SOME CONCEPTS OF SAMPLE SURVEYS IN DEMOGRAPHIC INVESTIGATIONS

P. C. MAHALANOBIS

1. **Non-sampling errors in sample surveys:** Sample surveys can supply, with speed and economy of cost, demographic information with sufficient accuracy for most practical purposes and with the possibility, in a properly designed and properly conducted survey, of making a valid estimate of (what is usually called) the margin of errors of sampling. The total margin of error, however, includes not only the theoretical errors of sampling but also "non sampling errors" which arise from the personal bias of different investigators, or from differences which arise at the stage of collection of processing of the data. Such non-sampling errors are often large and may be even larger than the theoretical errors of sampling. The object of this paper is to draw attention to some recent techniques for the study of "non-sampling errors" in sample surveys.

2. **The design of "inter-penetrating network of sub-samples" (IPNS):** This method can be used to study the differences arising from two or more sources of errors, such as two or more investigators or parties of investigators, forms of questionnaires etc. The logic of the IPNS design is simple. The sample (*i.e.* the total number of sample-units to be investigated) is selected in the form of two or more independent sub-samples with replacement, each sub-sample having full coverage of the whole population under survey. The information is collected or processed for each such sub-sample in such a way that one such sub-sample is assigned to each source of non-sampling error. For example, in order to study the non-sampling bias due to two investigators or two parties of investigators, information for one sub-sample would be collected by one party of investigators and the information for the other sub-sample by the other party of investigators; differences, if any, in the results based on the two sub-samples or components of the IPNS may then be ascribed to (non-sampling) differences between the two parties of investigators. An appropriate analysis of variance would in such cases supply relevant information for the study of non-sampling errors.[1]

3. **Four demographic characters in India:** *May—November* 1955.—I shall give numerical results for four demographic characters, namely, (1) sex-ratio (number of males per hundred females), (2) proportion of population in labour force, (3) proportion of population of age 50 and above, and (4) proportion of population literates, based on the National Sample Survey, round 9: May to November, 1955.

4. **Sample design and procedure:** The design was two-stage stratified with IPNS. Rural India was divided into a number of 'strata', each stratum consisting of 'districts' or 'district groups'; within each stratum, sample villages were selected with probability proportional to population and with replacement; in each selected sample village, 10 households were selected with random start. Urban areas were divided into 'blocks' as first-stage units; sample blocks were selected systematically with four independent random starts; and in each selected sample block, 16 households were selected systematically with random start.

The survey was conducted by 'two parties of investigators' in each State, each party working in two periods' of three months each, giving four different estimates for each demographic character for each State.

5.1. **Analysis of Variance, by States:** For each of 19 different States, an analysis of variance was done with the following partition into three degrees of freedom (d.f.): one d.f. between 'time', one d.f. between 'party' and one d.f. as error ('party' x 'time') with a total of 3 d.f. In this way, 19 F-ratios for each of 'party/error' and of 'time/error' were obtained for each demographic characteristic for the 19 States of India.

1. The IPNS design, which is being extensively used in India since 1935 or 1936, has been discussed in my two papers "On Large Scale-Sample Surveys" in *Phil, Trans. Roy. Soc., London* 1944, and "Recent Experiments in Statistical Sampling in the Indian Statistical Institute" in *Jour. Roy. Stat. Soc.*, Vol. 109, Part 4 (1946), and also in the reports of the National Sample Survey of India published in *Sankhya, the Indian journal of Statistics*. A brief description will be found in the United Nations publication on "Recommendations for the Preparation of Sample Survey Reports" (Provisional issue) June 1964.

5.2. The following Table 1 gives the analysis of variance of proportion of literate persons for each of the 19 States of India.

