal to be able to find out as soon as possible whether very small grids of a fraction of an acre or of moderate size of the order of 5 or 10 acres or fairly large grids of the order of 100 acres would be suitable for the purposes in view. A rough idea of the nature of the variance function is useful in planning subsequent stages of the survey. It is possible to get a broad idea of the cost function from even a very small pilot survey, and this is of great help for budgetary purposes.

309. Once such rough ideas are obtained it becomes possible to make plans for the next stage with considerable economy. The expanding method of work gives opportunities for

studying recording mistakes, output and accuracy indices of workers and questions of field organization (such as the composition of field units, arrangements for inspection, etc.) on an experimental basis; and finally makes it possible to build up the necessary human agency by giving training to suitable men—a point of great importance in a country like India. In the exploratory method the whole scheme remains flexible and under complete control, so that changes in programme can be made without any difficulty.

310. I may now say a few words regarding the actual planning of pilot surveys. Results of complete enumeration of fairly large tracts of area are very useful for studying the variance function by model sampling experiments in the Laboratory. In the jute survey this proved most helpful and cleared up many things at a fairly early stage.

311. After some experimentation it was found more convenient and economical to study the variance function and the cost function separately. It has also been found that it is usually possible to split up the cost function into a number of components, some depending on the size but not on the density of grids, some on density but not on size, and others independent of or depending upon both size and density of grids. Usually there is no interaction between different components. In 1938 and 1939 we intentionally used combinations of factors to study this question but found that interaction, if any, was negligibly small.

312. There are broad principles of planning the exploratory survey which often save a good deal of time and unnecessary expense. These are analogous to the design of experiment in certain respects but with important differences. For example, in the case of sample surveys considerations of cost make it impossible to work with an equal number of grids for different sizes or densities of grids. Experiments involving, say, grids of size 1 acre would be naturally much less expensive than a similar experiment with the same number of grids of say size 40 acres. Similarly, cost per square mile would be roughly proportional to the density of grids, so that high densities can only be used sparingly.

313. There are other factors also which introduce elements of asymmetry. For example, there are intrinsic differences in weather and working conditions in different regions or even in the same region at different times of the crop season, and also in previous training, experience or ability of investigators. All these things have to be taken into consideration in planning the exploratory stages of sample surveys, and the different factors have to be balanced as far as possible. General principles of planning are thus gradually emerging which it is hoped to discuss in a later paper.

## 8. Planning of large-scale sample surveys

314. I shall now consider briefly a number of questions regarding the planning of the final stage of a large-scale sample survey. Usually the total cost would be specified by administrators. It would also usually be stated whether the margin of error of the estimate over the whole survey is to be made a minimum or whether the margin of error or the percentage error in different regions was to be kept constant and a minimum. This would decide the type of solution required.

315. *Zones.* The first thing requiring to be settled is the question of zoning. For this a guess has to be made as to the value of $p$ (proportion of land under jute or the crop under

survey) likely to occur in the survey year. Where data for a number of years happen to be available (and this will usually be the case in crop surveys which are continued from year to year) time-series analysis may be of help but must be supplemented by other factors. For example, changes proposed to be brought about under any compulsory or voluntary scheme of regulation must be taken into consideration. The value of $p$ would be naturally settled in the light of all available information.

316. It would often be possible to assume that the shift in different zones would be roughly proportional to the average change in $p$ over the whole area. This was found to be broadly true in the case of acreage under jute. If the total area under jute increases in one season say by 20 %, then in most of the zones the area under jute would increase roughly by the same amount.

317. *Value of g.* Next a guess must be made as to what is likely to be the numerical value of the parameter $g$ in the survey season. In the case of jute it has been seen that the value of $g$ can be obtained with a fair degree of approximation as a linear function of $p$; and the value of $p$ having been already settled it is therefore possible to obtain the expected value of $g$ from the appropriate regression equation.

318. *Form of the survey.* It is next necessary to settle the statistical controls. It must be decided whether the survey is to be made in the form of one single sample or whether in the form of two or more subsamples. As already mentioned, there are various possibilities. There may be two or more interpenetrating but completely independent subsamples, or there may be linked half-samples (A) and (B) as in the case of the jute survey in 1941. Arrangements may also be made for a part or whole of a certain proportion of the grids being enumerated by more than one set of investigators. Having settled these questions the appropriate optimum solution of size-density distribution of grids can be worked out by straightforward numerical-graphical methods.

319. It must be admitted, however, that the actual solution would depend largely upon what values of $p$ and $g$ are considered as most likely to occur in the survey year. This is inescapable. The real question is whether by adopting any other procedure better results are likely to accrue (in the sense of more reliable estimates at lower costs) in the long run, that is, say, over a series of annual surveys. The only procedure independent of zoning and of values of $p$ and $g$ is unrestricted random sampling. (Even in stratified sampling it is necessary to have some knowledge of the zonal values of $p$.) It is thus found that the real choice is between (*a*) configurational sampling on the lines described in the present paper on the one hand, and (*b*) pure unrestricted random sampling on the other hand. From the discussion given in this paper it is, I believe, sufficiently clear that for a series of annual surveys the present approach would be more efficient in working with fields of a non-random type.

