SOME THOUGHTS ON STATISTICAL INFERENCE ABOUT FINITE POPULATIONS*

By
G. R. Seth

Introduction

The simplest form of statistical inference is making statements about a population on the basis of a sample, such statements being accompanied with some numerical measure of their reliability. When the latter is absent, still it may be called scientific inference but it will not be statistical inference. This means that for statistical inference a population must be precisely defined as well as the sampling rule. This is surely happening when one is dealing with actual populations in survey sampling and this is its distinguishing feature as also Barnard remarked at the symposium on Recent Developments in Survey-sampling, held at Chapel Hill, W. C., U.S.A. in 1968. To quote him: "A distinguishing feature of some of the populations which are dealt with in survey sampling is that they have a real existence (not excluding super populations in a certain sense). So when we talk of a random sample of such a real population we have in mind a physical process like that of drawing balls from an urn. And we surely do accept that 'random' is an attribute which can meaningfully be attached to a physical process". Thus survey statisticians are concerned with the study of finite or 'actual' populations, whereas 'theoretical statisticians have developed statistical inference' mainly for infinite hypothetical populations. General statistical theory so developed has not so far been fully utilised in the developments of methodology of inference about finite populations called sometimes sample survey methodology. This is not surprising as theoretical results generally do take some time before these can be put to any practical use. But recently doubts have been raised on the optimality of the survey sampling methods developed and practiced and it may be worthwhile to examine the extent to which these doubts are justified and the extent to which it is

* Technical address delivered during the 33rd Annual conference of the Society held at Kerala Agricultural University, Manuthy on 21st December, 1979.

possible to apply the statistical inference methodology developed on the basis of infinite population models to sample survey theory. Probability sampling is the fundamental tool in sampling studies and this approach is also being questioned especially by the Bayesians as randomness and the design of an experiment or of sample survey does not find any utility in statistical analysis in Bayesian theory of statistics. There are also certain aspects of survey sampling theory which should receive much more attention by survey statisticians in order to increase the effectiveness of sampling methodology. I will like to touch upon some of these aspects in my presentation. Statistical inference about finite populations is a much broader aspect than that relating to only sample survey theory but my remarks will be confined to only survey sampling.

Developments in Survey Sampling

To begin with, let us look at the developments in survey sampling. Some cases of use of sampling for inference about actual populations have been reported before the nineteenth century but it was recognised only in the nineteenth century that one could make inferences concerning an actual (finite) population characteristics such as its mean on the basis of sampled observations with the help of calculus of probability. For example, Laplace is reported to have utilised the sampling approach for the estimation of the total human population in France. In 1906 Bowley has reported on the status and application of this method prevailing at that time. Broadly speaking complete enumeration was the method generally utilised for the study of finite populations, and where sampling was resorted to, it was generally a 'purposive' sample. It was the epoch-making paper of Neyman (1934) which established the superiority of the 'random' sample over the 'purposive' sample. For the first time, he formulated the theory of stratified random sampling, cluster sampling and the use of 'confidence intervals' for making statements about the various characteristics of the population under study. He also utilised the results of the Gauss-Markoff theorem developed for infinite populations in the general statistical theory to obtain uniformly minimum variance unbiased linear estimates for finite populations. After the publication of this result there were rapid developments in the theory and application of random sampling methodology. A distinct new feature of sampling with varying

Neyman, J. (1934) : On the two different aspects of the representative method—the method of stratified sampling and the method of purposive sampling. *J. Roy. Statis. Soc.* Vol. 97, P. 558-625.

[1] Hansen and Hurwitz (1943): On the theory of sampling from finite populations. *Ann. Math. Statist.* Vol. 14, P. 333-362.

probabilities was introduced by Hansen & Hurwitz (1943) for the development of more efficient sampling designs. In forties, great progress was made in sampling theory for the development of designs which were to provide estimates of characteristics with minimum variance for given resources or the use of minimum resources for obtaining estimates with given margins of error. With the increased use of sampling in the study of populations, survey statisticians were faced with the questions of non-response and measurement errors. For the control of enumerator's biases and their measurement, interpenetrating checks were introduced in India for the first time by Prof. Mahalanobis in sample surveys undertaken by Indian Statistical Institute which become later a regular feature of the National Sample Surveys in India and this method is being used in various other countries in the world. Problems of non-response were initially studied by Hansen and Hurwitz (1946). Models for the study of measurement errors were formulated by Hansen, Hurwitz and Marks (1951) and Sukhatme and Seth (1952). In 1952 Hurwitz and Thompson developed a general linear estimate which took into cognisance of the 'label' of the unit and/or structure of the sample of units drawn. Earlier general linear estimation theory was based by ignoring the 'labels' of the units. In 1955 Godambe made a fundamental contribution to the sampling theory by indicating that application of Gauss-Markoff theorm to inferences about best linear unbiased estimates for finite populations was strictly not appropriate. Following up the general linear estimate given by Horvitz and Thompson he indicated that more general type of linear estimates were possible for finite populations than that for the infinite populations (See Appendix I). He proved that for this more general class of linear estimates, there did not exist a uniformly minimum variance unbiased linear estimate of the population total for finite populations. This result was later generalised by Godambe and Joshi (1965) to prove the non-existence of uniformly minimum

[2] Hansen and Hurwitz (1946) : The problem of non-response in sample surveys. *J. Amer. Stat. Assoc.*, Vol. 41, P. 517-529.

Hansen and Hurwitz and Marks (1951) : Response errors in surveys. *J. Amer. Stat. Assoc.*, Vol. 46, P. 147-190.

Sukhatme and Seth (1952) : Non-sampling errors in surveys. *J. Ind. Soc. Agri. Stat.*, Vol. 4, P. 5-41.

Godambe, V. P. (1955) : A Generalization of sampling without replacement from a finite universe, *J. Amer. Stat. Assoc.*, Vol. 47, P. 663-685.

