TABLE 33

Calcutta Public Preference Survey, 1941. *Inter-group marriages by educational groups*

Type of inter-group marriage	Educational groups	Total number	Percentage of			
			Unconditional support	Conditional support	Unconditional opposition	Indifferent
(1)	(2)	(3)	(4)	(5)	(6)	(7)
(*a*) Between sub-castes	Non-matrics	367	19·35	27·79	38·42	14·44
	Undergraduates	658	31·00	27·66	30·55	10·79
	Graduates	436	44·50	30·96	16·06	8·48
	Total	1,461	32·10	28·68	28·20	11·02
(*b*) Between castes	Non-matrics	367	17·17	22·88	45·78	14·17
	Undergraduates	658	23·56	25·84	40·42	10·18
	Graduates	436	38·53	30·05	22·48	8·94
	Total	1,461	26·42	26·35	36·41	10·82
(*c*) Within same *gotra*	Non-matrics	348	12·64	15·23	54·89	17·24
	Undergraduates	635	16·70	20·94	51·18	11·18
	Graduates	409	26·40	21·03	41·08	11·49
	Total	1,392	18·53	19·54	49·14	12·79
(*d*) Between provinces	Non-matrics	348	10·92	22·99	53·16	12·93
	Undergraduates	635	15·91	27·87	47·87	8·35
	Graduates	409	26·65	30·32	32·27	10·76
	Total	1,392	17·82	27·37	44·61	10·20
(*e*) Between communities (Hindus and Muslims)	Non-matrics	367	8·17	10·90	67·03	13·90
	Undergraduates	658	9·57	14·13	66·42	9·88
	Graduates	436	22·25	19·50	46·55	11·70
	Total	1,461	13·00	14·92	60·64*	11·44
(*f*) Between nationalities	Non-matrics	367	6·27	9·53	67·85	16·35
	Undergraduates	658	10·03	17·17	62·16	10·64
	Graduates	436	19·50	21·79	45·64	13·07
	Total	1,461	11·91	16·63	58·66	12·80

TABLE 34

Calcutta Public Preference Survey, 1941. *Religious instruction in colleges by educational groups*

Educational group	Number of persons				Percentage of persons			
	In favour	Not in favour	Indifferent	Total	In favour	Not in favour	Indifferent	Total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Non-matrics	288	22	87	397	72·54	5·54	21·92	100·00
Undergraduates	422	116	125	663	63·65	17·50	18·85	100·00
Graduates	179	144	115	438	40·87	32·88	26·25	100·00
Total	889	282	327	1,498	59·34	18·83	21·83	100·00

Model Sampling Experiments

I must refer to another special feature of the Institute—namely, the extensive use of model sampling experiments in applied work. For example, in the Bengal crop survey, information

was collected about the crop grown on each individual field for large tracts, and these were recorded on maps. It was then possible to carry out, in the Statistical Laboratory, a large variety of model sampling experiments using grids of different sizes and densities. Such studies proved most helpful in developing the actual sampling technique.

In the case of space distributions, it has been shown elsewhere that the patch number supplies a useful concept. Consider a map of, say, an agricultural area with fields which are assumed, for the sake of simplicity, to be of the same size. Suppose each field which is shown with paddy is coloured yellow, and fields without paddy are left blank or white. The map of any such region would be then broken up into a number of yellow and white patches, depending on the actual distribution over space of fields under paddy. If alternate fields in both directions are sown with paddy, then the pattern will be something like that of a chessboard; and in this case the total number of patches will be equal to the total number of fields. On the other hand, if half the whole area is sown with paddy in one compact bloc and the other half is left without paddy, then the patch number will be simply two. In any actual situation the patch number will lie between these two limits. If the fields to be sown with paddy (or to be marked yellow) are selected purely at random, then the patch number would show a characteristic distribution.

It is obvious that the design of the sample survey can be made more efficient if something is known about the distribution of the patches. For example, if the patch-pattern is of the chessboard type, then it will be quite sufficient to explore (in the ideal limit) only two adjoining cells to secure a complete picture. If the patch number is only two, then the best plan would be to try to settle the boundary between the two heaps. In any case, when there is any appreciable tendency towards patterning, it is often possible to have recourse to configurational (which is sometimes called " systematic ") sampling.

For two colours (*i.e.*, for a binomial distribution) in one dimension the theoretical solution is known. For more than two colours in one dimension, or for two or more colours in two dimensions, complete theoretical solutions have not yet been given. Raj Chandra Bose is carrying on mathematical researches on the sampling distribution of the patch number in the case of a binomial field in two dimensions, and has obtained a number of useful results.