320. At the same time it is recognized of course that if the survey is to be made only once, or in case there is no previous information regarding the nature of the field or of the existence of any zonal differences, then the method of unitary unrestricted or the purely random sample survey would be the best course to adopt. But where some previous information is available, or if the survey is proposed to be continued from year to year, then the method of grid sampling should prove more suitable.

321. *Field organization.* Besides the statistical plan it is also necessary to settle the organization of the field units. In large-scale surveys covering fifty or sixty thousand square miles it is impracticable for the same investigator to go over the whole area within the limited time at his disposal. The whole area has to be broken up therefore into a suitable number of blocks and subblocks. In the jute survey it was found convenient to work with subblocks of about 2000 sq. miles, each consisting roughly of from fifteen to twenty zonal units. As the survey was arranged in the form of two interpenetrating subsamples, two parties of investigators worked within the same subblock, and each field party had to cover about 2000 sq. miles.

322. It was also found convenient to have field units each consisting of six, seven or eight investigators in charge of one inspector, and to allot to each investigator a group of villages in such a way that he had to go over a good proportion of the whole subblock. A number of subblocks constituted a block, each of which was placed in charge of a chief inspector; this arrangement was convenient for administrative purposes. The composition of field units, the arrangement of subblocks and blocks, and the other details of field organization are, however, matters which naturally have to be settled to suit local conditions.

323. *Field inspection.* In the jute survey there was one inspector for each group of six, seven or eight investigators who was responsible for the accuracy of the field work done by the investigators under his charge, and was supposed to check a certain proportion of grids from time to time. Over a number of inspectors was placed a chief inspector whose duty was to go round the whole block and check the work done by his men. Besides this there were five field units whose only duty was to go over the whole area and check the field work by paying surprise visits to the different zones. The total proportion of checking which could be done in this way was, however, small, and in practice would be something of the order of say only 1 or 2%. The real value of the inspection system thus does not lie in the actual amount of checking work done on the field, but on the psychological effect it produces on the mind of the investigators.

324. *Flexible control of output.* A special device was adopted for controlling the output f work in the field survey which, as has been seen, differs widely from one investigator to another. The number of grids marked on the maps was intentionally made appreciably larger than the number the investigators were expected to be able to complete on an average every week, but inspectors were instructed to reduce the allotments to suit the capacity of individual investigators. Thus if a particular investigator was unable to complete more than say 70% of the original allotment then his share would be reduced to say 80% in the first instance, and he would be asked to omit all grids whose serial numbers are divisible by 5. As the grids are numbered serially at the time of marking, and as such marking is done in a purely random manner, it is clear that the omitted grids (and hence also those which are retained) would have purely random locations on the maps. The present method would, therefore, preserve in full the random nature of the sample, and yet, provided the inspectors do their work in the proper way, make it possible to secure practically the maximum output from each individual investigator.

325. *Tabulation by subsamples.* The subsample method has also been used with advantage in organizing statistical work. A certain fraction (say one-twentieth or one-tenth or one-

fourth) of the total number of sample grids in each zone is drawn at random and is tabulated in the first instances and the process is repeated as many times as necessary. In practice, instead of working with a single subsample it is usual to work with two subsamples at a time, for which calculations are made by different sets of computers. The advantages of such a method are obvious. First of all it supplies very quickly (with practically one-tenth or one-eighth of the total labour involved) a good approximation to the final estimate. Secondly, a comparison of the results for different subsamples supplies a useful check on computational work. The standard rate of output at different stages of the statistical work can be settled on the basis of subsample records, and necessary changes can then be made in computation arrangements in advance to keep within the scheduled time programme.

326. There are various other devices and methods which have to be developed in organizing and controlling large-scale surveys. The above examples will, however, give some idea of the type of work. The problem, as noted in the introductory section, is something like an engineering project in which abstract theory and concrete statistical methods as well as organizational and administrative work all play their part.

*Effect of changes in cost function*

327. I may now consider in general terms the effect of changes in the cost function on the sampling technique. Any shift in the general price level or the purchasing power of money will, of course, leave the optimum solution undisturbed for any assigned real cost. The nominal or money level of the cost would change, but this merely means a relabelling of the money scale of costs and needs no further discussion.

328. A differential change in the cost of field and statistical work, on the other hand, may produce certain distortions. The magnitude of the disturbance would depend on the amount of the differential change, and when this is small the optimum solution would remain more or less steady. Fortunately, large differential changes are not likely to occur in normal circumstances, at least this was our experience in the jute survey during the period 1937–41. The cost in rupees per hour for field and statistical work is given in col. (4) and (5) of table 19. Using the total number of hours of work in 1941 and the minimum rate for field work together with the maximum rate for statistical work and vice versa, it was found that the maximum change in total cost (keeping hours of work constant) was something like 3 %. Remembering that this was a period of abnormal and even violent price changes in India it appears that differential changes in the cost of field and statistical work are not likely to be large enough to introduce serious complications, and may be ignored for all practical purposes.