Godambe, V.P. (1955) : A unified theory of finite populations. *J. Roy. Statis. Soc.* B, Vol. 17, P. 269-278.

Godambe, V.P.. Joshi, V.M. (1965) : Admissibility and Bayes estimation in sampling finite population. I—*Ann. Math. Statist.*, Vol. 36, P. 1707-22.

variance estimate (linear as well as non-linear) for finite populations. This result attracted the attention of theoretical statisticians and since then they have paid considerable attention to the foundations of sample survey theory. Work has been done both in the interpretation of survey sampling theory in a way that it would fit within the framework of the general statistical theory and also by extending the statistical theory with new models and corresponding formal criteria of optimality and appropriateness. Among the many workers contributing to the developments to the foundations of the survey sampling theory, mention may be made of Basu, D., Ericson, Godambe, Hartley and Rao, Pathak, Royall and Sarndal. Bayesian as well as non-Bayesian approach have been adopted in these developments. This has raised considerable controversies regarding some basic issues and these are not as settled nor there is any hope of their settlement. But, I am sure, these will all lead to a very healthy development of the subject. While the sample-survey theorists were and are busy straightening out the kinks in the sample survey theory, the traditional survey statisticians like Cochran, Hansen and Hurwitz with their workers at US Bureau of the Census, Kish, Yates, Hartley and Rao, Panse and Sukhatme and workers at Indian Statistical Institute and Indian Agricultural Statistics Research Institute (IASRI) went on developing in their own pragmatic way better and cheaper sampling designs and considerably strengthened the practical base of survey sampling. One may make a note of the swing towards very restricted randomization in the choice of units almost bordering on 'purposive' selection. The effort is to make the realised sample as much 'representative' of the population as possible but having enough randomization to provide estimates with reasonable margins of error. In this connection, special mention may be made of the designs developed at IASRI for estimating marine catch, of integrated surveys for estimation of livestock products, fruits and vegetables, cost of production studies, large scale fertilizer field trials and statistical assessment of Agricultural Development projects. For the marine fish catch surveys, a combination of the random and systematic sampling over space and time has been developed to yield reliable estimates of catch. .In large scale fertilizer trials knowledge of sample survey practice and experimental designs were suitably integrated to get the response of fertiliser applications under actual farmers' conditions. Sampling designs were simple enough for operation under difficult field conditions. Most of these surveys have been of an enumerative type. Very little analytic work has been attempted. Other developments worth mentioning is the work initiated by Patterson (1950) and Tikkiwal (1951) on the designs to deal with sampling on succesive occasions, controlled selection as well as balanced sampling by Kish and randomised response models initiated by Warner (1965). Work on the construction of optimum number of strata, stratification points and allocation problems was also intensified. Attempts have also been made for development of optimum designs for the study of multivariate characters and stratification for multiple characters. It is not possible for me to give a detailed review of the recent developments. A review paper by Delanius (1962) summarises broadly the position at that time and the proceedings of the symposium 'New developments in survey sampling' held at Chapel Hill in 1968 provide a good summary of the developments in this fields. Work on non-sampling errors has not made much progress and this work has been reviewed in a recent symposium on 'Survey sampling and measurement' held at Chapel Hill again in 1977.

Distinction between 'actual' and 'hypothetical' populations:

Infinite populations are hypothetical with their distributions depending upon a number of parameters in the parametric case, whereas 'actual' populations are real. The latter can also be studied on the basis of a complete enumeration, sampling is artificially introduced only to save on resources of time, money and personnel whereas infinite populations can only be studied through 'realised' sample of units. A chance mechanism is supposed to operate in the case of infinite populations yielding the 'sample' observations whereas sampling of finite population is done artificially. A sample design is developed for drawing a sample from a finite population such that any sequence (s) of units has given pre-assigned probability, $p(s)$, such that $p(s)\geqslant 0$ and

$$\sum_{all_s} p(s)=1.$$

Finite populations are usually 'labelled' or can be 'labelled' so that each unit is identifiable, so also a set of units called the 'sample' is identifiable. We can impose any sampling design on the finite

Patterson, H.D. (1950) : Sampling on sucessive occasions with partial replacement of units, *Journal Royal Stat. Soc.* Series B, Vol. 12, P. 241-255.

Tikkiwal, B.D. (1951) : Theory of successive sampling, unpublished Thesis for Diploma. ICAR, New Delhi.

Warner, S.L. (1965) : Randomised response—A survey technique for elimination of evasive answer bias *JASA* Vol. 60, P. 63-69.

Delanius, T. (1962) : Recent developments in sample survey theory and methods. *Ann. Math. Statistics.* Vol. 33, P. 325-349.

population, whereas the chance mechanism which gives rise to a sample for an infinite population is hypothetically supposed to be fixed though unknown. In a sample from a finite population, there can be repetition of units whereas this is not so possible in case of infinite populations. Usually a finite population is parameterised by making the unknown values of the units constituting the population as parameters *i.e.* if the finite population contains N units, then the parameter space is a N-dimensionenal space for each character. This is a rather unusual formulation for all the observable values are considered as parameters. Thus if we want to know all the parameters, we have to enumerate the population completely. The sample space is generated by various samples $[s, x_{i1}, x_{i2}, ... x_{in}]$ where s is the sequence of $(i_1, i_2 ..., i_n)$ labels drawn and $x_{i1}, x_{i2} ...... x_{in}$ are the observations made on the character X under study, has a close relation with the parametric space. After the draw of a sample the components of parameter space corresponding to the units drawn in the sample become known and the dimension of the unknown parameter space is reduced by k, the number of distinct units drawn in the sample. More the number of distinct units drawn, more the number of components of the parametric space revealed. Ultimately when all of the population are observed in the sample space, the dimensions of the unknown part of parametric space reduces to zero and the parametric space reduces to a known point. This reduction in the range of the parameters in case of infinite populations is not possible except in the case of irregular distributions where the range of variation of observable depends upon the parameters like a rectangular population with range $(\theta, 1+\theta)$. Ignorance of this distinction has led Basu (1971) to give a somewhat wrong simple proof of the non-existence of uniformly minimum variance unbiased estimates. This proof runs as follows :

