SYMPOSIUM ON 'RECENT DEVELOPMENTS IN SAMPLE SURVEY TECHNIQUES'

A symposium on "Recent Developments in Sample Survey Techniques" was held during the 27th Annual Conference of the Indian Society of Agricultural Statistics jointly organised with the 1st Annual Conference of Agricultural Research Statisticians at the I.A.R.S., New Delhi, on 26th April, 1974 under the Chairmanship of Dr. P.V. Sukhatme, Emeritus Scientist, Gokhale Institute of Politics and Economics, Poona. The Convener was Dr. M.S. Avadhani, Associate Professor of Statistics, I.A.R.S., New Delhi. Four articles were contributed and presented at the symposium. Their extended summaries are given in the following paragraphs :

M. S. Avadhahi :*

Theory of Sampling from finite Populations

2. This paper gives a comprehensive and critical review of the work done by various research workers on sampling from finite populations during the past decade. The main problem of sampling consists in devising a suitable selection procedure for securing a representative sample from a given finite population and then building up the estimate of the parameter of interest on the basis of sample observations. A sampling procedure which assigns probability proportional to size of the units may be expected to provide a representative sample. A broad over-view of the work done on unequal probability sampling reveals that the work has mostly been concentrated on evolving sampling procedures such that the usual ratio estimate or Horvitz-Thompson estimate provides an unbiased estimate of the population total Y with an unbiased estimate of its variance also. Of all the procedures of varying probability sampling reviewed those of Lehri-Midzuno-Sen (LMS), randomised systematic sampling and of Rao, Hartley and Cochran facilitate selection of sample of any given size in practice. However, the latter two procedures have limitations of one kind or the other. The remaining are all IPPS sampling procedures which lack simplicity even for selection of the sample.

* I.A.R.S., New Delhi.

3. The basic principle governing the formation of strata is to ensure that the units within the strata are as homogenous as possible. Since in practice the variate values for the population units are not known, the use of a-priori information on auxiliary characters, both qualitative as well as quantitative, introduces some uncertainty in the homogenisation of units within strata which distorts the accuracy of the estimate. For populations having a bi-variate frequency distribution with g model two procedures of stratifying the population have been suggested and the author has shown that the variance of the estimate depends very much upon the departure from the regression within strata. It has also been observed that the suggested strategy is always superior to Lehri-Midzuno-Sen ratio strategy and is more efficient than Hartley and Rao with g less than 1 or g greater than 2. Another important feature of the suggested strategy is that the variance of the suggested estimates can be obtained before hand which is of immense value in practice.

4. The technique of controlled selection with equal and unequal probability of selection has been investigated in detail and the problem of exercising post selection control over non-preferred units has also been studied by the author with the help of a generalised version of the technique proposed by Keyfitz (1951).

5. Many times our interest lies in securing data for estimation of several parameters simultaneously. The stratification for multiple characters presents complex problems. It would be desirable to have a common sample for all the characters within any stratum. The principal difference between the approaches used for stratification lies in the types of restraints based on the individual variances. It is yet to be seen as to how far these criteria lead to a reasonably satisfactory allocation and also examine their relative merits from efficiency and administrative convenience point of view.

6. For comparing systematic sampling with simple random sampling Cochran (1946) assumed that the finite population is a random sample from a super population and pointed out that the results proved under this model do not apply to any single finite population but to the average of populations that can be drawn from the super population. However, he justified its use by saying that any sampling method is used in practice on a series of finite populations. The assumption underlying the super population model is exactly similar to the Bayesian assumption regarding prior distribution the only distinction being that the Bayesian requires the exact form of the prior distribution whereas in the super population approach it need not be known exactly. Both the approaches suffer

from lack of verifiability. To overcome these difficulties and yet make meaningful comparison the author presented two models of which one is appropriate to discrete variates and the other to the continuous ones. One important feature of these models is that they are realistic and the results proved under them hold good for the population at hand. Further, in the comparison of PPS or IPPS sampling studies with LMS ratio strategy it is interesting to know that the conditions for the former to be better than the LMS ratio vis-a-vis under the models in question are exactly similar to those obtained under the super population set up. The author's model is also amenable to analytical treatment as has also recently been established by Cassel and Sarndel (1972).