Investigations are also proceeding on purely experimental lines under the leadership of Jitendra Mohan Sen Gupta. A number of different sets of fields each of size 100 × 100—*i.e.*, consisting of 10,000 cells arranged in the form of a square—were constructed and filled with numbers from 0 to 9 at random (or, to be more strict, what are believed to be arranged at random) in two dimensions. Some were based on well-known tables of random numbers. A large number were produced mechanically with the Hollerith equipment by the British Tabulating Machine Co., Ltd. By assigning different digits to different colours, it is possible to prepare random distributions of patches on these fields and to study the patch number experimentally. Work is being done on a large scale, and has already yielded results which, it is believed, would supply useful information for the guidance of practical work, as well as valuable clues for mathematical researches.

I shall give one concrete example. Consider an $n \times n$ field. Let p be the probability for a cell to be black, and q to be white ($p + q = 1$). In counting black patches contact is recognized only along sides, but in counting white patches contact is permitted both along sides and at corners. Let B be the number of black patches and W the number of white patches (as defined above), but embedded within black patches. R. C. Bose has given the following results:

$$\text{Expectation } E(B - W) = p + 2(n-1)pq + (n-1)^2(pq^2 - p^2q)$$

$$\text{Variance } V(B - W) = n^2pq - 4(n-1)p^2q - (6n^2 - 10n + 4)p^2q^2 - (6n^2 - 40n + 50)p^3q^2 - (14n^2 - 56n + 54)p^3q^2 + (32n^2 - 104n + 84)p^4q^2 - (9n^2 - 30n + 25)p^4q^4$$

The results of one series of model sampling experiments is given in the following two tables, 35 and 36. The observed average number of (black minus embedded white) patches and the expected number calculated by the above formula are shown in cols. 2 and 3 of Table 35 for various values of p. The difference between the observed and expected number divided by the standard error of the difference (based on the theoretical value of the variance as calculated from the formula given above) is shown in col. 4. Although the distribution may not be strictly normal in the present case, it is likely that the figures in col. 4 may behave approximately as normal deviates. It will be noticed that out of thirty-six comparisons no fewer than ten are significant at 1 per cent. level and three at the 5 per cent. level.

TABLE 35

Model sampling experiments. Comparison of expected and observed average number of (black minus embedded white) patches in a binomial field

p	Average number		Normal deviate	Average number		Normal deviate
	Observed	Expected		Observed	Expected	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	10 cells × 10 cells (n = 1000)			20 cells × 20 cells (n = 250)		
0·1	7·99	8·21	−3·07 †	31·54	32·44	−0·35
0·2	12·81	12·93	−1·56	49·78	50·18	−1·34
0·3	14·44	14·46	−0·32	54·83	54·52	0·95
0·4	13·30	13·21	0·32	48·34	47·67	1·76
0·5	10·28	10·06	2·17 *	33·62	32·56	2·51 *
0·6	5·29	5·70	−4·09 †	11·97	13·19	−2·93 †
0·7	0·96	1·25	−3·38 †	− 7·37	− 5·72	−4·70 †
0·8	−2·06	−2·02	−0·63	−18·34	−18·53	0·74
0·9	−2·89	−2·66	−4·85 †	−18·34	−18·75	2·03 *
	50 cells × 50 cells (n = 200)			100 cells × 100 cells (n = 50)		
0·1	199·93	201·24	−1·67	797·48	802·98	−1·76
0·2	305·87	307·84	−2·40 *	1,215·70	1,233·68	−4·45 †
0·3	328·44	328·45	—	1,294·54	1,297·93	0·98
0·4	278·42	277·47	0·87	1,090·06	1,082·91	1·64
0·5	177·40	175·06	1·93	681·74	662·56	3·93 †
0·6	41·03	47·17	−5·17 †	122·82	142·21	−4·05 †
0·7	− 77·50	− 74·52	−3·00 †	−352·26	−348·78	0·87
0·8	−151·82	−152·55	1·03	−654·86	−657·71	1·01
0·9	−143·17	−143·70	0·92	−605·84	−607·50	0·71

* Significant at 5 per cent. level. † Significant at 1 per cent. level.

A similar comparison of observed and expected values of the standard deviation is shown in Table 36. Using large sample theory, the differences between observed and expected values are found to be significant in three out of thirty-six cases at the 5 per cent., and three out of thirty-six cases at the 1 per cent. level of significance.

TABLE 36

Comparison of observed and expected values of standard deviations of the number of (black minus embedded white) patches in a binomial field

