Statistics and Applications **{ISSN 2452-7395(online)}**
Volume 14, Nos. 1&2, 2016 (New Series), pp 131-143

Historical Perspective and Some Recent Trends in Sample Survey Applications

A.K. Srivastava
ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi

Received September 27, 2016; Revised September 29, 2016; Accepted October 05, 2016

Prologue

I am thankful to the Society of Statistics, Computer and Applications for providing me an opportunity to remember and pay my tributes to our respected teacher Dr. M N Das. Das Sahib (as we students used to call him fondly), was a revered teacher, guide and mentor to a vast community of students. Generations of his students' fraternity have earned respect and recognition both in academic fields and responsible positions in official statistics system in India and also internationally. His simplicity and an intuitive approach to every problem he faced, whether in research, official responsibilities or in his routine life made him a distinguished personality amongst his peers and contemporaries. The Society of Statistics, Computers and Applications was not only established by him but was practically sailed all along under his guidance. The present lecture is a humble effort on my part to offer my respect to "Das Sahib".

Abstract

A brief overview of developments in sample survey theory and its applications is presented in its historical perspective. The establishment of random sampling as a dependable tool for data collection and analysis; the discussions relating to foundational and inferential aspects of survey sampling; development of analytical tools for survey data analysis, adoption of improvements in information and communication technology for data collection and analysis are some of the themes discussed in the paper.

Key words: Random sampling; sampling from finite populations; Varying Probability Sampling; inferential aspects in survey sampling; survey data analysis from complex surveys; small area estimation

1. Introduction

The present talk is based on some of my own experiences and observations in the field of sample surveys during the last five decades. Purpose is not to provide an exhaustive review of literature, but to share my observations in the development of sample survey theory and practice in a historical perspective. Some experiences related to agricultural surveys, particularly in the Indian context are also touched upon.

2. Early Developments

Although efforts for the use of samples in data collection were initiated by the end of 19th century (Kiaer, 1895-96, 97, 1903), the real watershed in the development of sampling theory

Corresponding Author: A.K. Srivastava
E-mail: arunsrivast@gmail.com

and practice was observed with the emergence of a paper by Neyman (1934). To start with, the very idea of making inference about a population on the basis of observing only a part of it, generated a genuine skepticism. The initial efforts were focused on the considerations and criteria to be used for the selection of a good sample. The four important principles involved in Kiaer's approach were:

1) Representativeness
2) Lack of subjectivity
3) Reliability of the results should be assessed
4) Complete specification of the method of selection be included with the results of any sample survey

It may well be realized that how important these considerations have been in shaping the future of sampling theory and its applications. Representativeness was conceptualized as adequate representation of identifiable groups in the population, identifiable in principle and for which population figures were known. It was somewhat akin to "balanced sampling" in which sample means are close to population mean for identifiable known characteristic. Kiaer's method of selection for ensuring representativeness was close to proportional stratified multistage sampling (of course, without random selection). Bowley (1906) provided a theory of inference for survey samples and using Edgeworth's Bayesian version of the Central Limit Theorem, he was able to assess the accuracy of estimators made from large samples drawn by simple random sampling from large finite populations. His theoretical analysis showed that very often quite small samples are good enough and census was not always necessary. Bowley's method was, however, limited to simple random sampling only.

In the field of Agricultural Statistics, an interesting development had taken place in India in early 1920s. J, A. Hubback had conducted crop cutting experiments on paddy crops in Santhal Pargana of Bihar State, where 400 samples were harvested on an area of 100 square miles, followed by its extension to Santhal Paraganas District and also in the state of Orissa. Subsequently, when random sampling methods were established on sound footing, the crop-cutting approach was developed by P. C. Mahalanobis (1938, 1940) as a method for estimating crop acreage and crop yield of Jute crop in Bengal. For Food crops, the crop estimation methodology based on crop – cutting approach was developed in 1940s and 50s in Indian Council of Agricultural Research under the leadership of P. V. Sukhatme and V. G. Panse (1948, 51).