"Let $\theta^*(s)$ be the uniformly minimum variance unbiased estimate of $T(\theta)$. Let θ^*_0 (s) be its value at $\theta=\theta_0$ where θ is the parameter $[X_1, X_2 ...... X_N]$ and θ_0 is a specific value of θ, then $\phi(s)=\theta^*(s)-\theta^*_0(s)+T(\theta_0)$ is an unbiased estimate and $E\,[\phi^*(s)=T(\theta)$ for all θ. But at $\theta=\theta_0$, $\phi^*(s)$ reduces to $T(\theta_0)$, a fixed constant, and thus it has zero variance at $\theta=\theta_0$. Thus corresponding to any unbiased estimate, we can construct another unbiased estimate which has zero variance at any given $\theta=\theta_0$. Thereby proving the non-existence of UMV unbiased estimate". But the error in the

Basu, D. (1971) :Foundations of statistical inference Edited by Godambe and Sprott—Holt, Reinhort and Wintson P. 222.

argument is that $\phi^*(s)$ as constructed is only a constant equal to $T(\theta_0)$ and $\phi^*(s)$ a biased estimate except for $\theta=\theta_0$. For $\theta^*(s)$ and $\theta^*_0(s)$ are exactly the same as $\theta(s)$ depends only upon the revealed components of $\theta=[X_1, X_2,\ldots\ldots X_N]$ in the sample and thus its value for any θ_0 (consistent with s) does not change.

Labelling of units has another implication. A unit can only speak far itself and cannot give information on other units, except that a function of the observed components in a random sample can be used to estimate a function of the unknown unit *e.g.* $\bar{X}_n$, the sample mean of units, observed in a simple random sampling (without replacement) can estimate the mean of the remaining N-n units as $\bar{X}n$. This helps to estimate the total of the population as

$$\mathrm{n}\bar{X}_n+(\mathrm{N}-\mathrm{n})\,\bar{X}_n=\mathrm{N}\bar{X}_n$$

In finite populations, as mentioned earlier, estimates especially linear estimates can be formed by giving a weight to each unit observed depending upon the serial number of its draw, label of the unit and the structure of the sample actually drawn, where as for infinite populations only serial number of the draw for samples observed has any meaning. Thus a broader class of linear estimates can be formed for finite populations against those for the infinite populations based on weights dependent on serial number of the draw only. In case we ignore the labels, $(i_1, i_2,\ldots\ldots i_n)$ and structure of the samples is also ignored, the class of linear estimates, reduces to the class of linear estimates for infinite population. Prof. Neyman had probably this in mind when he extended the scope of the application of Gauss-Markoff theorem to unbiased linear estimates for finite populations. This view is supported by considering the various T-classes (7 in number) proposed by Tikkiwal (see Appendix I). Weighting coeff. in T^1-class estimates depend only upon the serial number of the draw where as the other 6 '*T*-classes' depend also on labels and/or structure of the sample. For T_1-class, uniformly minimum variance (UMV) linear unbiased estimates do exist for some sample designs for which unbiased estimates exist. Sample mean, for example, is UMV linear unbiased for simple Random Sampling (SRS) without replecement. Also subsequently Cochran, Sukhatme and other authors have ignored the labels while forming linear estimates there by studying estimates in the T_1-class of Tikkiwal. Further, the concepts like that of unbiasedness, admissibility, sufficiency, minimal sufficient statistic, likelihood function, maximum likelihood estimates, strictly developed for infinite populations in the general statistical theory are also applicable to survey sampling of finite populations. However, some concepts like those of asymptotic efficiency, consistency, limit theorems, linear regression and correlation coefficients are not strictly applicable to finite populations. Survey satisticians have at their back of their mind assumption of normality of the mean or its asymptotic normality when they give confidence intervals for their estimates. So also there is an implicit assumption of an infinite population when they use the linear regression technique to build resgression estimates. Perhaps there is a need to develop suitable finite population theory of regression and correlation. In this connection one may say that the blind use of the assumption of normality or central limit theorem can lead to misleading results. It has been shown by some Monte-Carlo studies that in case exact confidence intervals based on criteria derived from the sample survey desigen are not utilised, the results obtained by using normal theory can be very much off the mark. One such study by Stenlund and Westlung (1975) concerned the study of three populations of size 200 with differing variences and skewness and three sample designs : simple random sampling, sampling with varying probabilities (three types) and stratification (strate 2 or 3 or 4) and sample size 4, 12, 20 and 30 indicated that the hypothesis of normality was justified only in a few cases and that too mostly when sample size is 30, With the increase in skewness, the actual confidence coefficient differed greatly from 95% (or 99%) coefficient fixed for normality-based (confidence intervals. Another important distinction between finite and infinite populations is that observation obtained can be generally verified for their accuracy by re-sampling of the finite populations as is done in post-enumeration checks.

In finite populations for sampling with replacement one can use a sufficient statistic which is not minimal sufficient statistic for getting an estimate better than the estimate based on the observed sample only. This will be useful where it is difficult to obtain the estimate as a function of the minimal sufficient statistic. This happens because the probabilities of selection of all possible samples are known once the sample design is fixed. This has been shown by Tikkiwal. a mathematical formulation for which is provided in the Appendix II. This is not usually possible in the case of infinite populations.

Non-existence of optimal estimates for finite populations :

Sarndal (1972) has examined the implication of the above assertion by Godambe. To quote, he writer 'In spite of rigour of presentation, statistical public may feel somewhat hesitant about the

Stenlund, H. and Westlung, A. (1975) : Bull. Inter. Stat. Inst., Vol. 46, No. 4, P. 365-368. A Monte-Carlo study of three sampling designs.