7. For the purpose of illustration some populations are generated so that the g model for discrete ancillary variate holds. For estimability of the mean of populations so generated on the basis of a sample of size 4 the relative performance of the unordered estimators of Des Raj and Das under PPS sampling without replacement, Horvitz-Thompson estimator under Durbin's IPP Scheme and the classical ratio estimator under LMS sampling was investigated. It has been concluded that in all situations where the regression of the study variable on the ancillary variable is linear and passes through the origin and the population size is not large, LMS ratio strategy is to be preferred unless there is a strong evidence that g lies between 1 and 2. The superiority of LMS ratio strategy to that of Durbin when $g=2$ is of special significance in that it does not confirm the conclusion of Godambe (1955) drawn under super population model.

B. D. Tikkiwal :*

Sampling from a finite or infinite Population on Successive Occasions

8. The purpose of this paper is mainly to review the theory of single stage univariate successive sampling starting with the initial work of Jessen (1942) for sampling on two occasions, its generalisation for more than two occasions by Yates (1949), Patterson (1950) and Tikkiwal (1950) have been discussed. The assumption on common units between any two occasions as considered by Tikkiwal is observed to be more general than those by Yates and Patterson.

9. The above studies were carried out under the assumption that the population under study is infinite and the various correlations and regression cofficients occurring in the estimator are known.

* University of Rajasthan, Jaipur.

For infinite population when various constants involved in the estimator are not known and are estimated from the sample the results were developed by Tikkiwal (1956) under the following normality condition. Let $\underset{\sim}{Ye}$, $(Y_{e1},\ldots\ldots Y_{eh})$, where Y_{ei} denotes the variate value of the unit drawn on the ith occasion for $i=1, 2,\ldots\ldots,h$. The $\underset{\sim}{Ye}$ for different e is assumed to follow independently a non-degenerate h-variate normal distribution with mean vector whose ith element is μ_i, the population mean to be estimated on the ith occasion and a certain covariance matrix whose diagonal elements are σ^2_i the variance of Ye_i. When the population is finite and constants are known the corresponding results follow from the results developed by Tikkiwal (1967) while working on multiphase sampling.

10. The theory of regression and double sampling estimation, for finite populations when the regression coefficient is estimated from the sample has been derived by Tikkiwal (1960) by assuming that the bivariate finite population is itself a random sample from a non-degenerate normal population. This approach has been extended by Prabhu Ajgaonkar (1962) to derive the results in the case of univariate successive sampling from a finite population when the different constants are estimated from the sample.

11. All the results discussed above have been obtained under a specific correlation model as considered by Tikkiwal (1950). The results presented by Prabhu-Ajgaonkar and Tikkiwal (1961), on univariate sampling on successive occasions under arbitrary correlation pattern are discussed here in detail. It is shown that various properties of the MVLUE as developed for specific correlation pattern, still hold under arbitrary correlation pattern provided the common units on occasions, greater than or equal to two, form a sub-sample of the new units on the preceding occasion i.e. units are common only between two consecutive occasions. When this condition does not hold good, it is shown that the estimator no more remains minimum variance estimator, but it is still a convenient estimator to use. Therefore, an exact variance formula for this estimator is obtained under an arbitrary correlation pattern.

12. The various results due to Eckler (1955) on rotation sampling can be obtained as special case of those of Patterson (1950) and so of those of the author, since the Patterson's relevant results are special cases of those of the author. The rotation sampling as considered by Hanson, Horvitz, Nisselson and Steinberg (1955) is different from that of Eckler and is studied when the population under study is infinite. Rao and Graham (1964) considered a more

general pattern of rotation sampling than that of Hanson et-al. for finite populations. They gave the variance of their composite estimator under the general situation and tabulated the efficiency of the estimator for the gain in efficiency of the estimator for optimum Q over the efficiency of the classical sample mean estimator for different r, p, and m where Q is the weighting coefficient occurring in the composite estimator, r is the number of occasions a unit stays in the sample after its selection, ρ the correlation between any two consecutive occasions and m is the number of occasions a unit stays out of the sample before its re-entry. It is seen that the gain in efficiency of the estimator for fixed ρ and optimum Q is maximum for $r=2$, $m=\infty$. For these values of r and m, the rotation sampling becomes a special case of the situation where the matched units on an occasion form a subsample of the new (unmatched) units on the previous occasion

13. Rao and Graham's results regarding the efficiency of the composite estimator of the change of the population mean from the preceding occasion to current occasion is of interest. It is observed that efficiency of the composite estimator rapidly increases with the increasing value of r and decreasing value of m.