p	10 cells × 10 cells (n = 1,000)			20 cells × 20 cells (n = 250)			50 cells × 50 cells (n = 200)			100 cells × 100 cells (n = 50)		
	S.d.		Normal deviate	S.d.		Normal deviate	S.d.		Normal deviate	S.d.		Normal deviate
	Observed	Expected		Observed	Expected		Observed	Expected		Observed	Expected	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
0·1	2·24	2·27	−0·60	4·31	4·47	−0·80	11·17	11·08	0·16	21·85	22·09	−0·11
0·2	3·42	3·40	0·40	5·20	4·70	2·38*	12·90	11·65	2·19*	25·71	23·18	1·09
0·3	2·62	2·55	1·17	5·75	5·13	1·70	13·20	12·89	3·61†	30·26	25·84	1·71
0·4	2·99	2·91	1·14	6·05	6·01	0·15	13·67	13·34	0·43	32·16	30·89	0·41
0·5	3·16	3·20	−0·57	6·35	6·68	−1·10	12·45	12·12	0·38	41·77	34·52	2·09*
0·6	3·19	3·17	0·29	6·38	6·58	−0·69	16·76	16·81	−0·06	29·71	33·85	−1·22
0·7	2·66	2·72	−1·00	5·55	5·55	0·00	13·57	14·03	−0·66	25·16	28·16	−1·06
0·8	1·82	1·99	−4·25†	3·95	4·01	−0·50	10·05	10·00	0·10	17·85	19·08	−1·06
0·9	1·70	1·50	6·67†	3·36	3·19	1·21	8·49	8·23	0·63	17·01	16·64	0·22

* Significant at 5 per cent. level. † Significant at 1 per cent. level.

It is clear that expected values of the patch number or of the variance are not confirmed in a rigorous manner. The general agreement between observed and expected values indicates, however, that the theoretical results are probably not wrong. The failure is, therefore, most probably due to some or all the fields used for experimental sampling not being of a sufficiently random character. This often happens in practice. In fact, as pointed out in the paper on large-scale sample surveys (in the *Phil. Trans.*), the concept of degrees of randomness is particularly useful in situations similar to the present one. I may conclude this section by stating that we have a big programme of work for model sampling experiments, but progress is slow for lack of resources.

Sample survey of the economic background of the Bengal famine

A sample survey was undertaken in 1944–45 to collect information relating to the after-effects and the economic background of the Bengal famine of 1943. A first report by Ramkrishna Mukherjee, Ambika Ghosh, and myself has been published in *Sankhya* (the Indian Journal of Statistics), Vol. 7, Part 4, and it is not necessary to enter into details, but a few typical results may be of interest in the present connection. The enquiry covered 15,769 families selected at random from 386 villages, which themselves were selected at random from 41 (out of 86 rural) sub-divisions (administrative units) covering about 60,000 sq. miles in Bengal. The design was zonal, with stratification of sub-divisions by intensity of incidence of famine conditions; and randomization of villages was completed separately within each sub-division.

From the sample survey it appeared that the land position was precarious even before the famine, with one-third of all rural families having no paddy land, while two-fifths had less than 2 acres, so that about three-fourths of all rural families had either no paddy land or owned less than 2 acres. With an average production of 820 lbs. of rice per acre, an average consumption of about 320 lbs. per head per year and an average family size of 5·4 persons, the subsistence level would be about 2 acres of paddy land per family on an average. The actual over-all average for the province was, however, found by the sample survey to be about 1·8 acres of paddy land per rural family, which was below the subsistence level. It is not surprising therefore that, averaged over a number of years before the war, there was a small but net import of a little over 1 per cent. of total production of rice and other cereals into the province.

The sample survey showed that the classification of sub-divisions by amount of paddy land owned per family before the famine was roughly parallel to the degree of incidence of famine conditions, indicating that sub-divisions in which there were more families with paddy land below subsistence level were more vulnerable to the famine.

It was found that about 1·6 millions of families (about one-fourth of the number who had owned paddy land before the famine) had either sold in full or in part, or had mortgaged their paddy land during the famine period. About a quarter of a million of families were obliged to sell all their paddy land, and were thus reduced to the rank of landless labour.

The net loss of plough cattle was about a million, or 13 per cent., during the famine period. Sales of cattle largely exceeded purchases, showing that transfers had taken place not merely from one rural family to another, but that large purchases to the extent of about 600,000 head of cattle had been made by outsiders (possibly by contractors for the supply of meat for army consumption).

The economic deterioration was measured by the number of families transferred from occupations at a higher economic level to occupations at a lower level. Assessed by such methods, it was found that about 700,000 of rural families had suffered a lowering of economic status during the famine. It was also found that even at the beginning of 1943 (before the advent of famine conditions) there had been already an increase of 150,000 of destitute persons. There was a further increase of about 330,000 destitutes during the famine itself, so that nearly half a million persons were rendered destitute under war and famine conditions in Bengal.

Even in the pre-famine period (January 1939 to January 1943) about 6·84 per cent. of rural families had suffered economic deterioration against 3·32 per cent. who had improved their economic level, while the position of 1·07 per cent. was not clear. Economic deterioration had thus set in definitely in the pre-famine period. Rates of change became more rapid during the famine period. Improvement in economic status during the famine period was relatively twice as great as that in the pre-famine period; but this was offset by a three times greater rate

of economic deterioration and twelve times greater rate of destitution. The famine period was thus one of greatly accelerated economic changes. Improvement of economic conditions was quicker, but was restricted to a comparatively small number of families. Deterioration and destitution were even more accelerated, and were shared by a large number of families. The poorer sections of the community, especially landless labour, fishermen, and village craftsmen, were most seriously affected, and many were rendered destitute; the middle group, who had land of their own and other assets, were naturally less vulnerable; and a comparatively small number of families in the upper stratum had remained immune, and sometimes even became prosperous.