3. Neyman (1934) and Subsequent Developments

Representativeness of samples and objectivity in the selection process were the major concern. The two methods of random sampling and purposive selection represent logical developments of the methodology presented by Kiaer. It was in the same pursuit that Neyman (1934) came up with his landmark paper "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection". In this paper, Neyman developed a theory of inference, based on confidence intervals, which is suitable for use with finite populations of a kind which are commonly encountered in survey situations. He demonstrated that it was possible, using the idea of confidence intervals to define a representative method of sampling and a consistent method of estimation.

The striking feature of Neyman's approach was that the randomness introduced through the method of selection was the basis of inference for sampling from finite population. The probability structure associated with the sampling design provided the framework for confidence intervals, which became the basis of measuring reliability of estimators. There was no need to make assumptions about the random nature of the population distributions.

However, the confidence statement was a frequency statement about a mode of behavior based on all possible random samples that could be drawn. Neyman introduced the concept of efficiency through the shortest confidence interval in a class of intervals. Using this approach of shortest confidence interval and minimum variance linear unbiased estimators, he deduced that stratified sampling would usually be more efficient than simple random sampling and in the process introduced optimum allocation (known as Neyman's optimum allocation).

Neyman's arguments not only provided a sound inferential base for random sampling, but it established the probability sampling approach as a viable and preferred alternative for survey sampling practice in comparison to purposive sampling. It also liberated sampling from its equal probability straight jacket. Some of the standard sampling tools and techniques, which are not based on equal probability rationale, are a natural outcome of Neyman's approach. In optimum allocation for stratified sampling, the idea of unequal selection probabilities for units from different (unequal) strata, was implicitly inbuilt. Similarly, for cluster sampling with unequal cluster sizes, the idea of unequal probability was a natural consequence.

The next fifteen to twenty years witnessed a very rapid all round growth based on random sampling approach. Various probability sampling methods were developed and refined during this period. A theory for systematic sampling was developed (Madow and Madow (1944)). Probability proportional to size (PPS) sampling was introduced (Hansen and Hurwitz (1943)). It may be remarked that although PPS was introduced by Hansen and Hurwitz in 1943, Mahalanobis (1938) was aware of the PPS selection in 1937 and it was used in agricultural surveys in the form of grid sampling. A comprehensive theory for Ratio and regression methods, which were used earlier as traditional methods, was provided by Cochran (1942) in the context of unequal clusters. The concept of double sampling or two phase sampling was introduced by Neyman (1938). Sampling for repeat surveys or sampling over successive occasions was introduced by Jessen (1942) and further developed by Patterson (1950). Sub sampling or multistage sampling, which is a natural outcome of cluster sampling (sub-units are selected within each cluster as a second stage of selection), was also developed.

An interesting feature of the developments during this period was that methods developed were simultaneously tested through actual surveys. In fact, the need for various methods which were developed, was emerging from practical considerations only. In India at Indian Statistical Institute, Calcutta, various methods like Interpenetrating Network Sampling (IPNS) and cost functions etc. were developed. The Jute Survey conducted in Bengal by Mahalanobis (1937, 38) is an elegant example for conducting pilot sample surveys. The methodology for crop estimation surveys through crop cutting techniques was developed at Indian Council of Agricultural Research. By 1950, quite substantial developments had taken place, which are consolidated in the form of various text books (Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurwitz and Madow (1953) and Sukhatme (1954), Dalenius (1957). These books were written by persons, who were not only involved in the development of sampling theory but were instrumental in establishing the practice of survey sampling as a powerful tool for development of statistical systems in their respective countries. Most of these books had subsequent editions and a follow up of some valuable text books such as by Kish (1965), Murthy (1967), Des Raj (1968) etc.