329. The effect of changes in the parameters of the cost function from year to year will now be considered. In general this would affect the optimum solution for the distribution of size and density of grids. But here also it was found that for the three years 1939, 1940 and 1941 (when working conditions were changing abruptly from year to year) a joint fitting of the variable terms gave quite satisfactory results, and that the form of the cost function was steady. In the jute survey the differences in the constant terms themselves were not large, so that the disturbing effect also was comparatively small.

330. But suppose the effect had been larger, would that have justified abandoning the present approach to the sampling problem? This is the real issue. In the jute survey work was done with an *ad hoc* field staff recruited afresh every year. But suppose a permanent staff had been employed (as I wanted), then the efficiency of field work would have improved and the cost of operations would have decreased from year to year. If the money cost is kept the same and there is no general change in the purchasing power of money (or if the real money cost is kept constant), then owing to improved efficiency of work the optimum solution would change from year to year, passing from one appropriate to a lower labour-cost level to one appropriate to a higher labour-cost level. Using the flexible control of output (which I have already described), this would increase the precision of the final estimate without any increase in the money cost.

331. But suppose no further increase in precision is required. It should then be possible to reduce the money cost, but this can only be done by making a guess of the numerical values of the cost parameters. Just as in the case of zonal values of $p$- and the $g$-parameters, the crux of the problem is the possible (or even likely) time shifts in the values of the cost parameters.

332. A stationary field (in which the field parameters $p$ and $g$ as well as the cost parameters remain steady) is an ideal situation for which a critically sharp optimum solution can be obtained. In any real situation conditions are, however, not steady but changing. A dynamic element is thus introduced which is inherent in the very nature of things. Therefore, an estimate of the appropriate values of $p$, $g$ and the cost parameters which are likely to occur in the survey year must be made, and on this basis plans prepared. If the estimates are good, solutions will be obtained which are very near the true optimum; if not, the solution adopted would differ appreciably from the true optimum. The risk has to be taken; but provided there is a rational basis for estimating the field and cost parameters, then, in the long run, the present method should turn out to be more economical. Also, of course, with the gradual accumulation of experience, it should be possible to develop statistical methods for forecasting the values of $p$, $g$, and the cost parameters on the basis of available knowledge relating to previous years. In other words, as in the case of all problems involving time shifts, it is necessary to investigate the dynamic aspects of the fluctuations in the values of the parameters. This is closely linked, for example, with one question of considerable practical importance. Should all the grids be selected afresh every year? Or should a certain proportion of old grids be used from year to year? This is being studied at present and will be discussed in subsequent papers after a sufficient volume of material has accumulated.

### *Concluding remarks*

333. From the account given above it will be seen that the general lines of the sampling technique appropriate for the estimation of crop areas have become fairly clear. The practical usefulness of the method described here has been demonstrated on a country-wide scale. Fresh theoretical problems (which may be broadly described as space generalizations of the theory of sampling distributions and estimation) have been opened up. Much, however, still remains to be done, and active work is proceeding on both theoretical and

experimental lines in Calcutta on both uni-stage and multi-stage sampling. In conclusion, I should again like to emphasize the essentially co-operative nature of the undertaking, and the fact that the credit for the advances already made belongs to the large group of my fellow-workers in the statistical as well as field branches of the project.

## BIBLIOGRAPHICAL NOTE

(*Added* 15 *January* 1943)

334. I have already mentioned that, at the time of writing the paper, I did not have access to the Statistical Library which had been removed from Calcutta as an evacuation measure. Recently, I had the opportunity of looking up a number of papers during a visit to Giridih where the books are stored at present. I did not have time for an exhaustive search, and I shall refer in this Note to such papers only as were readily available, and to such portions of these papers only as have any bearing on the topics discussed in the present paper.

335. Recent work on the technique of the sample survey may be considered to have started with the *Report on the Representative Method in Statistics* prepared by a Commission appointed in May 1924 and presented before the International Institute of Statistics in 1926, together with a memoir on theoretical aspects of the subject by A. L. Bowley. This report described two methods, the first of which was called 'inquiry by random selection' in which 'we are concerned with a definite population or universe of persons or things to all of which we have access, and that we make a selection at random of some of these persons or things, in such a way that every unit in the universe has an equal chance of being selected, while the method of selection is completely independent of the characteristics to be examined' (p. 364). This is what has been called the unitary unrestricted type of sampling.