The above summary gives an idea of the kind of information which it was possible to secure by the sample survey. The margin of error of the sample estimates in the cases investigated varied roughly between 3 and 5 per cent. Comparison with other available statistics also showed differences of the same order. For example, extrapolating from the cattle census figures of 1930 and 1940, the calculated number of plough-cattle in rural Bengal in the beginning of 1943 was about 8·3 millions. The sample estimate was about 7·9 millions, giving a difference of about 5 per cent. The sample estimate of 18·6 millions of acres of paddy land was also not inconsistent with other available estimates. On the whole, the above enquiry showed the possibility of using the sample survey in a quick and efficient manner to obtain information about economic conditions in rural areas with a margin of error sufficient for many practical purposes.

A multi-purpose survey in Bengal, 1946–47

I may briefly refer to an extensive survey (which is actually in progress in Bengal at the present time, in 1946–47) as an example of a multiple survey. The size of the grid (sample unit) in this case is just 1 sq. mile, and the information is being collected in the form of two independent but interpenetrating networks of sample units. In the first stage of the survey all households falling within each grid were surveyed with a rather short schedule covering a number of basic items, such as sex and age composition of the family, caste and community, occupation, agricultural land, number of cattle, total indebtedness, etc. Information has been already collected for 84,370 families from 475 grids scattered over about 60,000 sq. miles of rural Bengal. The information is being tabulated on Hollerith machines, and certain preliminary estimates have been made on the basis of about half the number of families surveyed.

The total rural population based on this portion of the material is found to be 50·16 millions, roughly as in March or April 1946, with a calculated standard error of 5·26 millions. The corresponding census population in rural Bengal in March 1941 was 47·185 millions. Owing to uncertainties in census counts it is difficult to calculate reliable rates of growth of population, especially in Bengal. It is also difficult to assess the exact effects of the Bengal famine of 1943 on the growth of population. We may, however, make some rough calculations. The total population (including rural and urban areas) of Bengal was about 46·7 millions in 1921, 50·1 millions in 1931, and 60·3 millions in 1941. Very rough estimates by linear extrapolation can be made on the above basis. Adopting the rate of growth during 1931–41, the extrapolated rural population as in March 1946 would be about 51·9 millions. There are, however, reasons to believe that the census count in 1931 was abnormally low on account of the non-co-operation movement. Adopting the average rate of growth during the twenty-year period 1921–41, the extrapolated figure for the rural population at the end of 1944 would be about 50·4 millions. The observed sample estimate of 50·1 millions is thus of the right dimensional order; and is fairly satisfactory with a sampling fraction of about 1 : 250.

The enquiry just completed is, however, only the first fold or layer of the survey. Arrangements are now being made to select about 20 per cent. of the families included in the first survey for a more detailed study of rural indebtedness and agricultural labour. If everything proceeds smoothly it should be possible to amplify the results for the second survey for making estimates for the sample covered in the first survey, and hence for the rural population of Bengal as a whole. After the second survey is over, if funds permit, it is proposed to reduce the number of families much further and carry out a detailed enquiry into family budget and other socio-economic conditions. At each successive survey the families would be selected at random, but with appropriate zoning and/or stratification, so that it should be possible to amplify the results of one survey to obtain estimates for the families covered in an earlier survey, and hence, finally, for all rural families in the province.

Other work in statistical sampling

A good deal of other work on statistical sampling has been and is being done in the Indian Statistical Institute. In most cases these have definite practical ends in view. The important point which I should, however, like to emphasize is that such practical studies have continually given rise to theoretical problems of fundamental interest. The subject of crop surveys itself raised many questions of theoretical interest, to some of which I have already referred. I have also mentioned that probability problems relating to space distributions are being tackled on fruitful lines with the help of topological and combinatorial concepts by Raj Chandra Bose. The close integration of applied work and theoretical researches has been such a valuable feature of the work of the Institute that it seems worth while giving a few more examples.

Analysis of anthropometric measurements. The statistical analysis of anthropometric measurements—work which may be properly considered to be purely of an applied nature—had led, about 20 years ago, to interesting theoretical developments in the formulation of the generalized distance (D^2-statistic). It was the starting point of a good deal of mathematical research in the theory of sampling from multivariate correlated population by Raj Chandra Bose and Samarendra Nath Roy, whose work in this subject is already well known.