Table 1

Analysis of variance of proportion of literate persons, by States

State	Mean square			F-ratio*	
	Party[1]	Time[1]	Error[1]	Party	Time
1	2	3	4	5	6
1. Uttar Pradesh	1·73	10·99	2·39	1·38(r)	4·60
2. Bihar	8·67	1·65	0·38	22·82	4·34
3. Orissa	29·59	6·55	2·50	11·84	2·62
4. West Bengal	1·48	1·65	0·20	7·40	8·25
5. Assam	89·02	0·04	0·94	94·70	23·50(r)
6. Andhra	8·61	0·32	4·78	1·80	14·94(r)
7. Madras	0·08	12·04	7·24	90·50(r)	1·66
8. Mysore	12·08	0·11	6·62	1·82	60·18(r)
9. Travancore & Cochin	22·04	34·28	30·75	1·40(r)	1·11
10. Bombay	17·64	2·34	7·02	2·51	3·00(r)
11. Saurashtra	46·31	83·63	3·48	13·31	24·03
12. Madhya Pradesh	0·90	2·25	7·03	7·81(r)	3·12(r)
13. Madhya Bharat	1·09	0·01	13·59	12·47(r)	1359·00(r)
14. Hyderabad	0·22	2·53	0·28	1·27(r)	9·04
15. Vindhya Pradesh	7·02	4·75	0·14	50·14	33·93
16. Rajasthan, Ajmer	0·11	0·07	1·37	12·45(r)	19·57(r)
17. Punjab, Delhi	22·56	0·42	2·54	8·88	6·05(r)
18. PEPSU	3·76	63·68	25·91	6·89(r)	2·46
19. Jammu & Kashmir	6·13	0·32	7·04	1·15(r)	22·00(r)

5.3. It will be noticed that none of the 19 F-ratios is statistically significant showing that, firstly, there was no significant difference between results based on the information collected by the two 'parties' or between those based on information collected in two different time periods. The survey in respect of proportion of literates thus did not show any significant non-sampling errors due to either 'party' or to 'time' factors. Also out of 19 F-ratios, 9 values, those marked (r) within brackets, were cases of F 1 (less than or equal to one) and 10 were cases of F 1 (Greater than one).

*Significant at 5% level : F 5% (1, 1)=161·4 ; at 1% level : F 1% (1, 1)=4052.

1. Each with 1 d. f. The symbol (r) indicates error/party or error/time in F-ratio.

5.4. There are three other similar tables of F-ratios for each of the three other characteristics, not shown here. The combined frequency distribution of observed values of the F-ratio for all four demographic characteristics are shown in cols. (2) and (5) in the following Table 2.

Table 2

Frequency distribution of observed F-ratios

Demographic characteristics	F=party mean square / error mean square			F=time mean square / error mean square		
	F1	F1	Chi-squares	F1	F1	Chi-square
1	2	3	4	5	6	7
1. Sex-ratio (19) proportion of population	8	11(9·5)	0·4737	9	10	0·0526
2. In labour force (19)	9	10(9·5)	0·0526	6	13	2·5789
3. Age 50 and above (19)	14	5(9·5)	4·2632	11	8	0·4737
4. Literate (19)	9	10(9·5)	0·0526	9	10	0·0526
5. 4 Characteristics (76)	40	36(38)	0·2105	35	41	0.4737
6. 4 Characteristics each for 'party' and 'time'(152)				75	77(76)	0·0263

5.5 Half the F-ratios are expected to be less than or equal to one (that is, cases of F 1) and half greater than one (that is, cases of F 1). The expected numbers of F-ratios are shown within brackets in col. (3); the expected numbers would be the same for col. (6). The corresponding values of chi-square, to test the agreement between observed and expected numbers of F-ratios are given in columns (4) and (7) for 'party' and 'time' respectively.

5.6 It is clear that 'party' differences were not significant, showing that the survey was conducted under satisfactory statistical control. 'Time' differences were also non-significant showing that these demographic characteristics were not affected by the difference in the time of collection of information.

6.1 **Combined analysis of variance:** It is possible to make a combined analysis of variance between 'state', 'party' and 'time' with either three interactions (party x time, state x party, and state x time) or with a single interaction (party x time). With three interactions, the error would be based on 18 degrees of freedom, and with one interaction on 54 degrees of freedom.

6.2. Results for such an analysis of variance are given in Table 3 for proportion of literates.

Table 3

Analysis of variance of proportion of literates between State, party and time

source	degrees of freedom	sum of squares	mean square	F-ratio
(1)	(2)	(3)	(4)	(5)
with three interactions				
1·1 State	18	7984·00	443·56	64·86**
2·1 Party	1	0·35	0·35	19·60(r)
3·1 Time	1	13·06	13·06	1·90
4·1 Party x time	1	0·77	0·77	8·91(r)
5·1 State x party	18	278·69	15·48	2·26*
6·1 State x time	18	214·58	11·92	1·74
7·2 Error(1)	18	123·43	6·86	—
8 Total	75	8614·88	—	—