336. The principle of zoning was also explicitly stated: 'In the case of heterogeneous population some additional security can be obtained if before making our choice of units to be observed, we divide the population into more homogeneous groups (e.g. into urban and rural communes or in a town into different districts or parishes), and then select at random the same proportion from each of these groups' (p. 365). This was called 'the method of stratification' on p. 5 of Bowley's memorandum, and corresponds to what has been called zonal unrestricted sampling in the present paper, with, however, a distribution of sampling units purely proportional to the total number of units in each stratum or zone.

337. The second was the method of purposive selection. 'In the foregoing we have taken it for granted that we make up our sample by selecting a *number of units* at random in such a way that every single unit has an equal chance of inclusion. In many cases, however, it will be possible to save time and labour by proceeding more summarily. Instead of making up the sample of units one will form it out of *groups of units*' (p. 367). This resembles to some extent what has been called configurational sampling with, however, important differences. The groups of units are not chosen at random but in such a way as is assumed 'to give the sample the same characteristics as the whole' (p. 368), which explains the name purposive selection.

338. Next there is a very important paper written by Jerzy Neyman in 1934 which contains a great deal of matter of both theoretical and practical importance, but I shall confine my references to only those points which have immediate relevance.

339. Neyman developed the idea of stratified sampling not merely by individuals but also by groups. He pointed out that especially in the case of human populations the individuals are almost invariably grouped. 'These groups are then grouped again and again' (p. 568). 'If there are enormous difficulties in sampling individuals at random these difficulties may be greatly diminished when we adopt groups as the elements of sampling' (p. 569).

340. Bowley had distinguished between purposive selection and random sampling in the following way: 'In purposive selection the unit is an aggregate, such as the whole district, and the sample is an aggregate of these aggregates, while in random selection the unit is a person or thing which may or may not possess an attribute or with which some measurable quantity is associated' (p. 570). Neyman pointed out that the fact that 'the elements of sampling are not human individuals but groups of these individuals, does not necessarily involve a negation of the randomness of the sampling' (p. 571). He therefore considered this to be a 'special type of random sampling by groups', and did not think that the nature of the elements of the sampling should be 'considered as constituting any essential difference between random sampling and purposive selection' (p. 571).

341. Neyman's method of stratified sampling by groups resembles but is different in one respect from what has been called the zonal configurational type. In the present paper discussion has been restricted to what has been called the overlapping system of grid sampling (paragraph 60), while Neyman uses what has been called here the system of 'exclusive' sampling units in the footnote to the same paragraph. The essential distinction is this: In the present method the same basic cell or the same individual may form a part of more than one grid or sampling unit; in Neyman's method, on the other hand, the same individual or the same basic cell cannot be included in more than one sampling unit.

342. In deriving the best solution Neyman assumes that the total number of sampling units drawn would be kept the same. This amounts to the assumption that the total cost of operations is simply proportional to the total number of sampling units. Neyman, however, does not explicitly consider the cost aspects of the problem.

343. In the discussion which followed, R. A. Fisher gave a concise description of the basic concepts which are essentially similar to those used in the present paper. In Fisher's language 'the smallest unit that need be considered, the unit of *measurement*, as it might be called, consisted, in the case of a cereal crop, of, perhaps, 10 inches or 24 centimetres measured along a drill row. Again it might consist of a single plant, as with potatoes or sugar beet' (p. 615). This clearly corresponds to what I have called the basic cell or the quad. Next: 'A number of units of measurements, usually four, fixed in a relative position, but not necessarily adjacent, constituted a *sampling unit* which would, therefore, contain in all one metre length of drill row, taken however, in practice, from four different rows. Since the parts of a sampling unit were fixed in a relative position, the positions of all were determined simultaneously by a single act of random sampling, i.e. by the choice, by a

physically random process, of the particular sampling unit used from among all those available in the *sampling area*' (p. 615). Fisher's sampling unit is what in the present paper has been called either a sampling unit or more concisely a grid. His sampling area is simply called here the field or zone under survey.

344. Fisher pointed out that 'it was essential that there should be at least two independently located sampling units in each sampling area, since it was from the differences between these, or the variance among them, if they were more than two, that the error of sampling was estimated. The variance among the units of measurement within the same sampling unit served a different and subsidiary purpose. It was essential to the study of what structure or size the sampling unit should have, and by analysing the variance within and among sampling units, one could ensure that the sampling units were so chosen as to give the maximum precision in return for the labour expended' (p. 615). Fisher further pointed out that 'the error of random sampling, on the other hand, should be ascertained with high precision from every experiment to which the sampling method was applied, for on it one relied for judging the *number* of sampling units which could with advantage be taken from the growing crop' (p. 615). This is exactly the point of view adopted in the present paper. In fact the fundamental criterion used in obtaining the optimum solution is identical with that stated by Fisher, namely, 'maximum precision in return for the labour expended', which is conveniently measured in practice in terms of the total cost.