We had occasion recently to take up an anthropometric study of an extensive series of physical measurements covering 3,000 individuals belonging to twenty-two different castes and tribes in the United Provinces of India. All the measurements were taken by one individual observer, Dr. D. N. Majumdar of the Lucknow University, and were therefore free from differences of personal equation. This made the material particularly suitable for comparative purposes. The use of the D^2-statistic has, I believe, yielded information of great interest and significance. As the paper is in the press, and will be shortly available, I need not enter into details.

I may mention one result which is of some methodological interest. From general considerations I reached the conclusion that, when correlations between variates are taken properly into consideration (as in the D^2-statistic), the use of indices (such as the cephalic or the nasal index) do not supply any additional information. This has been fully confirmed by C. R. Rao by actual numerical calculations. In fact, the present study has given rise to a number of interesting theoretical developments on which Rao has done some very useful work.

More recently, in 1945–46, Dr. D. N. Majumdar, working under the auspices of the Institute, has collected a large volume of material relating to physical measurements and blood-group tests of about 4,000 individuals belonging to various communities, castes, and tribes of Bengal. The analysis of this material should yield valuable results. A survey of blood pressure which was concluded some time ago is perhaps worth mentioning in the present connection.

Design of experiments. At one time a great deal of work relating to the design of agricultural experiments had been done in the Institute. This had led to significant theoretical researches in which the concepts of the Galois field and finite geometry were used with great success by R. C. Bose and, under his leadership, by workers like K. Raghavan Nair, K. Kishen, C. R. Rao, Harikinkar Nandi and others.

Circulation of rupees and rupee notes. At the request of the Reserve Bank of India, experimental surveys were made and methods were devised for estimating the circulation of silver rupees of various dates by sample counts at a number of receiving centres, like banks or railway stations. A method was also devised, based on sample counts of rupee notes received back in the Reserve Bank, to estimate the average life of such notes. It may be mentioned, incidentally, that this project gave rise to an interesting problem of measuring the magnitude of the difference between samples drawn from multinomial populations which was tackled on fundamental lines on the theoretical side by Anil Kumar Bhattacharyya, with promising results. A brief note was issued in the Report of the Reserve Bank of India on Currency and Finance for the year 1940–41, pp. 49–54.

Sampling for yield of cinchona bark. I have already mentioned the project for estimating the yield of cinchona bark as an example of multiple sampling. It is interesting to note that the final choice of physical measurements was made in this case from considerations of cost. It was found, for example, that although the measurements of the surface area of the standing plant gave the highest correlation with the yield of bark (0·865 in one case against the next highest 0·690 for a different character), the use of three simple physical measurements like the " standing

vertical height of the plant," the " girth at height of 6 inches above ground level," and the " number of stems of the plant at ground level," gave a multiple correlation of 0·848. Theoretically the use of surface area would be more efficient, inasmuch as observations on eighty-nine plants would give the same information as the measurement of the three other characters on 100 plants. The time and trouble to measure the surface area for eighty-nine plants would be, however, far greater than that for measuring the other three characters on 100 plants. This is a simple example of the use of cost considerations in settling the sampling programme.

This particular project also has given rise to many interesting problems of the estimation of appropriate errors in the case of multiple sampling of various kinds. Mrs. Chameli Bose has obtained a number of useful results and is working on the subject.

Population enquiries. After the Indian Census of 1941, the Government of India had decided, as a matter of economy, to cut down most of the standard tables; even age or occupational tables were not prepared. Fortunately, Mr. M. W. M. Yeatts, as Census Commissioner, had issued instructions for the preservation of a 2 per cent. sample consisting of every fiftieth individual census slip. In certain areas, mostly in Indian States, full tabulation had been carried out. A comparison of results based on the complete count and on the 2 per cent. Y-sample in such areas showed that it should be possible to reconstruct most of the tables with sufficient accuracy for practical purposes. About 7 millions of the 2 per cent. slips have been brought over to the Statistical Laboratory in Calcutta from all over India, and the work of transferring the information to Hollerith cards has already started. In only one province—namely, Bihar—the original census slips had also been preserved; and *ad hoc* sampling studies on the basis of this material are also proceeding.

The above projects were sanctioned on the recommendations of the Population Data Committee which was appointed by the Government of India in May 1944, and which submitted its report in June 1945. A comprehensive programme relating to population census and demographic statistics generally was prepared by this Committee, and it was definitely recommended that a continuing sample census of the Indian population should be carried out from year to year.

Statistics of road development. Mention may also be made of the use of sampling methods for traffic count on roads and connected socio-economic variates in the traffic catchment. The object in this project is to lay down a scientific foundation for preparing programmes of road development. K. B. Madhava and Satyabrata Sen are at present actively engaged in this work.

Postal traffic and revenue. An enquiry is proceeding at present to explore the possibilities of using the sampling method for estimating in advance the volume of traffic and revenue in different postal sectors. Exploratory surveys under the guidance of Satyabrata Sen have already yielded encouraging results.