*N. S. Sastry** :

Major Developments in Theoretical and Experimental Research on Non-Sampling Errors

14. The classical theory of sampling is based on the assumption that whenever a unit is included in the sample it contributes a unique value. In practice this is not true and errors in response, measurement or processing introduce errors which together are described as non-sampling errors. Considerable researches on these problems, both theoretical and experimental, have been conducted in the last three decades.

15. Failure to collect data on some of the units selected in the randon sample frequently gives rise to biases. The remedy lies in making repeated attempts to collect data on a subsample of the non-respondents and researches in this field by Hansen, Hurvitz, Deming, etc. were aimed at examining alternative procedures for call backs and subsampling fractions at successive stages to arrive at the optimum strategy. El-Badry suggested several waves of questionnaires to be mailed to non-responding units as a cheap way of increasing response until further wave is not effective. Foradori gave a generalisation of the approach to multi-stage designs. Ericson

* Central Statistical Organisation, New-Delhi.

suggested a sequential procedure in which optimal second sample is drawn conditional on the results obtained from the first (initial) sample. Srinath, Rao and Ghongurde studied the problem further. Dalenius showed the possibility of selecting a subsample of non-respondents in such a way that the time lag between the first and the second survey is eliminated. Hartley, Politz and Simmons developed methods more efficient than call-backs in large samples when biases from early calls are substantial. Bartholomen developed a technique useful in interview surveys in which the enumerators arrange for the second call subsample from the not at homes during the first visit.

16. If similar surveys are repeated frequently, a method proposed by Kish and Hess which consists in adding to the sample a subsample of non-respondents from an earlier survey may be economical. Whenever bias from early calls shows a systematic pattern adjustment for bias may be made using extra-polation methods etc. as discussed by Herdricks, Clansen and Ford and Finlener. Fellegi reported that in Canada computer methods are being developed to identify non-responding units and carry out imputations for them.

17. Omissions, duplication and inaccurate recording of data introduce imperfections in the frames which give rise to non-sampling errors. Two papers by Hansen and others and Szameital and Schaffer may be regarded as major contributions in the field of problems presented by imperfect frames.

18. Refusal to respond or intentionally misleading replies are more frequent when the respondents are questioned about sensitive or highly personal matters. Warner developed an interesting interviewing procedure designed to reduce or eliminate such biases. The respondent is asked to choose randomly one of the two statements viz, I am a member of A Group or not and answer Yes/No truthfully. The interviewer does not know which statement is affirmed or denied and thus the privacy of the raspondent is protected. Provided the respondents follow the procedure intelligently and truthfully a sample of replies provide unbiased information regarding the character under study. Greenberg brought out and measured the advantage in terms of efficiency accruing from Warner's technique. Abdul-Latif, A. Abut Ela and others extended the technique to the general case where the population is divided into more than two groups. The method was applied in North Carolina to estimate proportion of mothers in the groups married, premature and unmarried mothers from among those reporting live birth

between October, 1964 and October 1965. The study indicated that the technique is potentially more efficient than direct interview. The method was modified by Simmons and later developed by Greenberg in which the respondent is to choose confidentially one of two questions, one sensitive and the other innocuous, and then answer it truthfully. Greenberg further extended the technique from the situation where the reponse is categorical to where it is quantitative.

19. Erikson recently gave another version of the randomised response model for quantitative data. Gould and others developed model taking into account the possiblities of respondents failing to respond according to plan. Warner discussing the subject later established a general linear randomised response model and showed that all existing randomised response procedures are special cases of the general linear model. James Press suggested an alternate approach based on direct questioning coupled with a procedure for assessing the probability that the respondent is lying and on this basis adjusting for the bias. Hendricks suggested a Bayesian approach for the estimation of category, probabilities when certain a-prior information was available which might reduce the error bias caused by unreliable data.

20. Errors in recording numercial data may depend on enumerator and give rise to enumerator variability. The method of inter-penetrating samples developed by Mahalanobis permits testing significance of this effect by means of analysis of variance. A systematic theoretical treatment of the problems was given by Hansen, Sukhatme and Seth. Uncorrelated response errors reduce the precision of estimates, however if all units are affected by a bias no method is available for revealing such a bias. Fellegi analysed further the model due to Hansen and developed estimators for several parameters of the model and derived their biases. Koch and God Nathan studied multivariate model analogous to Hansen's. Both Madow and Desraj considered the use of sub-sampling primarily to reduce response bias. Koch also demonstrated that the same principles may be used to develop response error models for certain non-linear estimators in 2×2 contingency tables. Cochran studied effects of measurement errors on multiple correlation. Chai explored the effect of correlated measurement errors on the estimator of the linear regression coefficients.