345. Since the publication of Neyman's paper a number of other papers have been published on this subject by Neyman (1938), Stephan (1939), Hansen & Harwitz (1942), and other workers dealing, however, mainly with investigations relating to human populations, although the theory has had applications in other fields as well. Besides a rigorous deduction of the various formulae Neyman's theory covers a good deal of ground which is somewhat different from the topics discussed in the present paper and need not therefore be considered in greater detail here.

346. The mathematical theory has been given in a concise form by Allen T. Craig (1939) in a paper 'On the mathematics of the representative method of sampling', which gives a clear exposition of Markoff's method on which Neyman's work is based. The 'best' solution is defined by two conditions, namely, that (*a*) it should be a linear unbiased estimate with (*b*) minimum variance. The total number of sample units being kept constant is an implicit condition involved in the optimum solution.

347. In Neyman's method of stratified sampling (or what in the present paper is called zonal unrestricted sampling) the optimum solution depends on a knowledge of the true standard deviation of each zone. Sukhatme (1935) investigated the effect of estimating the standard deviation of the different zones by a preliminary inquiry, i.e. the effect of using $s_k$ instead of $\sigma_k$ for the $k$th zone.

348. The use of sampling methods for estimating the yield of crops has received a good deal of attention in recent years. Estimation of yields by sampling of individual plots of replicated experiment of cereals had been practised since 1929 at Rothamsted and at associated centres. A comprehensive paper on the subject was published by Yates & Zacopanay in 1935. This belongs essentially to multi-stage sampling and is beyond the

scope of my present paper, but I may note that 'grid' has the same meaning as 'sampling unit' as used by Yates & Zacopanay as well as Fisher; Wishart & Clapham (1929) also use 'sampling unit' in the same sense. In figure 1 of p. 553 of Yates & Zacopanay's paper various types of sampling units are shown which are examples of grids of different patterns in our language.

349. In 1938 a special discussion 'On crop estimation and its relation to agricultural meteorology' was opend by J. O. Irwin, and a number of other workers participated in it. In this paper I find a reference to the use of the sampling method in U.S.A.: 'Forecasts of acreage are made partly subjectively by asking reporters to express the acreage as a percentage of the usual acreage, partly by a sampling process of securing reports from individual farmers who report the acreage on their farms under each crop in current and previous years, and partly by field-count methods such as counting by means of a special speedometer the number of feet along the road under each kind of crop' (p. 4).

350. W. G. Cochran's contribution to the discussion has certain points of interest in the present connexion. Following Yates (1936–7 *a*, *b*) he divided the problem into two parts 'according as the estimates are required before harvest or at harvest', and emphasized the value of the sampling method 'both for determining the yield of a field and of a district'. 'Preliminary research would be needed to develop a good sampling technique, to train observers in its use and to assess the amount of time and labour required to estimate the mean yield of a country with a given degree of accuracy' (p. 13). Cochran mentioned that in the case of wheat the coefficient of variation between fields was roughly of the order of 22%, from which he concluded that 'an estimate of the mean yield of the country with standard error of only 1% could be obtained by sampling 500 fields' (p. 19). All these questions, however, refer to what I have called multi-stage rather than the uni-stage sampling and need not be further considered here.

351. Cochran discussed in 1939 the use of analysis of variance in crop estimation on the lines of Yates & Zacophanay but with a good deal of fresh material and suggestive observations. As regards nomenclature, Cochran used 'subdivisions' for Neyman's 'strata' and 'zones' of the present paper. Cochran points out that Neyman's term 'stratum' has a definite geographical flavour, while 'subdivision' is more neutral. This is quite true, but I have not considered it advisable to use the word subdivision in the sense of zone or stratum, as a subdivision is a recognized and familiar administrative unit in India; subdivisional headquarters, S.D.O. for subdivisional officer, etc., are in everyday use.

352. I have already mentioned that Wishart & Clapham (1929), Yates & Zacopanay (1935), Fisher (1934) and Cochran (1939) used the word 'sample unit' in the same sense in which both 'sampling unit' and 'grid' have been used here. In 1939, Cochran, however, uses the word in a modified sense which has an important bearing on the terminology adopted by me. He writes: 'the sampling unit need not be measured completely; it may itself be enumerated by subsampling. A common example of the method occurs in the estimation of the yield of field crops. Here the sampling unit is usually a single field, the yield of the field being estimated by taking several small samples from the field, instead of by harvesting the entire crop in the field. From the point of view of the analysis of variance this method might be regarded as a case of incomplete subdivision; the material is grouped by

fields, but not all fields are sampled, so that the sampling error consists partly of variation between fields and partly of variation between subsamples within fields' (pp. 493–494).