Although a comprehensive review of sampling literature is not attempted here, a mention of some review papers at different points of time should be worthwhile to put the developments in proper perspective. Some of such review papers are by Sukhatme, P. V. (1959), Dalenius (1962), Murthy, M. N. (1963), Smith, T. M. F. (1976), O'Muircheartaigh

and Wong (1981), Ghosh, J. K., Maiti, P., Rao, T. J. anad Sinha, B. K. (1999), Rao, J. N. K. and Bellhouse, D. R. (1990), Rao, J. N. K. (1990).

4. Varying Probability Sampling

Although sampling with unequal probabilities was conceptualized in 1930s (Mahalanobis (1938)) and formalized in 1940s (Hansen and Hurwitz (1943), the real impetus to the concept was provided through the papers by Narain, R. D. (1951) and Horvitz and Thompson (1952). These papers led to a series of studies on various methods of varying probability sampling without replacement. But an important impact was on future direction of research in the sampling theory. Horvitz and Thompson considered three classes of linear estimators termed as T_1, T_2 and T_3. The Horvitz and Thompson estimator (HTE) was a central theme in many of the subsequent developments in sampling literature. However, calculation of inclusion probabilities for varying probability sampling without replacement was, quite often, not so simple and was considered a limitation in its application. In the mid-fifties and early 60s a number of sampling schemes and estimators were proposed which either avoided the calculation of inclusion probabilities or simplified its calculation. Some of the important contributions in this direction are Rao- Hartley and Cochran sampling procedure (1962), Des Raj ordered estimators (1954), M. N. Murthy's un-ordered estimators (1957) etc. These estimators had an appeal of simplicity as they did not require calculation of inclusion probabilities.

The usual HT estimator of population total Y is

$$\hat{Y} = \sum_{i \epsilon s} y_i / \pi_i$$

Where y_i and π_i are study character value and the inclusion probabilities for ith unit belonging to the sample s. If inclusion probabilities are proportional to the size measure of selection, then in varying probability without replacement sampling schemes, $\pi_i = np_i$, where n is the sample size and p_i are the initial probabilities, the HT estimator reduces to

$$\hat{Y} = \frac{1}{n} \sum_{i \epsilon s} y_i / p_i.$$

This simplicity of the HTE initiated a quest for sampling schemes with inclusion probabilities proportional to size, which are termed as IPPS schemes or πps schemes. Some of the important IPPS schemes are Probability Proportional to Size Systematic (PPS systematic) sampling Goodman and Kish (1951), Hartley and Rao (1962), Sampford (1967) etc. A comprehensive account of developments in varying probability sampling is available in a Springer-Verlag publication by K. R. W. Brewer and Mohammed Hanif (1982).

5. Inferential Aspects in Survey Sampling

Another direction in which the research in sampling theory progressed was attempts to bring estimation in sampling theory closer to the estimation problem in usual statistical inference. Concepts like sufficiency, likelihood, admissibility etc. were well established in usual statistical inference. However, they were not finding a place in the usual sampling literature. In the process of looking for a unified theory of sampling from finite populations, Godambe (1955) provided a formal framework for finite population inference as follows:

Consider a finite population each of whose elements are labelled, with label set U= {1, 2,…., N}. With each $i \in U$ is associated a value Y_i. The population parameter is the vector Y= (Y_1, Y_2, …..,Y_N). Y is a member of the parametric space $\mathfrak{Y}$, which contains all possible populations and $\mathfrak{Y} \in R_N$. Prior information exists on which survey design can be based. A sample, s, is a

subset of U, usually of n units, so that s= {i_1, i_2, ...,i_n}. A sampling scheme assigns a probability p_s to each s. S is a set of all possible samples and $\sum_S p_s = 1$. *The data comprise the labels, s selected by the scheme p_s and the associated values Y_i . Thus data= (s, Y_i; $i \in s$). An estimator e_s is a function of the data and any relevant prior information.*

Within this framework, Godambe proved nonexistence of uniformly minimum variance (UMV) estimator. This was a negative but very powerful result in the quest of best estimators. Incidentally, a more elegant proof of the non-existence theorem was provided by Basu, D. (1971). Non- existence of the UMV estimators led to choice of best estimators in restricted classes. In the absence of UMV estimators concepts like necessary best estimators were also considered. In the process of searching for good estimators in a particular class, various concepts like admissibility, hyper admissibility etc. were considered. Quite a good number of papers appeared in 1960s on these aspects. However, most of these concepts had little success and Basu (1971), observed that these concepts are totally unrealistic and that no real progress is possible along these lines.