Sarndal, C.E. (1972): Sample survey theory vs general statistical theory, estimation of the population mean. *Satis. Ints. Review*, Vol. 40 P. I-II.

proper content of much publicized, but rather elusive phrases like 'there do not exist optimum estimators in sample survey theory'. However, it seems unfortunate to the present author if sample survey theory were to acquire the image of an area of 'General non-existence of optimum estimators quite in contrast to 'General Statistical Theory'. He shows that the difference between the two theories with regard to optimum estimation are possibly not that great after all. He cites the example of estimation of population of all absolutely continuous distributions (F) with regard to Lebesgue measure and having finite mean μ and variance σ^2, $\xi_1, \xi_2, \ldots\ldots, \xi_n$ are n observations from F. Then $\xi_{(1)}, \xi_{(2)}, \xi_{(3)}\ldots\ldots\xi_{(n)}$ (the ordered sample of $\xi_1, \xi_2, \ldots\ldots\xi_n$) is known to constitute a complete sufficient statistics. $E(\zeta_1/\zeta_{(1)}, \xi_{(2)}\ldots\xi_{(n)}) = \sum_{i=1}^{n} \xi_{(i)}/n = \sum_{i=1}^{n} \xi_i/n$ and is uniformaly minimum variance unbiased estimate and a fortiori, the sample mean is also uniformly best linear unbiased estimate. But if we restrict the distribution to distribution functions completely known except for certain parameters, say $F\left(\frac{x-\mu}{\sigma}\right)$, where F is a distribution of specified form with μ and σ being the location and scale parameters both unknown. Lloyd (1952) has shown that a uniformly best linear unbiased estimate (UBLUE) for estimating μ can be developed which has a smaller variance than the sample mean for various values of μ and σ^2 for all n with equality holding if and only if the sample mean is identical to the UBLUE.

As mentioned earlier, Gauss-Markoff technique is not equipped to handle successfully the more general class of estimators which are generated following the work of Horwitz and Thompson, Godambe, Koop and Tikkiwal. The non-existence theorems of Godambe (1955) and Godambe and Joshi (1965) show that conventional optimum properties cease to exist as the class of estimators is made sufficiently wide. Kempthorn also comments on the artificial example given by Godambe to indicate the non-existence of Best linear unbiased estimates, saying that the example is sterile from a practical point of view, since rarely one would have the prior knowledge of the inequality in the parameters needed in order to know when the suggested estimator is superior to the sample mean. If there is a prior knowledge of any inequality (or any relationship) between the parameters, then this knowledge can be used to search for a better alternative to the convetional estimators.

Llyod, E.H. (1952): Least square estimation of location parameters using order statistics, *Biometrica,* Vol. 39, P. 88-95, Sarndal (1972) gives such an example :

Let $x_1, x_2......, x_N$ be the units with their x-values. Suppose it is known that x_N considerably exceeds the population mean and x_1 falls much below the population mean such that $x_N - x_1 > A > 0$. Consider the following estimate :

$$e(s:x) = \bar{x} + \frac{A}{2} \text{ if unit with label 1 included in the sample but not label } N;$$

$$= \bar{x} - \frac{A}{2} \text{ if unit with label } N \text{ included in the sample but not label 1}$$

$$= \bar{x} \quad \text{if labels 1 and } N \text{ are both included or both excluded.}$$

Then $e\,(s, x)$ is uniformly better than $\bar{x}$ over the corresponding subset of parameter space corresponding to $\left(X_N - x_1 > \frac{A}{2}\right)$ and it has a relative gain of efficiency compared to the sample mean of no less than $\frac{A^2}{2n\sigma^2}$.

Sarndal concludes that in spite of the fact we can find estimators 'better than the best', in the case of general statistical theory, the concepts of General Statistical Theory are kept clear and unmuddled by non-existence theorems. This handicap being not peculiar to sample survey theory, sample survey theory alone should not be assailed on this account. On the other hand, we should try to find out estimators better than the mean in case actual population is known to belong to a known subset of the parametric space. This possibly should be the pragmatic approach in further developments in sample survey theory.

Interpretation of Survey-sampling theory to fit in with the actual Statistical Theory

Interpeting survey-sampling theory in such a way that it would fit within the frame-work (model) of the general statistical theory has been attempted by several workers. Basu (1969) says that the statistical model for sample surveys proposed in a series of interesting papers, has been so formulated as to confuse conventional statistical mathematicians into the belief that the analysis of survey type data falls outside the mainstream of the theory of statistical analysis. In these formulations, one sees a sample space S (of all possible samples

Basu. (1969) : Role of sufficiency and likelihood Principles, in sample survey theory, Sankhya Series A, Vol. 31, Part 4.

s) with just one probability measure p on S (as against a family of prob. measures for General Statistical Theory). The pair (S, p) is called the sampling design.. A typical sample $s \in S$ is a subset or a finite sequence with its members drawn from a fixed population of individuals 1, 2, 3,......N. The parameter is an unknown vector $\theta=(x_1.x_2......x_N)$, a statistic is a very special kind of function of the sample s and the parameter θ. In his view it is really not necessary to formulate the survey model in the above unfamilar manner. He formulates the survey model with the usual trinity [X, A, ρ_o] where X is the sample space, A is the set of all subsets of X and ρ_n the family of measures defined for all sets of A. Likelihood function so developed unfortunately is uniformative about the part of the parameter space not revealed by the sample and thus is a constant so far as unknown components of parameters are concerned and thus he comes to the conclusion that the person who is responsible for analysis of data need not even know the sampling design that produced the data. Even though sampling design may not be used in Bayesian interpretation of statistical inference but an efficient design involving the choice of observations will be necessary to produce optimum estimates. Basu very briefly examines the randomization principle and comes to the conclusion that there is very little (if any) use for it in survey design, and thus emphasizes the need for a careful design of a survey to get a good (representative) sample.