353. It will be noticed that Cochran is adhering to the scheme of uni-stage sampling inasmuch as the sampling units are selected by one single act of randomization. He then contemplates the estimate for the selected sampling units being obtained by a number of subsamples, and calls this 'incomplete subdivision'. For present purposes, it is preferable to call this two-stage sampling. In the present terminology sampling unit has a somewhat generalized connotation. In multi-stage work there is, in fact, a hierarchy of sampling units at successive stages. The units selected in the first stage of randomization may be called sampling units of the first stage; within each such selected sampling unit smaller sampling units are selected by a second act of randomization, and these are sampling units of the second stage (these are called subsamples by Cochran); within each second-stage sampling unit smaller sampling units are drawn at random by a third act of randomization, and so on. Present terminology is thus more general and more convenient for dealing with this type of sampling.

354. Cochran further writes: 'An important practical consideration in subsampling is the division of resources between the amount of subsampling per sampling unit and the number of sampling units chosen. For a given expenditure, one can only be increased at the expense of the other. The best compromise will depend on the relative costs of increasing the number of subsamples per sampling unit and of increasing the number of subsamples per sampling unit, and on the relative increase in accuracy obtained by such changes' (p. 494). Cochran also discussed what has been called the size and pattern of grids. He writes: 'the size and structure of the sampling unit plays an important part in determining the accuracy of sampling. In an areal survey for instance, the choice may lie between a section, a square block of four sections, or a township. Here again the problem is to strike the most effective balance between the amount of work and the statistical efficiency, since for a given percentage sampled, a few large sampling units are usually less expensive to collect than a large number of more widely scattered small sampling units' (p. 494). This amounts to the joint use of cost and variance functions for obtaining optimum solutions as advocated in the present paper. Cochran further observes that a thorough study of the relative effectiveness of different sizes of unit cannot as a rule be made without a special investigation for that purpose. This is what has been called the exploratory method. Cochran has also referred to what has been called recording errors: 'Errors in counting are bound to occur in any large-scale investigation, and though they are not usually differentiated from the sampling errors they will contribute to inaccuracy in any means which are calculated' (p. 507), but has not discussed it any further.

355. Cochran has also given considerable attention to what he calls double sampling. 'The first is a large sample, in which the second character alone is enumerated, and the second a small sample in which both characters are usually enumerated' (p. 495). This method has been used in the Calcutta Statistical Laboratory in the case of crop yields for several years. For example, in the case of jute, the weight of green plants is collected on a large scale, and the weight of green plants together with the corresponding weight of dry fibre after retting in a comparatively small number of cases. In the same way in the case of paddy the crop is weighed on an extensive scale immediately after harvesting, while

only in a small proportion of the samples obtained in this way the grains were weighed again after drying.

356. I may note here that Cochran's paper was presented before the American Statistical Association on 29 December 1938, and was published in June 1939. As already noted the present work commenced in 1937, the first report was ready by July 1938, and a paper was presented before the Lahore session of the Indian Science Congress on 5 January 1939, or only a week later than the date on which Cochran read his paper. The work in U.S.A. which Cochran describes and the work in India were thus developed independently. There were, however, striking points of resemblance which only show that the procedure adopted in both countries were on sound and therefore parallel lines.

357. I must now refer to an important paper by Fairfield Smith published in 1938. Although this does not deal with large-scale sample surveys it contains many ideas which are intrinsically similar to those adopted in the present paper. For ordinary field experiments he reached the same form of the variance function as used by me. He found that 'in general the regression of the logarithm of standard deviation on size of plot is substantially linear' (p. 9). His work covers the yield of a wide variety of crops like wheat, maize, sorghum, mangolds, beets, potatoes, sweet potatoes, soybeans, pineapples, natural pastures, oranges, lemons, walnuts and apples grown in Australia, Hawaii, U.S.A., Egypt and England. Excepting only a small number which had too few plots to give reliable results or in which data were not available in a suitable form for his purposes, Fairfield Smith included in his investigation the results of all published work available at the time in Canberra library. The value of what he calls the $b$-coefficient (here called the $g$-coefficient) was less than 1 in every case of crop yield considered by him; roughly half the values occurred between 0·40 and 0·55.

358. Secondly, he explicitly used the method of cost function to obtain the optimum solution. He assumed the cost of operation per plot to be of a simple linear form $k_1+k_2(x)$, where $(x)$ is the size of the plot. He also investigated the cost of using plot sizes other than the most efficient, and gave numerical examples. This is a good example of the cost-variance methods used in the present paper.

359. Thirdly, he discussed the effect of correlation between adjoining plots on what has been called the variance function. He noted that 'since the regression of variance on plot size is a function of the correlation of adjacent areas, it appears theoretically inevitable that the shape of plot should have some effect, since the correlation of ends of long narrow strips must usually be less than that at the opposite sides of a square of equal area...it might be worth while to make use of two $b$-coefficients [which we call $g$-coefficients] with assigned directions to describe the heterogeneity' (p. 22).

360. Fairfield Smith refers to two papers of J. A. Harris (1914, 1920) in which he had demonstrated that 'a hypothesis of zero correlation between adjacent areas cannot be regarded as probable' (p. 2). In fact, Harris appears to have proposed using the intra-class correlation coefficient of yields from adjacent areas as a coefficient of heterogeneity (p. 3).