The concept of sufficiency was also tried in sample survey situation. Basu (1969) showed that in sample survey situation, sufficient statistics is the set of labels, s and the set of values Y_i, corresponding to *distinct labels.* An application of Rao- Blackwell theorem provides the result that sample mean based on distinct units is a better estimator for the population mean than the mean based on all the sample units (repetition included). The result was known earlier, but sufficiency concept provided an alternative and more elegant proof. Similarly, the result of Murthy (1957) that the unordered estimators in case of unequal probability selection were better than the ordered Des Raj estimator (based on the ordering in which units are selected) was proved using the result that the set of observations in the sample after ignoring the ordering of selection is a sufficient statistics.

It was observed that the likelihood function in the sample survey frame-work is flat and, therefore, well-known Maximum Likelihood Estimation (MLE) is hardly of any use in sample survey situation. One of the important views which was emerging was that sampling design should not have a role to play in the estimation of population parameters. This was, indeed a deviation from traditional approach propounded by Neyman.

Another important development during 1960s and 70s was the use of models in survey sampling for estimation purposes. Models were being used in survey sampling during 1940s and 50s also, but the application was mostly limited to comparison of variances for different sampling designs. Cochran (1946) used a model for auto-correlated populations, for comparing relative accuracies of systematic and stratified random samples. During 1950s and 60s models were frequently used for comparing varying probability estimators with those for ratio estimators. Comparison among the methods were also made from an infinite population model in which finite population was considered as a random sample from an infinite super-population. A commonly used model was

$$y_i = \alpha + \beta x_i + e_i; \qquad E(e_i|x_i) = 0$$
$$V(e_i) = V(y_i|x_i) = cx_i^g, \quad \text{where } 1<g<2$$

This model was commonly known as g-model. Comparison of various estimators for equal and unequal probabilities are available in most of the text books, e.g. Cochran (1977).

An alternative approach for estimation of population totals was made by Brewer (1963) and Royall (1968, 70, 71) in the form of predictive approach, which made explicit use of models for estimation purposes. The population total Y may be written as

$$Y = \sum_{s} y_i + \sum_{\bar{s}} y_i$$

Where *s* is the sample drawn and $\bar{s}$ consists of population units not belonging to the sample. For the sample s y – values are observed and therefore for estimation of Y, the remaining un-observed units in $\bar{s}$ need to be predicted. For this purpose, certain models have been used whose parameters are estimated through the observed samples. A model based predictor of Y may be written as

$$\tilde{Y} = \sum_{s} y_i + \sum_{\bar{s}} \tilde{y}_i$$

$\tilde{y}_i$ is the predicted value of y_i on the basis of a specified model. Complete description of such models is available in literature such as Cassel, Sarndal and Wretman (1977). In this approach, the sampling design has practically no influence on the way the predictors are developed. Therefore, several well established and interesting features of the sampling designs, such as unequal probability selection was losing its ground. Baysian approach in survey sampling was also leading towards similar conclusions. In fact, Basu (1971) observed that *"it is not easy to understand how surveyors got messed up with the idea of unequal probability sampling"*. This was the period when the theoretical base of survey sampling based on Neyman's sampling design based approach was being questioned. However, there were apprehensions about performance of model based predictive approach was under model failures. Attempts were made to study the robustness of such predictors under certain type of model failures (Royall and Herson (1973a, 1973b). It was also observed that randomness does provide certain safeguards against model limitations. This led to sampling investigations related to model assisted survey inference which was a mix of random sampling approach assisted with inference from model based arguments. The books by Cassel, Sarndal and Wretman (1977) and Sarndal, Swensson and Wretman (1992) provide a comprehensive account of model based and model assisted inferential aspects of survey sampling.