Integrated programme of work

The need of handling computational work on a very large scale has been salutary in teaching us the importance of efficient organization of the human agency. J. M. Sen Gupta and N. T. Mathew and others have done valuable work in this direction. An account of this work, especially from the point of view of cost accounting, would, I think, be of considerable interest. Unfortunately, we have not yet had sufficient time or money to undertake this in a systematic manner.

Along with computational work may be mentioned the preparation of statistical tables of various kinds (which often require the calculation of numerical values of Bessel and other mathematical functions), which has been proceeding for many years, and offers scope for training in a very useful branch of statistical work. The names of J. M. Sen Gupta, Purnendu Kumar Bose, and Raja Rao may be mentioned.

The contact between statisticians and field workers has also been fruitful in every way. The credit for the successful organization of the field side is almost entirely due to the organizing leadership and the ability to appreciate statistical needs on the part of Nihar Chandra Chakravarti and workers in the Field Branch, like Dhirendra Mohan Ganguly, Pronay Kumar Chatterjee and others. Much valuable work on the field has also been done by statisticians like J. M. Sen Gupta, Birendranath Ghosh, N. T. Mathew, Purnendu Kumar Bose, Harikinkar Nandi, and others. I may perhaps also mention that during the last few years the usual practice was

for all theoretical workers, including R. C. Bose and S. N. Roy, to live in camp on the field for a few weeks in connection with the crop-survey work.

Training in statistics has also been an important part of Institute activities from the very beginning. One-year post-M.Sc. courses, as also occasional special courses in particular subjects, are being arranged for a long time. The Institute workers have also been intimately associated with the two-year course leading to the M.A. or M.Sc. degree and the recently introduced two-year course leading to the Honours B.Sc. degree in Statistics in the Calcutta University.

In the light of the experience gained during the last fifteen years I am convinced that the close integration of applied work, theoretical research, and training offers the soundest line of advance in statistical work, especially on the professional side. It offers scope for developing the spirit of team work among persons with widely varying interests which alone can make it possible to tackle many problems on a scale large enough to ensure success. I have been fortunate in having secured a large group of able and enthusiastic workers, each of whom in his own line has much greater knowledge and ability than myself. Whatever success the Indian Statistical Institute has achieved is in fact essentially due to the spirit of co-operative effort.

References

[1] Bhattacharyya, A.: On a measure of divergence between two multinomial populations. *Sankhyā*, Vol. 7 (4), 1946, pp. 401–6.

[2] Deming, W. E.: Errors in Surveys. *American Sociological Review*, Vol. 9 (4), 1944, pp. 359–69.

[3] Hubback, J. A.: Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166, 1927, and reprinted in *Sankhyā*, Vol. 7 (3), 1946, pp. 281–94.

[4] King, A. J., and Jessen, F. R.: The master sample of agriculture. *Journal of the American Statistical Association*, Vol. 40, No. 229, 1945, pp. 38–56.

[5] Mahalanobis, P. C.: "A statistical report on the rupee census" published in the report on currency and finance, 1940–41, by Reserve Bank of India, June 1941, pp. 49–55.

[6] Mahalanobis, P. C.: Sample surveys. Presidential address, Section of Mathematics and Statistics. *Proceedings*, Indian Science Congress, 1942.

[7] Mahalanobis, P. C.: On large scale sample surveys. *Phil. Trans.*, Vol. 231 (B), No. 584, pp. 329–451.

[8] Mahalanobis, P. C., Mukkerjee, R. K., and Ghosh, A.: A sample survey of after effects of Bengal famine of 1943. *Sankhyā*, Vol. 7 (4), 1946, pp. 337–400.

[9] Sukhatme, P. V.: Bias in the use of small size plots in sample surveys for yield. *Nature*, Vol. 157, No. 3993, p. 630.

[10] Wald, A.: Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, Vol. 16 (2), 1945, pp. 117–86.

[11] Yates, F.: Some examples of biassed sampling. *Annals of Eugenics*, Vol. 6 (2), 1934, pp. 202–13.

[12] Yates, F.: A review of recent statistical developments in sampling and sampling surveys. (Read before the Royal Statistical Society, January 23rd, 1945.)

Discussion on Professor Mahalanobis's Paper

Dr. F. Yates: I have much pleasure in proposing the vote of thanks. Professor Mahalanobis has given us a most interesting description of the work for which he has been responsible in India. His address has provided a vivid picture of the Institute which he directs, and at the same time has shown what enormous strides have been made there, and how extremely able his direction has been.

There is not time to discuss all the many interesting points that Professor Mahalanobis has raised, but I should first like to take the opportunity of expressing my admiration for the way in which he has handled the analysis of the social survey and broadcast survey data, the discussion of which formed the last past of his address. Those who have had experience of the analysis of data of this type will know that while it is comparatively easy to present tables giving overall percentages, etc., the critical analysis of such material does present very great difficulties. Professor Mahalanobis is certainly to be congratulated on the penetrating methods of analysis he has developed in connection with these surveys. There is no doubt that the statistical analysis is the most difficult part of social survey work of this kind, and it is much to be hoped that further efforts will be made to develop the methodology.