361. It would be seen that the basic ideas underlying Fairfield Smith's work are similar to those adopted in the present paper, although the actual spheres of application are widely different. But here again the work was done independently. Although Smith's paper had been received for publication in December 1936 it was actually published in

January 1938 and reached India several months later. In the meantime, working on a widely different problem, namely, estimating the acreage under the jute crop in Bengal, a technique had been developed which involved essentially similar basic concepts.

362. Finally, there is a recent paper by Snedecor & King, published in the *Journal of the American Statistical Association* in March 1942, which reached India in August 1942, in which an account is given of the sample survey work on crop acreage done in the United States which appears to be very similar to the work done in India. The name 'grid' is used for sampling unit as in this paper; and the variance function and the cost function are called by the same names and are jointly used for obtaining the optimum solution as in the present work. The work in the United States was started a little later than the work in Bengal but was probably developed quite independently. Some of the earlier reports on the jute census work were sent to the U.S.A., but I am unable to say whether these had any influence on the work there. From the broader scientific point of view the question of priority is not of any importance; and what is really gratifying is to find that work in both countries appears to have been developed on parallel lines. From the point of view of India this is certainly encouraging.

363. I have already referred to extensive work on time series which corresponds to space distributions in one dimension. H. T. Davis has recently given a full account in a book on *Analysis of Economic Time Series*; as the problem discussed here is essentially two-dimensional in character further details are unnecessary.

364. As regard what has been called associated space distributions in two dimensions, I have not been able to find in the short time at my disposal much of earlier work. Kendall in 1941, in a paper in *Biometrika*, distinguished between samples drawn one at a time and also drawn in a block of $n$ units which he calls a 'clutch'; but he has not developed the subject in greater detail.

365. H. Todd in a 'Note on random association in a square point lattice' (1940) calculated certain probabilities which have some bearing on the present method of 'patch number'. He visualized the field as consisting of 'points' (similar to square cells), and calculated the probability of occurrence of doublets, triplets, and quadruplets (similar to patches of 2 cells, 3 cells, and 4 cells, in contact with one other along a line or at a point, that is, recognizing diagonal contact, as used here). This, however, is not enough to determine the patch number in two dimensions.

[*Note added in proof on* 24 *July* 1944]

366. I must refer to a paper by J. A. Hubback on *Sampling for Rice Yield in Bihar and Orissa* published in 1927 (Bulletin No. 166 of the Agricultural Research Institute, Pusa, Government of India) which is of great importance in the present connexion. Unfortunately, this paper had got mislaid at the time of evacuation of the Statistical Library from Calcutta and could not be traced at the time I was writing the Bibliographical Note.

367. It is interesting to note that as early as at least 1923 Hubback had been convinced of the need of using random samples for crop estimation work. After criticizing methods which were at that time (and in India still are) in existence he wrote:

> The only way in which a satisfactory estimate can be formed is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal element

of the experimenter, but also makes it possible to say what is the probability that the result of a given number of samples will be within a given range from the true mean (p. 4).

In 1923 he actually obtained 400 random samples of cuts of paddy in a part of Santal Parganas district in Bihar. As far as I know this was the first occasion on which the random sampling method was used in crop-cutting work in India. Similar work was continued by Hubback in 1924 and 1925, results of which are discussed in considerable detail in the present paper.

368. At that time the size of the cut in crop-cutting work on paddy was usually about 1/10 of an acre, Hubback however deliberately used very small cuts 1/3200 of an acre (in the form of an equilateral triangle) for reasons which he explained in clear language:

It is no advantage to take a large number of samples from places very close together, where the crops will naturally be very much the same on the same day. The degree of accuracy is not seriously improved by such practice. This explains why there is no need to take large samples instead of the handy samples obtained by my method. A sample of one-tenth of an acre is merely 320 of my samples taken together in juxtaposition. It simply gives a determination of the mean yield of that particular field, which is not more effectively accurate than that given by say four small samples. Even four samples instead of one are not worth while, because in the great majority of cases they do not differ among themselves enough to affect the mean or the standard deviation of the whole set of samples. This may be illustrated from the columns for 'all classes' and '1st cutting' for Santal Parganus, Godda thana 1924. Technically speaking, there is very high correlation between the individuals of such groups of samples, which makes the ordinary rule, that the standard deviation divided by the square root of the number of samples, quite inapplicable. (p. 9.)

This appears to be the earliest recognition of the variance function having a non-normal form in crop-cutting work, and ante-dates Fairfield Smith's paper (1938) by eleven years.

369. Hubback's paper is also remarkable in giving the earliest description of an attempt to use the linear sampling method in 1925 to estimate crop acreage:

The plan was to make the sampler march from centre to centre across country as near as he conveniently could in a straight line. After certain intervals of time he had to count 100 paces and note how many of these ended on harvested rice land....Each count of 100 would give a definite percentage, in many cases 0, in many nearly 100, the difference from the full 100 being due to the presence of field ridges. The mean of these percentages would give the percentage of harvested rice land in the total area of the tract sampled. (p. 10.)