It is observed that despite all the doubts raised regarding the Neyman's design based approach of sampling during 1960s and 70s, it is being practiced in almost all large scale sample surveys conducted by different countries. Most of the data needs under various sectors of economy are being met through sample surveys conducted with the traditional sample survey approach. However, it goes to the credit of galaxy of survey statisticians during this period, who got into a healthy debate regarding foundational aspects of survey sampling. These discussions not only clarified various doubts and queries raised from the early days of sample survey applications, but also paved the way for future developments in analysis of survey data.

6. Analysis of Survey Data

During 1980s and onwards, the technological developments in computing capabilities, ushered in a new scope for data analysis. Most of the literature on sampling theory deals with estimation of population parameters such as population means, totals and ratios along with their standard errors. In the context of complex survey data analysis, attention has been paid to complex descriptive parameters such as domain (subpopulation) totals and means, quantiles, regression and correlation coefficients etc. I would only like to refer to some of the trends in the analysis of survey data during the last two to three decades.

Estimation of domain parameters, particularly for smaller domains has gained importance in view of growing needs for micro level planning. The main problem in the estimation of domain parameters is that sample sizes in sub-populations are too small to provide reliable estimates with the help of direct estimators. Hence it becomes essential to borrow information from related or similar areas through implicit and explicit models to produce indirect estimators which increase the effective sample sizes. We shall come back to some methods of small area estimation in a subsequent section.

One of the assumption in classical statistical data analysis i.e. units of the sample are independently and identically distributed (i.i.d.), no longer holds in case of complex survey data as these are collected through complex sampling designs. For these purposes, even sampling designs like stratified and cluster sampling are treated as complex ones. Regression coefficients are estimated with the help of multiple regression technique, which assumes independence of observations. Consequently, use of standard Ordinary Least Square (OLS) technique to survey data for estimating regression coefficients provides misleading results due to sample units being not independently drawn. Efforts have been made to incorporate the effect of sampling designs for appropriate estimation procedures.

In the context of survey data analysis, interest has also developed in analytical surveys i.e. investigating relationships among variables in the survey. The analysis of categorical data from survey data is an example of analytical data analysis. Here also i.i.d. assumption is violated and standard statistics such as chi-square tests need adjustments to ensure valid conclusions.

Area of variance estimation has gained importance in view of complex survey designs. The traditional approach has been to derive a formula for sampling variance and its estimator for each estimator of interest. This becomes quite cumbersome in case of complex data analysis. In case of non-linear estimators, it may not be even sometimes possible to work out such a formula. The earliest example of a simple approach for estimation of variance in a complex situation is available in the form of interpenetrating net-work of sub-sampling (IPNS) by Mahalanobis (1937). Although this approach was developed in the context of non-sampling error for estimating enumerators' effect, but it has been used quite effectively for variance estimation. Some of the commonly used variance estimation methods are Taylor Series Linearization, Jack=knife method, Balanced Repeated Replication method and Boot-strap methods. It may be noted that IPNS has been a pre-cursor to several Resampling methods for variance estimation. Some of these techniques, such as Boot-strap method are highly computer intensive. Also, every survey data analysis computer packages has got one of these variance estimation techniques as an essential component. Along with the sample selection methods and analytical tools, there has been considerable improvement in data collection and data processing tools and techniques. The traditional approach for data collection has been through survey schedules/questionnaires canvassed through personal interviewing. With the improvements in communication technology and computational facilities, there have been efforts to adopt improved methods of data collection. Methods like Computer Assisted Telephone Interviewing (CATI) have been in practice in developed countries for more than three decades. It helps in reducing both measurement and data processing errors. During this period, the outreach of communication as well as computing facilities have grown phenomenally in developing countries as well. An improved version of Computer Assisted Personal Interviewing (CAPI) is now being used in many of the developing countries. Hopefully, it will have an enduring effect on cost and quality of data collection. Presently, mobile phones technology has surpassed the traditional telephone lines in its reach even in rural areas. With smart phones getting more and more popularity and

spread, it has got great potential of getting ground toward improvements in data collection methodology.