(r) indicates error/party, or error/party x time. significant at 1% level**, and at 5% level*

Table 3—*contd.*

(1)		(2)	(3)	(4)	(5)
	with one interaction				
1·2	State	18	7984·00	443·56	38·84
2·2	Party	1	0·35	0·35	32·63(r)
3·2	Time	1	13·06	13·06	1·14
4·2	Party x time	1	0·77	0·77	14·84(r)
7·2	Error (2) (5·1 +6·1+7·1)	54	616·70	11·42	
8	Total	75	8614·88		

6.3 Differences between states are definitely significant whether the analysis is made with three interactions or with one single interaction. Different states clearly have different proportions of literates. On the other hand differences between results collected by different parties or at different periods of time are not significant, neither are the interactions.

Fractile graphical analysis

7.1. With the IPNS-design it is also possible to use a very simple graphical analysis for detailed investigations of demographic characteristics, for example, the change in the proportion of adults with increasing level of living. In a household survey of expenditure it is usual to record the age and sex composition and the size of the household, and also the total consumer expenditure during a suitable reference period, for example, in 30 days in the case of the National Sample Survey of India, round 8: July 1954-March 1955. For each household, the *per capita* expenditure can be then found by dividing the total expenditure by the number of persons in the household; and this *per capita* expenditure may be used as an approximate indicator of the level of living.

7.2. Data for percentage of adults in households are given in the following Table (4) separately for rural and urban areas of India based on the National Sample Survey, round 8.

Table 4

National Sample Survey of India: *Round* 8: *July* 1954-*March* 1955.
Ranking of households by per capita *total consumption expenditure*:

Percentage of adults

Decile per cent	Sub-sample 1	Combined	Sub-sample 2	Sub-sample 1	Combined	Sub-sample 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Rural			Urban	
0—10	51·1	49·4	48·6	58·4	55·0	53·3
10—20	51·4	54·3	55·6	52·8	56·8	59·0
20—30	58·6	56·0	53·8	60·9	62·5	56·3
30—40	55·6	57·1	58·9	66·3	57·4	54·3
40—50	59·0	58·0	56·5	63·5	64·8	59·0
50—60	60·9	61·3	61·5	69·8	59·3	63·3
60—70	59·0	61·0	63·1	58·4	66·8	64·0
70—80	62·4	60·9	59·0	68·6	64·5	65·5
80—90	61·3	61·1	63·1	65·7	68·7	72·4
90—100	67·3	67·9	66·8	74·6	75·2	75·0
0—100	58·6	58·7	58·7	64·0	63·0	62·2
No. of sample households	931	1869	938	224	447	223

7.3. In the IPNS-design such information will be available for at least two sub-samples and also for the combined sample. The sample-households in each sub-sample and also in the combined sample are ranked in, say, ascending order of *per capita* expenditure. The number of sample-households is multiplied by the appropriate probability factor to give the estimated population, and the total (estimated) number of household is divided into a number of equal divisions, for example, ten decile groups. The proportion of adults in each decile is then calculated and plotted on the y-axis corresponding to the successive decile groups shown at equal distance apart on the x-axis; successive points are finally joined by straight lines to give graph (1) for the first sub-sample. Graph (2) may be then plotted in the same way for the second sub-sample; and also the combined graph(1, 2) for the combined sample. The area included between graph (1) and graph (2) supplies a geometrical or graphical error-area associated with the combined graph (1, 2).

7.4. Two sets of graphs, each consisting of graph 1 for sub-sample 1, graph 2 for sub-sample 2, and graph (1,2) for the combined sample, for urban and rural areas are shown in the accompanying chart.

7.5 It will be seen that the percentage of adults is higher in urban areas in comparison with rural areas and increases with increasing level of living in both rural and urban areas.

7.6 The error-area is given by the area enclosed within graph 1 and graph 2 in each case. The error-area is larger for the urban area partly at least on account of the smaller size of the sample. Also, the error-areas overlap to a large extent up to about 70 per cent of the population (ranked according to increasing level of living) showing that differences in the percentage of adults in rural and urban areas are not statistically significant in this range on the basis of the available evidence.

7.7 There is however no overlap in the error-areas for the percentage of adults in the top twenty per cent of population (ranked in accordance with the level of living) indicating that for the richer households the percentage of adults is significantly higher in urban areas. A proportionately larger number of adults in the household very likely implies a proportionately larger number of earners and hence a larger *per capita* income.