Not only this. Hubback actually proceeded to investigate whether the result would be sufficiently accurate to justify the labour consumed:

Six men were employed for nearly a month each and between them made 1971 counts. From these a mean of 47·22 per cent. was obtained. The standard deviation was 37·10. This means that it is about 21 to 1 that the true percentage is between 49·4 and 45. The accuracy obtained is hardly sufficient to justify the employment of six men for a month, in each district. Further the work is strenuous and tedious, and it is probable that it would be in practice shirked, and results fudged. The difficulty of excluding large patches of thick jungles or other inaccessible country both from the sampling and the sampled area has yet to be overcome. But it is still possible that some method on similar lines may prove practicable. (p. 10.)

370. Hubback also gave time-estimates of the labour which would be required for attending a mean yield per acre "correct within one maund in about 95 per cent. of cases." The basic concepts of both the cost and the variance functions thus occur in an incipient form in his paper much earlier than any paper cited here.

## References


Bowley, A. L. 1926 Memorandum (pp. 1–62) incorporated in the Report on the Representative Method in Statistics. *Bull. Inst. Int. Statist.* **22**.

Bowley, A. L. 1934 Discussion on paper by J. Neyman, 'On the two aspects of representative sampling'. *J. R. Statist. Soc.* **97**, 558–625.

Cochran, W. G. 1939 The use of the analysis of variance in enumeration by sampling. *J. Amer. Statist. Ass.* **34**, 492–510.

Craig, A. T. 1939 On the mathematics of the representative method of sampling. *Ann. Math. Statist.* **10**, 26–34.

Davis, H. T. 1941 *The analysis of economic time series*, xiv + 620 pp. Indiana: Principia Press.

Fisher, R. A. 1934 Discussion on paper by J. Neyman, 'On the two aspects of representative sampling'. *J. R. Statist. Soc.* **97**, 558–625.

Hansen, Morris H. & Harwitz, William N. 1942 Relative efficiencies of various sampling units in population inquiries. *J. Amer. Statist. Ass.* **37**, 89–94.

Harris, J. A. 1914 *Amer. Nat.* **49**, 43–55 (referred to by Smith 1938).

Harris, J. A. 1920 *J. Agric. Res.* **19**, 279–314 (referred to by Smith 1938).

Hubback, J. A. 1927 *Sampling for Rice Yield in Bihar and Orissa.* Bulletin No. 166, Agricultural Research Institue, Pusa, Government of India.

Irwin, J. O. 1938 Crop estimation and its relation to agricultural meteorology. *J. R. Statist. Soc.* **5**, Suppt. 1–45.

Kendall, M. G. 1941 A theory of randomness. *Biometrika*, **32**, 1–15.

Neyman, J. 1934 On the two aspects of representative sampling. *J. R. Statist. Soc.* **97**, 558–625.

Neyman, J. 1938 Contributions to the theory of sampling human populations. *J. Amer. Statist. Ass.* **33**, 101–116.

Report on the Representative Method in Statistics 1926 *Bull. Inst. Int. Statist.* **22**, 359–451.

Smith, H. Fairfield 1938 An empirical law describing heterogeneity in the yields of experiments. *J. Agric. Sci.* **28**, 1–23.

Snedecor, George W. & King, Arnold J. 1942 Recent developments in sampling for agricultural statistics. *J. Amer. Statist. Ass.* **37**, 95–102.

Stephan, F. 1939 Representative sampling in large-scale surveys. *J. Amer. Statist. Ass.* **34**, 343–352.

Sukhatme, P. V. 1935 Contributions to the theory of the representative method. *J. R. Statist. Soc.* **2**, Suppt. 253–268.

Todd, H. 1940 A note on random association in a square-point lattice. *J. R. Statist. Soc.* **7**, Suppt. 78–82.

Wishart, J. & Clapham, A. R. 1929 *J. Agric. Sci.* **19**, 600–618 (referred to by Yates & Zacopanay).

Yates, F. 1936–7*a* Applications of the sampling technique to crop estimation and forecasting. *Trans. Manchr Statist. Soc.* pp. 1–26.

Yates, F. 1936–7*b* Crop estimation and forecasting: Indications of the sampling observations on wheat. *J. Minist. Agric.* **42**, 156–162.

Yates, F. & Zacopanay, I. 1935 The estimate of efficiency of sampling with special reference to the sampling for yield in cereal experiments. *J. Agric. Sci.* **25**, 545–577.


*Note.* A list of publications from the Statistical Laboratory, Calcutta, is given in paragraph 8 on pp. 331–332.

PRINTED IN GREAT BRITAIN BY W. LEWIS, M.A., AT THE CAMBRIDGE UNIVERSITY PRESS