Our comments on Basu's formulation of the sample survey model is that no doubt it does bring it in line with the general statistical theory model but without making the likelihood function informative about the parameter space. He has merely reinterpreted the (S, p) model. According to (S, p) model, the probability of selection of the sample units do not depend upon the components of the parametric space with the result that likelihood function will be a priori uninformative about the parameters. (S, p) model has to be changed in its content in order to obtain informative likelihood function or some other approach for statistical inference should be adopted. Likelihood principle based on likelihood function even though an important tool in statistical inference is not the only one as ably demonstrated by Survey Statisticians. Informative likelihood functions can be developed by ignoring the 'labels' of the units as has been shown by Royall (1968). Similarly, Hartley and Rao (1968) achieve the same objective by introducing label independent or scale

Royall, R. (1968) : An old approach to finite sampling theory. Ann. Math. Stat. Vol. 63.

Hartley, H.P., and Rao, J.N.K. (1968) : A new estimation theory for sample surveys. *Biometrica*, Vol. 55, p. 547-57.

load estimators. These estimators are defined as mathematical functions which depend only on the sample frequencies attained by the various components of the scale vector $(y_1, y_2, \cdot y_T)$. It is assumed that a characteristic y can attain only T distinct values, y_t being attained by N_t unknown values $(t=1, 2, 3,...T)$. The parameteric space $(x_1, x_2...x_N)$ has been thus transformed into one of T-dimensions spanned by $(N_1, N_2,...N_T)$ and reduces the problem into a multinomial or hypergeometric distribution and the likelihood function becomes informative. This works as T is much smaller than N and the sample does not reveal the parameter space in the sense the other formulation does. Formulation of the likelihood function on the basis of scale load factors in fact can be justified as follows.

If for the actual population of all the N units were observed, as it is possible to do so, the population being real, then there is no statistical inference involved except the summarization of data. Now the summarization of data first attempted by ordering the units in ascending or discending order for continuous variates and then dividing this ordered set into a number of frequency classes. Number of frequency classes and their lower and upper limits can in fact be fixed in advance of any sample survey enquiry. Thus for the i-th class interval we have a frequency N_i (unknown) such that $\sum_{i=1}^{T} N_i = N$ where T is the number of frequency classes for the frequency distribution of the character Y for the population. With the formulation of the frequency distribution of X, the unknown parameters correspond to the unknown frequencies in the frequency classes $(N_1, N_2,..., N_T)$. Then Hartley and Rao's Y_i's are some typical values (mean or median) for each frequency class. Other attempts in this direction include. The one by Ericson (1969) who proposes a priori distribution assuming exchangeability and derives estimates for a stratified population which correspond to the usual ones based on stratified sampling estimation.

Extension of the Statistical Theory with a New Model for Survey Sampling :

As regards the extension of the statistical theory with a new model and corresponding formal criteria of optimality, attempts have been made by Godambe (1966) to provide new concepts like linear sufficiency, distribution-free sufficiency and concept of censoring to develop a new class of optimal estimates. Censoring techniques

Godmabe (1966) : A new approach to sampling from finite populations. *J. Roy. Statist. Soc.,* Series B. Vol. 28, P. 310-28.

prohibit the use of the knowledge of units not drawn in the sample for proposes of estimation. But so far it has not yielded any usable estimates other than those already known. In a way it provides justification for the use of the conventional estimates. But one main difficulty in their use is the verification in actual practice of the conditions prescribed for theoptimality of the estimate for real populations. Godambe has also suitably combined the Bayesian concept of a priori information and non-Bayesian concepts, like unbiased estimates, to produce hybrid estimates. Such estimates, however, are not welcomed by Bayesians. According to TIAO and Box (1973), while such an estimator serves some useful purpose, consideration of its sampling properties (like unbiasedness. etc.) has little relevance to Bayesian inference as the posteriori distribution of the parameter or parameters under consideration contains all the information about the parameter and this distribution should be appropriately summarised (described) for inferring about the parameter under study. Royall (1973) has used the polynomial regression model for obtaining regression and ratio estimates and under the condition of "balanced" samples, the model is robust for obtaining unbiased estimates, balancing being regarding the equality of the sample and the population moments upto k-th order where k ($k \neq 1$) is the degree of the polynomial regression. But in practice one is not clear about the method to obtain 'balanced' samples especially if we have to study simultaneously more than one character. Also the validity of the regression model for the full range of values of the auxiliary character x may itself be questioned and needs verification. J.N.K. Rao and others have also made studies assuming various types of super-population models.

Super Population Models:

Development of super-population models in sample survey theory is in line with the thinking of some of the statisticians (especially theoretical statisticians) that survey statisticians should fall in line with other statisticians like those concerned with comparative experiments in the use of super-population models for the study of finite populations. There is a pressure also from Bayesians regarding the utilisation of a-priori information to formulate a-priori distribution of the parameters for finding the posterior distributions of the parameters. Aggarwal (1959) was probably the first to construct Baye's estimates. Erison (1969) has developed a Bayesian approach for estimation for simple random sampling and stratified sampling.

Tiao, C.C. and Box, G.E.P. (73) : *American Statistician* Vol. 27, No. 1 P. 12-14.

Survey statisticians are hesitant to use the methods so far developed on the basis of the models and Bayesian teachniques. Cochran observes that theorists mostly are producing results for the theorists and the practical results can only be obtained in case these were developed in close collaboration with the survey statisticians.