21. Sometimes survey responses are compared with more accurate records on a case-by-case basis for checking. However errors may occur in matching. Neter et. al. developed two simple

models to study effect of matching errors. The two models, when applied to the data from a record check study indicated that relatively small imperfections in the matching process can lead to substantial bias in estimating the relationship between response errors and true values. The models may be applicable whenever data from different sources are being matched. The models may be extended in a number of directions.

22. In large scale sample surveys or censuses errors through various sources crop in and tests must be designed to indentify serious errors to take remedial measures, such editing and correction or imputation is considerably facilitated by the advent of computers, which can apply complex criteria for testing and carrying out the corrections rapidly and cheaply. Of late much effort is being spent on estiblishing rational principles of automatic detection and correction of non-sampling errors. Nordbatten summarised principles along with existing practices in this area and proposed some directions in which research should proceed. Szameitat and Zindler have presented a theoretical study of various correcting methods. The authors proceed on the assumption that the units in which errors occur are a random sample of the total sample units. In the rough methods of correction all the erroneous units are rejected and estimates are based on the remaining units by altering the weighting factor in the estimation procedure. This results in some waste of data. The refined methods considered and compared are (*i*) the Monte Carlo method, (*ii*) the 'cold deck' and (*iii*) the 'hot deck' methods. Conditions under which each of the refined methods works satisfactorily were given.

23. Considerable researches on the sources and methods of control of non-sampling errors have been made in the field of crop cutting surveys, notably by the Indian Statistical Institute under the leadership of Mahalanobis and by the Institute of Agricultural Research Statistics under the leadership of Sukhatme and Panse. A critical review of the experimental evidence gathered has been given by Zarkovich.

24. Valuable experimental data on the sources and magnitude of non-sampling errors in household surveys have been collected by the Indian National Sample Survey Organisation and U.S. Bureau of Census as on effect of length of reference period on the quality of data, effect of question length and form and schedule structure on completeness and accuracy of response and effects of reporting load on enumerators, work and respondents' cooperation. Experimental

results obtained in various house hold surveys have been summarized by Zarcovich, Som and Fellegi.

25. Brizker et. al. and Stuart summarised the experience of U. S. Bureau of Census and the U. S. Bureau of Labour Statistics on editing correction or inputation methods. Norbatten's simulation study on the efficiency of automatic detection or correlation of errors in individual observations and Fellegi's discussion on the same are also of major interest.

Daroga Singh :*

Applied Research in Sample Survey Techniques

26. The development of sample surveys as a procedure for describing large populations is certainly a story of success, especially for social sciences. Today it is essential to have a deep knowledge of the socio-economic structure of any country for running sound administration. Agriculture and industry play deminant role in the formulation of entire economic structure and therefore, it is also essential to have at hand rational and practical techniques in the collection of appropriate data.

27. In the early part of the present century, Fisher and Karlpearson gave a practical shape as to how to interpret about the population from a given sample. However, the development of modern random sampling technique started with Neyman (1934), who pointed out the superiority of random sampling procedures, over the non-random one. During the last few decades, many organisations have contributed substantially to the modern theory of sampling through the applications of sampling techniques to many important practical problems. Some of the most notable institutions are the Bureau of Census and Survey Research Centre, State University of Michigan in U.S.A., London School of Economics and Politics and Rothamsted Experimental Station in U.K., Indian Statistical Institute, Calcutta, Institute of Agricultural Research Statistics (I.C.A.R.) New Delhi in India, Statistics Division of the Food and Agricultural Organisation of the United Nations, Rome etc. In the present article different problems arising in survey sampling, like choice of selection procedures and method of estimation, the choice of items to be included in the survey and non-sampling errors, particularly in the field of agriculture, have been discussed.

28. In the application of sampling techniques for collection of data relating to agriculture, a number of investigations for estimation of area and yield of different crops were undertaken firstly by

* I.A.R.S., New Delhi.

Mahalanobis (1938-46) and subsequently by many research workers at the Institute of Agricultural Research Statistics. In these investigations, the efficiency of various methods of stratification, optimum size of samples and appropriate size of the sampling units were discussed. Since cultivation and harvesting practices differ considerably from crop to crop, modifications in the sampling techniques for different crops have also been suggested. Investigations have also been undertaken for estimation of area and yield of fruits and vegetables. Surveys have also been undertaken for determining the response of various agricultural inputs under the farmers conditions. And since, this response depends upon the fertility of soil, suitable sampling techniques for taking the soil samples from the field have also been evolved.