7. Small Area Estimation

Surveys are normally planned with specific populations in view. However, quite often interest also lies in parts of the population known as 'sub populations' or domains of interest. Domain parameters may be estimated satisfactorily within the framework of sampling theory provided the domains get sufficient representation of sampled units to provide reasonably reliable estimates. However, the problem arises when the 'sub populations' or 'domains' are so small that samples in these domains are too small to provide any reliable estimates.

Most of the small area techniques in the early stages were developed in the context of demographic studies. Most common of the various techniques was "Synthetic Estimation method", which had several variants. All the techniques utilized related information from various sources of data such as census, registers and samples. Also all the methods were indirect methods of estimation and were based on certain intuitive assumptions, in which models are in the implicit form. In recent years, there have been lots of developments based on models, which are termed as model-based small area estimation techniques.

We now consider some explicit model-based methods which are essentially mixed models and are used in specific situations based on data availability on the response variables of interest. These are

A. area level models where information on response variable is available only at the small area level; and ii) unit level models where information on the response variable is available at the unit level.
B. These models are described as follows.

Area Level Models

An area level model has two components:

Direct survey estimate of the parameter θ (which is a function of finite population mean), based on the sampling design, expressed as follows:

$$\hat{\theta}_q = \theta_q + e_q, \quad q=1,\ldots,Q$$

where the e_q's are assumed to be independent across small areas with mean zero and known variances χ_q. This model is a sampling model and χ_q is a design- based sampling variance. A linking model

$$\theta_q = X_q^T \beta + \nu_q \quad q=1,\ldots, Q$$

where the model errors ν_q are assumed to be independent and identically distributed with mean zero and variance σ_q^2. The model variance σ_q^2 is a measure of homogeneity of the areas after accounting for the covariates X_q. Combining these two models, the resultant mixed linear model is

$$\hat{\theta}_q = X_q^T \beta + \nu_q + e_q, \quad q=1,\ldots,Q$$

Using the data $\{(\hat{\theta}_q, X_q), \quad q=1,\ldots,Q\}$, we can obtain estimates θ_q^*, of the realized values θ_q from the model (3.3). Here e_q's and ν_q's are design-based and model-based random variables respectively.

Empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods have played an important role in the estimation of small area means $\overline{Y_i}$ under model as considered as above. EBLUP method has been used in many practical applications. One of the early applications of this method was due to Fay and Herriot (1979). In fact, this method was adopted by the U. S. Bureau of Census in 1974 to form Per Capita Income (PCI) estimates for small places. EBLUP method is applicable for mixed linear models and they do not require normality assumption of the random errors ν_q and e_q.

The other methods EB and HB are applicable under specific distributional assumptions. The inferences in HB methods are obtained through posterior distributions. EBLUP and EB are identical under normality assumptions. For EBLUP and EB, an estimate of MSE $(\tilde{\theta}_q)$ =E $(\tilde{\theta}_q - \theta_q)^2$ is used as a measure of variability of $\tilde{\theta}_q$, where the expectation is with respect to the model.

EBLUP estimate of θ_q is a composite estimate of the form

$$\theta_q^* = \hat{\gamma}_q \hat{\theta}_q + (1-\hat{\gamma}_q) X_q^T \hat{\beta} ,$$

where $\hat{\gamma}_q = \hat{\sigma}_\nu^2 \big/ (\hat{\sigma}_\nu^2 + \psi_q)$ and $\hat{\beta}$ is the weighted least square estimate of β with weights $(\hat{\sigma}_\nu^2 + \psi_q)^{-1}$ obtained by regressing θ_q on X_q: $\hat{\beta} = (\sum_q \hat{\gamma}_q X_q X_q^T)^{-1} (\sum_q \hat{\gamma}_q X_q \theta_q)$ and $\hat{\sigma}_\nu^2$ is an estimate of the variance component σ_ν^2. It may be noted that θ_q^* is a linear combination of direct estimate $\hat{\theta}_q$ and the model based regression synthetic estimate $x_q^T \hat{\beta}$, with weights inversely proportional to their respective variances. For the non-sampled areas the EBLUP estimate is given by the regression synthetic estimate itself.