Most of the estimates obtained by the Super-population models are the ones already in use and perhaps it provides justification for their use under a set of conditions postulated by the models and thus have not so far developed any new worth-while estimates. Also as regards the sample survey data, there are a large number of users and if one assumes a super-population model and provides estimates of parameters on the basis of the super-population, various users may not agree with the assumed super-population especially when the estimates are not robust with respect to the alternative super-populations making the results unacceptable to users there by making the super-population exercise useless and wasteful. Further statisticians concerned with comparative experiments have generally to apply their experimental results to future populations which they have not sampled and thus they have to use necessarily models for the future populations like the users of a new improved seed. Survey statisticians, on the other hand, generally have to describe the population actually sampled. Thus there is a genuine hesitation to lose a grip over the real population when one displaces it by a super-population *i.e.* when one has a tendency to talk of a population of populations instead of dealing with an actual population. Sample surveys also being generally multipurpose, construction and careful assessment of stochastic models for the analysis of a number of characters would be an immense undertaking. With the increase in the analytic uses of survey data, as well as in the conduct of surveys of a predictive nature and the use of non-random samples (like in medical studies), the use of models will increase in survey-sampling fields also. But a great care is to be done in their choice and these should be based upon the past available information as much as possible. Also for their use it will perhaps be necessary to draw a preliminary sizeable sample for testing the validity of the model, followed up by a large sample for estimating the parameters and or the 'causal' relationships. While modelling, one should somehow keep the identity of the population under study. The same arguments made above hold also against the adoption of Bayesian techniques. A-priori distributions are usually highly personal and subjective and its choice has a great influence on the posterior distribution of the parameters especially in case of small size samples. Bayesian estimates will thus be highly subjective and not acceptable to the various users unless they agree with the assumed a priori distribution. No doubt survey satisticians are using a priori information in a Bayesian sense to improve upon the design of the sample. This information is usually available on characters other than the character under study and is utilised in forming strata, selection of units with varying probabilities and in the use of ratio and regression estimates. Possibly a priori information available is not used to the full and there can be a case for its greater use in designing good samples and the randomization should be done only when all the priori information is utilised for making the sample as representative as possible. In this connection, one may emphasize the need of obtaining the distribution from earlier serveys of the distribution (*s*) for the character (*s*) to be studied in subsquent occasions. But formulation of a priori distribution may not be fruitful in all the cases. Perhaps the beginning can be made with the set-up in which repeated (successive) surveys are conducted over time to study the changes in the value of characters where distributions of observations in earlier surveys can be utilised to obtain a priori distribution (*s*) for the character (*s*) to be studied in subsequent occasions. Perhaps such studies could lead to an acceptable a priori distribution as the a priori distributions are no more hypothetical and subjective but based on the past experience of the results of the previous surveys. This will make the Bayesian estimates more acceptable to users

Purposive *vs.* Random Samples:

Everyone agrees that the samples should be 'representative' of the population for making inferences about the parameters of the population. The concept of 'representativeness' has, however, not been clearly defined. Ideally a representative sample should be such that when its size is expanded by the factor $\frac{N}{n}$, (n is the sample size fIom a population of size N) it should reproduce the population and this should apply not only for one or a few characters but for all the different characters under study of a population. As we do not have sufficient knowledge about the population, for that is why we are studying it, this ideal representive sample is not possible to construct. The next best thing is that a frequency distribution of 'representative' sample should be very close (in a sense to be clearly defined) to the corresponding distribution in the population for each character. Such a thing is also not usually possible to achieve and it is difficult as well to verify its representativeness for all characters.

Neyman (1934) gives a similar concept of representativeness when he says that a procedure is representative when the samples obtained by grouping units presumably give the same characteristics as the whole of the population. Usually it is some sort of quota sampling for different categories of some characters closely associated with the characters under study, the frequency distribution of these characters being roughly known and quotas (numbers) are then fixed for each combination of the various characters in proportion to the number of units for those characters in the different categories. After fixing quotas, it is left to the enumerator to select the number of units equal to the quota fixed from the total units in the category, the choice of actual units being left to enumerators. Such samples are thus representative of the population in respect of the characters for which quota are fixed and perhaps also for other characters which are highly positively correlated with some quota-fixed-character but may not be representative in respect of other characters. All large scale sample surveys cover multiple characters and thus these samples may not be representative in respect of a number of characters under study thereby making the estimates obtained highly biased. At present, there is also no mechanism to study the extent of biases produced by purposive samples. Further, the estimates cannot be utilised for making statistical inferences with prescribed margins of error. Thus Neyman (1934) defines representativeness also in another sense *i.e.* samples which can provide margins of errors for estimates based on the observation made in the sample will also be called representative. Random samples have been shown by Neyman (1934) to be such representative samples. A random sample is not 'representative' in the sense a 'purposive sample' is supposed to be, for it can consist of only extremely low or high values, but it sets up a mechanism for estimating margin of errors to be associated with the estimators (or one can make confidence interval statements on the basis of a probability sample). These are the two aspects of representative method—purposive and stratified random sample studied by Neyman in his (1934) paper. On the basis of the data of a complete human population census of Italy in 1921, Prof. Neyman showed that purposive samples provide estimates which are not consistent with the population values for a large number of characters as observed in the census whereas the stratified random sample provided satisfactory confidence intervals for the population parameters, thereby establishing the superiority of random samples over the purposive samples. Further representative samples of a large size are difficult to obtain by 'purposive sampling' especially in the case of a study involving multiple characters. Even when great care is taken to select purposive samples, the various users will still look upon it with doubt saying that subjective factors in selection have made this non-representative. This will specially be the case when the choice is left to a large number of field enumerators who cannot look on 'representativeness in the same sense and thereby introducing their own selection biases. In fact, this has been found true in practice in many cases. The more leeway is given to enumerators in the choice of the sampling units and their measurements, more biases and measurement errors occur. As it is not always possible to select ultimate units (for lack of updated frames) at headquarters of a sample survey organization, it will be a good strategy to ask enumerators select the units according to same prob. system from the available (or to be formed) frame of the ultimate units for their area of enumeration, the number of such samples to be fixed in advance by the headquarters. Random sampling safeguards against conscious or unconscious self-deception and the results of a survey are readily acceptable and convincing to users as these are free from personal biases. Randomization also increases the utility of the results as the data can be used to study the relative efficiencies of alternative probability models. In large scale surveys where biases (selection) play a bigger role than random errors, random samples can be safely used to produce data free from personal biases. Randomization will also enable the sampler to specify beforehand data analysis procedures which have definitely known communicable probabilistic properties e.g. unbiased. The estimates based on a random sample retain their nice properties (at least approximately) under a variety of circumstances in which the assumed model may not be strictly proper *i.e.* it is an insurance against model defects. In certain cases 'randomization' also validates an 'assumed' model. For example, 'exchangeability' is assumed by 'Ericson' in the development of Bayesian procedures in sample surveys. 'Exchangeability' and 'randomness' are equivalent in a certain sense and interchangeable. One should verify the 'exchangeability' before using it and this is very difficult to verify. But if one adopts 'randomness', 'exchangeability' is automatically assured. It is better to have an 'assured' exchangeability rather than 'assumed' exchangeability with the help of random sampling. Randomization is redundant in Bayesian set-up as the repeated samples (possible) are not utilised for judging the credibility of the Bayesian estimates, for the observed sample and the a prior distribution determines the posterior distribution which is the basis of inference. But, as mentioned above, it can help in the validation of certain assumptions made in the Bayesian methods.