29. The problems of sampling and measurement, in case of live-stock numbers and live-stock product, are different from those encountered in crop estimation surveys. A large number of investigations have been undertaken at I.A.R.S. for evolving suitable sampling methodology for live-stock products like milk, eggs, meat and wool. Appropriate selection procedure vary from survey to survey. Systematic sampling has been used extensively in the forest surveys by many workers, like Hasel (1943), Finney (1942, 48, 50, 53) Griffith and Mokashi (1954, 1945, 1946), Mokashi (1954).

The use of systematic sampling for estimation of the catch of marine fish was well demonstrated by Sukhatme, Panse and Sastry (1958) and the usefulness of this sampling technique for estimating the lactation yield of cows has been shown by a large number of studies undertaken at the I.A.R.S. Padam Singh (1970) suggested a procedure for estimating the variance of the mean of a systematic sample.

30. On account of operational convenience and cost, cluster sampling and sub-sampling have been most commonly used in sample surveys. I.A.R.S. made use of this sampling procedure mainly in live-stock and fruit and vegetable surveys and surveys undertaken to study the impact of milk supply schemes on rural economy, Goel (1973) made extensive study of problems arising in overlaping clusters. Since the efficiency of cluster sampling decreases as the size of cluster increases sub-sampling within each cluster enhances the flexibility of the sampling procedure. In fact almost every large scale survey used multi-stage sampling design or stratified multi-stage design. In numerous sampling investigations conducted at I.A.R.S., efficiencies of cluster sampling and sub-sampling have been extensively studied.

31. Another sampling technique known as aerial photheraply though costly and therefore not commonly used was tried for determining the land use and collecting area statistics in Goalpara district of Assam (India) which is not easily accessible by land transportation. However not much use of this technique has been reported in the field of agriculture.

32. When the units vary much in their sizes, the technique of selection of units with varying probabilties is expected to provide more efficient estimates. Though the technique of varying probabilities was introduced by Mahalanobis only in 1938, the idea was quite fascinating and lot of theoretical as well as empirical investigations have been undertaken by so many workers. But still the varying probability sampling schemes are not very popular with the survey practitioners mainly because of lack of simplicity both, in respect of selection of sampling units and estimation of parameters.

33. The ratio and regression methods of estimation have been commonly used since their introduction because of simplicity and their utilizing the ancilliary information quiet efficiently. With the increasing use of sampling techniques in all fields like, agriculture, economics, sociology, industry, etc, it is not only advantageous but sometimes very necessary to include a number of related items in a single enquiry. Though a number of problems faced by the sampler in surveys like that of stratification, optimum allocation of samples among strata, or optimum probabilities of selection of units, etc., become more complex in such surveys the use of multi-subject surveys has gained popularity and become a necessity in many fields. In India National Sample Survey Organisation undertook different socio-economic surveys to satisfy the demand of different social development programmes.. At I.A.R.S. data on a number of items are collected in many surveys especially in surveys for assessing the progress of Intensive Agricultural development programmes and the High Yielding Varieties programme. Effciencies of the alternative sampling design in relation to intensive Agricultural Development Programmes have been empirically examined by Singh and Singh (1973).

34. With the increasing awareness of the quality of data collected through surveys, the liabilities of statistician for a careful planning, proper handling of the field work and processing of data supported by appropriate quality checks has much increased and a lot of investigations, throughout the world have been carried out to obtain data free from non-sampling errors. Special mention may be

made to interpenetrating sub-samples introduced by Mahalanobis (1958) and costa ccounting methods of data collection by Panse (1955). A successful survey statistician is one who has some knowledge about the population and who makes the best use of this knowledge in the efficient planning of the survey. In a survey, the important constituent elements may be objectives of the survey, sampling frame, method of data collection, choice of sampling design, etc. Mahalanobis referred to this disintegration of elements and later their coherent assembling in the most efficient way as "Statistical Engineering" considerable work has been reported on studies of different constituent elements of sample survey. Pilot surveys, wherever necessary are undertaken to arrive at a suitable sampling design.

35. Lastly, some fields of study where not much work has been done and which need attention of the survey statisticians are development of sampling methods (*i*) to assess effect of pollution by industrial waste (*ii*) the use of chemicals in agriculture (*iii*) for estimation of non-sampling errors, especially in multi-factor surveys (*iv*) for assessment of fisheries and forestries, etc.