Under the model, the leading term of MSE $(\tilde{\theta}_q)$ is given by $\gamma_q \psi_q$ which shows that the EBLUP estimate can lead to large gains in efficiency over the direct estimate with variance ψ_q, when γ_q is small i.e. the model variance σ_ν^2 is small relative to the sampling variance ψ_q. Choice of good auxiliary data to provide a good model fit is, therefore the key to successful application of the small area technique.

An excellent example of application of this method is in a study on *Small Area Estimates of School-Age Children in Poverty* (Constance, F. Citro and Graham Kalton, Editors (2000)). Some recent applications of this method have been towards estimation of poverty ratios at district level in Uttar Pradesh (UP) utilizing data from Consumption Expenditure Surveys of NSSO. An application of this method was also made in crop surveys in Ethiopia for obtaining Woreda level estimates of crop areas.

Unit Level Models

Consider a population of N units with q-th small areas consisting of N_q units. Let y_{qj} and x_{qj} be the unit level y- value and correlated covariate x-value for j-th unit in the q-th small area. It is assumed that the domain means $\overline{X}_q$ is known. Consider the following one-folded nested error linear regression model

$$y_{qj} = x_{qj}^T \beta + \nu_q + e_{qj}, \quad j=1, \quad N_q ; \; q=1, \; \ldots \; ,Q$$

where the random small area effects ν_q have mean zero and common variance σ_ν^2 and are independently distributed. Also, e_{qj} are assumed to be independently distributed with mean zero and variance σ_e^2 and are also independent of area effects ν_q. This model was initially considered by Battese, Harter and Fuller (1988).

If N_q is large, the population mean $\bar{Y}_q$ is approximately equal to $x_q^T\beta + \nu_q$. The sample data $\{y_{qj}, x_{qj} \quad j = 1,\ldots,n_q;\ q = 1,\ldots,Q\}$ is assumed to satisfy the population model. This happens in equal probability sampling. This will also follow in probability proportional to size sampling when the size measure is taken as the covariate in the model. Assuming $\bar{Y}_q = \bar{X}_q^T\beta + \nu_q$, the EBLUP estimate of $\bar{Y}_q$ is of the form

$$\bar{y}_q^* = \hat{\gamma}_q[\bar{y}_q + (\bar{X}_q - \bar{x}_q)^T\hat{\beta}] + (1-\hat{\gamma}_q)\bar{X}_q^T\hat{\beta}, \qquad q = 1,...,Q$$

where $\hat{\gamma}_q = \hat{\sigma}_\nu^2 \big/ (\hat{\sigma}_\nu^2 + \hat{\sigma}_e^2 n_q^{-1})$ with estimated variance components $\hat{\sigma}_\nu^2$ and $\hat{\sigma}_e^2$, and $\hat{\beta}$ is the weighted least square estimate of β. It may be noted that the EBLUP estimator is a composite estimator combining the survey regression estimator with the regression synthetic estimator

Under the unit level model for the sample data, the leading term of MSE ($\bar{y}_q^*$) is given by $\gamma_q(\sigma_e^2 / n_q)$, which shows that EBLUP estimate can lead to large gains in efficiency over the survey regression estimate when γ_q is small. Battese, Harter and Fuller (1988) applied the nested error regression model to estimate area under corn and soybeans at county level in North-central Iowa using farm interview data in conjunction with LANDSAT satellite data.

For details of an exhaustive and thorough presentation of small area estimation an excellent reference is the book by Rao, J. N. K. (2003).