To summarise, randomization has worked in practice even though one may not be able to fully explain why it has worked. Even Prof. Savage, leading Bayesian Concedes that randomness has some part to play in statistical methodology especially in large experiments and large Scale Sample Surveys. Infact, a well-designed probability sample will always be better than a good judgement sample. It may increase the accuracy of the estimates by removing unsatisfactory biases even though estimates will have a larger variance. This is of special importance in large scale surveys where biases are more important than random errors. Randomization increases the range of utility of the data as it can be comfortably used by another worker and facilitates the testing of the various alternative probability models. But randomization has not to be used indiscriminately as a cure for all ills. How much randomization is necessary to achieve its different purposes (like validity of results and obtaining margins of errors) is a question to which a lot of attention is to be paid by the sampler. Its abuse is to be avoided. It is thus necessary to set up some standards for use of restricted random sampling.

Future of Survey Sampling

Forecasting is a hazardous affair in any field. Users as well as producers have a say in any future development. Theorists will continue to develop more efficient methods for the sample survey theory. One can only wish more collaboration between survey-statisticians and theorists to develop immediately usable methods and super population models. This interaction will indeed lead to more rapid and useful developments. One such field is the field of non-sampling errors. Now it is well recognised that non-sampling errors are indeed more important than even sampling errors as these can make the whole data non-sensical. Quality of data is really to be looked upon as statistical quality control over the statistical production of data. As quality of data can suffer through bad selection, non-response, measurement errors and processing errors, each stage of sample survey adds an error of its own type different in magnitude and direction. Care is thus to be taken at all the stages to control the quality of data. Here there is a real role of 'infinite' population models and simulation studies need to be undertaken to understand the mode of operation of errors in sample surveys of different types and for devising means for their control. Also methodologies need to be developed for analysis of survey data which is subject to large measurement errors. In another field where theorists can help is the estimates of frequencies for such sub-*domains* for which size of the samples becomes random. As the sample sizes are fixed generally for domains of study and not for all such domains, random number of units fall in a given sub-domain. Also sometimes sample units are missed in enumeration. When size of the sample becomes random, should the estimates be based on a conditional inference taken as the numbers observed in the sample as given or these should be averaged over all possible variations in the random size? Sometimes one method is better than the other and *vice versa.* Conditions under which one is to be preferred to the other need further research. No doubt samplers are using a priori information available about the populations under study to a large extent, but there is still scope for its increased use for further restricting random sampling. For example use of a priori distributions will be realistic based on past surveys will be realistic. Thusin such cases one should explore the possibility of setting up meaningful prior distributions, for that will extract the maximum a prior information for use of stastical inference. Successive sampling has made good progress but research should be undertaken how to make increased use of information in earlier surveys to improve estimates of future surveys. More analytic studies should be undertaken to study the structure of the population especially the inter-relation of various characteristics. For application of asymptotic results to sample survey, one should really fix the minimum sample size for various type of populations. As the need of storing information for future use is realised more and more sampling designs should be developed to store a part of the data which can reproduce the essential features of the original data. In surveys based on varying probability designs, frequencies of various characters of the population are represented by the frequencies in the observed sample. These estimates of frequencies are highly biased. A proper methodology needs to be investigated for this purpose. One should increasingly also study the role of 'controlled' selection and the 'balanced sample' techniques, both initiated by Kish. Development of optimum design for multivariate characters has not made much hardway and work in this needs considerable strengthening. There is a large scope of more work in capture release and recapture sampling in fields like estimating fish catch. One can mention many more topics like those suggested in the symposia held on 'recent developments in survey sampling' held at Chapel Hill in 1968, to which proper attention has not yet been paid.

In short, efficient restricted sampling techniques should continue to be developed while controlling the influence of selection and measurement biases. Wherever "model" or "Bayesian" approach can be useful, one should not hesiate to use them. There is a need of greater integration of sample survey theory with other areas of statistical inference. Some of the confusion regarding the foundations of sample survey theory can be removed by establishing a better dialogue between the theorist and practitioner and I end my talk with the hope that it will be achieved in the near future.