Professor Mahalanobis did not spend much time discussing his sampling methods. There are several points of mutual interest that I should like to have the opportunity of discussing with him personally, but which I will not trouble to bring up now. I fully agree with Professor Mahalanobis that the determination of the sampling procedure is often the easiest part of a survey, but I would stress that this ease is in fact based on experience, which has to be acquired by a thorough study of the variability of the material that is being sampled.

Professor Mahalanobis finds the sampling problems easy because his Institute has built up a body of experience in sampling of the particular types of material with which it is concerned.

This contains a moral for us all. We cannot devise good statistical methods merely by sitting in our studies and theorizing. Good statistical methods almost invariably result from contact between the mathematical statistician and the workers who are responsible for collecting the data, and who are interested not in the statistical methods themselves, but in the conclusions that emerge from their work. Consequently, the statistician must be in intimate contact with the actual numerical material. He must not think solely in terms of algebraic symbols. Professor Fisher once made what I consider a most revealing remark: "Most of the statistics which I have learnt, I have learnt on the computing machine."

Professor Mahalanobis has been concerned with the problem of securing numerical accuracy in the computations. In this connection I might mention the experience I had in the geodetic survey of West Africa. In that part of the world it was necessary to carry out much of the main framework of the survey by means of traverses instead of triangulation. The difficulty with traverses is that they involve a vast amount of computation which is not self-checking. In triangulation, once the base line is determined, the remainder of the computations are self-checking—it is impossible to make an error of any magnitude in a computation which will not reveal itself later in glaring discrepancies.

We had a very great struggle to get the numerical work of our traverses correct. We eventually evolved a most rigorous system of computation. All computations were done in the field in duplicate, the duplicates being compared only at certain points. A third computation was then done at headquarters using the "setting-out" taping, which was carried out in metres instead of feet, and bearings 45° different from the true bearings, so as to give different sines and cosines. This last computation was carried out so as to eliminate errors of the type that are likely to be made independently by different computers—*e.g.*, writing an odd taping length of 449·321 ft. as 499·321 ft., when practically all the taping lengths are just under 500 ft.

I was interested, after this system had been evolved, in finding that the U.S. Coast and Geodetic Survey laid it down in their standard instructions that triangulation was to be used in preference to traverses, if the cost was not more than 100 per cent. greater, simply because of the danger of computing errors. (It is now recognized that with modern invar tapes traverses are quite as accurate, possibly more accurate, than triangulation.)

Professor Mahalanobis stressed the importance of providing the results of sampling surveys on the required dates. We had an interesting example of this—which illustrates the flexibility of sampling methods—during the war, when we were making an estimate of the amount of standing timber in the country. The first estimates were required within about five months of the decision to undertake the survey. Initially the survey had been planned to be undertaken in two 5 per cent. samples, the sampling units being 6-inch Ordnance Survey Quarter Sheets (3 miles × 2 miles). Owing to the almost inevitable delays in organization, collection of staff, etc., it was clear by the time we were ready to start work, that the first 5 per cent. sample would not be completed and analysed by the given date. I therefore insisted that the first 5 per cent. sample should be divided into two parts. This was done by the simple process of drawing a diagonal line across each sampled 6-inch Ordnance Survey Quarter Sheet. There was, in fact, considerable opposition to this, as it was thought that the results of a $2\frac{1}{2}$ per cent. sample would not be sufficiently accurate for the required purpose. Fortunately I knew, from previous experience, the variability of the material, and was able to promise that the results would be of the required accuracy, and that moreover when they were obtained, estimates of the actual standard errors could be furnished so as to provide concrete evidence of this accuracy. I am not sure whether these standard errors really carried conviction, but what certainly did convince those who were initially in doubt was the close agreement between the first $2\frac{1}{2}$ per cent. sample and the second $2\frac{1}{2}$ per cent. sample, and between these and the second 5 per cent. sample.

This illustrates an important point in survey work. In general, overall estimates for the whole area are first required. It is on these estimates that the broad administrative decisions are based. Subsequently more detailed estimates for different parts of the area may well be required, in order to implement the administrative decisions. By carrying out a sampling survey in stages, as was adopted in the Forestry Survey, and as is also, I think, Professor Mahalanobis's practice, the administrators can be given the necessary information on which to base their decisions, while additional data for more detailed and accurate estimates are being collected.