APPENDIX I

TABLE 3.1

Various T-classes of linear estimators defined by Horvitz and Thompson, Godambe, Koop and by present authors incorporating the earlier work of Tikkiwal and Prabhu Ajgaonkar

Authors O	T_1	T_2	T_3	T_4	T_5	T_6	T_7	
Horvitz & Thompson	$\sum_{r=1}^{n} \beta_r X_r$	$\sum_{i \epsilon s_1} \beta_i x_i$	$\beta^{s_1}\left(\sum_{i \epsilon s_1} x_i\right)$	—	—	—	—	SWOR
Godambe	$\sum_{r=1}^{n} \beta_r X_r$	$\sum_{i \epsilon s_1} \beta_i x_i$	$\beta^{s_1}\left(\sum_{i \epsilon s_1} x_i\right)$	$\hat{T}_j = \sum_{i \epsilon s_1} \beta_{(s_1, i)} x_i$	—	—	—	SWOR
	$\sum_{r=1}^{n} \beta_r x_r$	$\sum_{i \epsilon s_1} \beta_i x_i$	$\beta^{s_1}\left(\sum_{i \epsilon s_1} x_i\right)$	$\hat{T}_j = \sum_{i \epsilon s_1} \beta_{(s_1, i)} x_i$	—	—	—	SWR
Koop	$\sum_{r=1}^{n} \beta_r x_r$	$\sum_{i \epsilon s_1} \beta_i x_i$	$\beta^{s'_k}\left(\sum_{i \epsilon s'_k} x_i\right)$	$\sum_{r=1}^{n} \beta_{ro} x_r$	$\sum_{i \epsilon s'_k} \cdot \beta_i^{s'_k} x_i$	$\beta^{s_1}\left(\sum_{r=1}^{n} x_r\right)$	$\sum_{r=1}^{n} \beta_{ro}^{s_1} x_r$	SWOR

Present authors	$\sum_{r=1}^{n} \beta_r x_r$	$\sum_{i \varepsilon s_1} \beta_i x_i$	$\beta^{s'_k}\left(\sum_{i \varepsilon s'_k} x_i\right)$	$\sum_{r=1}^{n} \beta_{ro} x_r$	$\sum_{i \varepsilon s'_k} \beta_i^{s'_k} x_i$	$\beta^{s_1}\left(\sum_{r=1}^{n} x_r\right)$	$\sum_{r=1}^{n} \beta_{ro}^{s_1} x_r$	SWR
	$\sum_{r=1}^{n} \beta_r x_r$	$\sum_{i \varepsilon s_1} \beta_i x_i$	$\beta^{s_1}\left(\sum_{i \varepsilon s_1} x_i\right)$	$\sum_{r=1}^{n} \beta_{ro} x_r$	$**\sum_{r=1}^{n} \beta_i^{S}(x_r = x_i)$	$**\sum_{r=1}^{n} \beta_r^{S} x_r$	$\sum_{r=1}^{n} \beta_{ro}^{s_1} x_r$	SWOR
	$\sum_{r=1}^{n} \beta_r x_r$	$**\sum_{r=1}^{n} \beta_i \ (x_i = x_r)$	$\beta^{s_1}\sum_{i=1}^{n} (x_r = x_i)$	$\sum_{r=1}^{n} \beta_{ro} x_r$ $\beta_{ro} = \beta$ if $x_r = x_{i_r}$	$**\sum_{r=1}^{n} \beta_i^{s}(x_r = x_i)$	$\bullet\bullet\sum_{r=1}^{n} \beta_r^{s} x_r$	$\sum_{r=1}^{n} \beta_{ro}^{s_1} x_r$	SWR

various notations used in the the table :

In SWOR : $s_1 = 1, 2, 3 \ldots\ldots \binom{N}{n} n!$; $S = 1, 2, 3 \ldots\ldots \binom{N}{n}$

In SWR : $s_1 = 1, 2 \ldots\ldots\ldots N^n$; $S = 1, 2, 3 \ldots\ldots \sum_{k=1}^{n} \binom{n}{k} M_k$

In SWOR : $s_k = 1, 2, 3 \ldots\ldots \binom{N}{n}$ $\quad \sum_{i \varepsilon s_1}$ denotes the summation over distinct units in the s_1th sample.

In SWR : $s_k = 1, 2, 3 \ldots\ldots \sum_{K=1}^{n} \binom{N}{k}$ $\quad \sum_{i \varepsilon s'_k}$ denotes the summation over distinct units in the s'_kth sample.

** denotes new classes of estimators given by present authors incorporating the earlier work of Tikkiwal and Prabhu Ajgaonkar
Source :- Bhargava, N.K. and Tikkiwal, B.D. (1978), Sankhya Series C Vol. 40, Part I, P. 18.

APPENDIX II

Let θ^* be any estimate of θ based on the observed sample $(i_1, i_2, \ldots\ldots, i_n)$ of fixed size n. Let $P(s)$ denote the probability of drawing the samples given by $i_1, i_2, \ldots\ldots i_n$. Then $\phi^* = E[\theta^* \mid S]$ has the same expectation as θ^* and has a lower variance than θ^*, where S is the set of all permutations in the sample s observed, E being the symbol for expectation.

$$E(\theta^*) = \underset{S}{E}\,[E(\theta^* \mid S)] = \underset{S}{E}\,[\phi^*] = E(\phi^*)$$

ϕ^* depends only on the observations and thus is itself an estimate

$$V(\theta^*) = \underset{S}{E}\,[V(\theta^* \mid S)] + \underset{S}{V}\,[E(\theta^* \mid S)]$$

$$= \underset{S}{E}\,[V(\theta^* \mid S)] + \underset{S}{V}\,(\phi^*)]$$

As first term is positive or zero

$V(\phi^*) \leqslant V(\theta^*)$

Further $\psi^* = E(\phi^* / \mathscr{E})$ has the same expection, where e is the set of all k-partition of n, as ϕ^* but an equal or a lower variance than that of ϕ^*.

$$E(\phi^*) = \underset{\mathscr{E}}{E}\,[E(\phi^* / \mathscr{E})] = \underset{\mathscr{E}}{E}\,[\psi^*].$$

Thus ψ^* has the same expectation as ϕ^*.

$$V\,(\phi^*) = \underset{\mathscr{E}}{E}\,[V(\phi^* \mid \mathscr{E})] + \underset{\mathscr{E}}{V}\,[E(\phi^* \mid \mathscr{E})]$$

2nd term in the above is the variance of ψ^* whereas the first term is non-negative. Thus $V(\psi^*) \leqslant V(\phi^*)$