Professor Mahalanobis mentioned the large biases that he has encountered when taking small areas for estimating crop yields. To a certain extent, we have had the same trouble in this country, but only to a limited degree. I suspect that many of these troubles will gradually fade away, as a permanent organization of trained people is built up. In our survey work in Africa, for example, much of the field work was undertaken by native surveyors who were trained in a Survey School which had been specially set up for the purpose. We had very little trouble with these surveyors. The very rigorous methods of booking observations which are always followed in survey practice were a great help. All observations must be booked direct into properly printed field note-books which have numbered pages. The surveyor is completely at liberty to reject any observations with which he is not satisfied, but a permanent record of such observations is always available, so that it is possible to find out what the surveyor has really done. Sometimes they did very funny

things. One surveyor—quite a conscientious man—had been taken to task because his field work was not reaching the required standard of accuracy. Subsequent to this he turned in a series of traverses in which the closing errors in his bearings (on astronomical azimuths) were far smaller than they should have been. At first it was thought that the results had been deliberately faked. Subsequent cross-questioning revealed, however, that having made a preliminary computation of his closing error and finding it larger than he thought it should be, he had gone back the next day and re-observed some of the angles, rejecting the old or the new observation of each angle according to which improved the closure. The methods of booking enabled him to prove that this was what had been done, and also enabled the matter to be set right without any re-observation.

DR. C. OSWALD GEORGE, in seconding the vote of thanks: Never have I so much enjoyed listening to an address of this character, not only because of its intrinsic interest, but also because of the particularly charming manner of its delivery.

The first point which strikes one is perhaps the difference between conditions in India and in this country. On the other hand, there are important similarities. Of the various headings Professor Mahalanobis has written on the blackboard, we in this country also have to consider problems of time and sampling errors, and—for there is a body here known as the Treasury—the cost of statistical enquiries. Calculating and recording errors are also not unknown here. Had I been speaking a fortnight ago, I might have spoken more confidently on the much higher standards of accuracy in this country, but I fear Professor Mahalanobis may have seen the unfortunate calculations of mean yields in a recent issue of what we had previously regarded as an almost infallible newspaper. So I will pass over that point.

What he calls errors from physical fluctuation, I assume, arise with crops such as jute, where one reporter may find two acres under the crop, while another reporter, visiting the same farm a fortnight later, may find no jute acreage whatever. Under the existing system in this country, where returns relate to a specified day, the same problem does not arise, but some of his sampling problems interest us. He tells of the various sampling methods he adopted, "unitary and zonal, configurational and unrestricted," but unfortunately he has not given us sufficient detail to permit any useful comment on their theoretical or practical merits. He frequently mentioned large-scale sampling theory, but if greater use had been made of stratified sampling, small sample theory would possibly have come more into the picture.

I was particularly interested by his detailed treatment of errors, for I think the question of errors (other than sampling errors) is of more importance than is sometimes believed. How many can lay their hands on their hearts and say that in their published work, errors—if not quite of the type most troublesome in India, nevertheless of serious import—have never reached the printers? But it is noteworthy that Professor Mahalanobis does not mention errors due to the machines: perhaps in the Indian climate machines are infallible!

Another point that struck me was the speaker's attitude towards a system of complete enumeration and his declaration, when comparing it, detrimentally, with sample surveys, that complete enumeration is entirely lacking in any kind of control and precludes any valid estimate of the margin of error of the final results. Without further explanation it is not easy to see why the underlying weaknesses are inherent in the one and entirely absent from the other, or why a system of duplicated surveys cannot be used in conjunction with complete enumeration. And whatever may be desirable in India, it is not likely that in this country complete enumeration will be entirely dispensed with, but rather that it will be used as a basis upon which sampling surveys may be developed.

Professor Mahalanobis has laid particular stress on the results of the Bihar survey shown in Table I, and it is to this that I should like to confine my final remarks. He says that his system of sampling was very successful, in that in the case of rice and pulses a very small sample of about 0·025 per cent. gave results within about 2 per cent. of the final result obtained from the total sample. We are not told what was the result of the small sample in the case of sugar-cane. But what strikes one most about Table I is perhaps the variations between the various estimates, particularly that between estimates 8 and 9, which suggest something more than sampling errors. Estimate 9 for rice gave 13,650,000 acres, a fall of 355,000 acres compared with estimate 8. It is noticeable that this is a larger difference than that between estimate 1 (based on a sample of one in 4,175) and estimate 9, although it resulted from a large increase in the sample (from 80·7 to 100·0 per cent., or 19·3 per cent. of the whole sample). If my arithmetic is not wrong, it seems that the total acreage in or about January, if it had been estimated from this 19·3 per cent. of the total sample, would be only about 12·2 million acres.

What is the explanation of such large discrepancies? We have not been given sufficient detail to decide this point. An apparent explanation would be that there are what Professor Mahalanobis referred to as physical fluctuations, which would presumably mean that the acreage under rice changes substantially from month to month, or even from day to day. But this can hardly be the explanation, for if it were, any similarity between the first estimate in October and the last in January could hardly be attributed to the merits of the sampling methods employed. The question would also arise of what was being estimated: was it the acreage under the crop in October, or