8. Concluding Remarks

Sample survey theory and its applications have come a long way in the past century. During this period, random sampling approach has been established as a dependable tool for collecting and analyzing the data. During this process, various conceptual issues have been raised and solutions obtained. The process has helped in refining the sampling tools and techniques. In the recent years, focus has been on improved techniques for analysis of data. The scale and volume of data generation and its availability has increased phenomenally due to technological improvements. The challenges for the role of sample surveys in this context are enormous. But the prospects of meeting such challenges looks bright. In the present talk, I have tried to share some of my experiences and observations relating to developments in sampling theory and practice.

I would like to conclude with my tributes and respect to “Das Sahib”.

References

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankhya*, **31**, 441-454.

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part I, In *Foundations of Statistical Inference, Holt, Rinehart and Winston, Toronto,* 203-242.

Battese, G.E., Harter, R.M. and Fulller, W.A. (1988). An error component model for production of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **88**, 28-36.

Bowley, A.L. (1906). Address to the economic science and statistics section of the british association for the advancement of science. *J. Roy. Statist. Soc.,* **69**, 548-557.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bull. Int. Statist. Inst.* **22**, Livre I.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Austral. Jour. Statist.*, 5-13

Brewer, K.R.W. and Muhammad, Hanif (1982). Sampling with unequal probabilities. *Lecture Notes in Statistics, Springer- Verlag,* New York, Heidelberg Berlin.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling, Wiley, New York.

Cochran, W.G. (1942). Sampling Theory when the sampling units are of unequal sizes. *J. Amer. Statist. Assoc.*, **37**, 199-212.

Cochran, W.G. (1953). *Sampling Techniques*, 2nd Edition (1963), 3rd Edition (1977), John Wiley and Sons, New York.

Constance, F. Citro and Graham Kalton, Editors (2000). *Small Area Estimates of School-Age Children in Poverty,* National Academy Press.

Des Raj (1954). On sampling with probabilities proportional to size. *Ganita*, **5**, 175-182.

Des Raj (1968). *Sampling Theory*, McGraw Hill Book Co., New York.

Dalenius, T. (1957). Sampling in Sweden, Almqvist & Wicksell, Stolkholm.

Dalenius, T. (1962). Recent Advances in sample survey theory and methods. *Ann. Math. Stat*. **33**, 325-349.

Deming, W.E. (1950). *Some Theory of Sampling*, Wiley, New York.

Fay, R. and Herriot (1979). Estimation of income from small places- an application of james-stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Ghosh, J.K., Maiti, P., Rao, T.J. and Sinha, B.K. (1999). Evolution of statistics in india. *Int. Statist. Review*, **67**, 13-34.

Godambe, V.P. (1955). A Unified Theory of sampling from finite populations. *J. Roy. Statist. Soc.*, B, **17**, 269-278.

Goodman, R. and Kish, L. (1950). Controlled Selection – a technique in probability sampling. *J. Amer. Statist. Assoc.*, **45**, 350-372.

Hansen, M.H. and Hurwitz, W.N. (1943). On the Theory of sampling from finite populations. *Ann. Math. Stat*. **14**, 333-362.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. John Wiley and Sons, New York. Vol. I and II.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann, Math. Stat*. **33**, 350-374.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.,* **47**, 663-685.

Hubback, J.A. (1927). Sampling for rice yields in bihar and orissa. *Imp. Agri. Res. Inst. Bull., Pusa, 166 (Reprinted in Sankhya (1946),***7**, 282-294).

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin No. 304.*

Kiaer, A.N. (1895-6). Observations of experiences concernant des denombrements representatifs. *Bull. Int. Stat. Inst.,* **9,** Livre 2, 176 183.

Kish, L. (1965). *Survey Sampling*, John Wiley and Sons, New York.

Madow, W.G. and Madow, L.H. (1944). On the theory of systematic sampling. *Ann. Math. Stat.*, **15**, 1-24.

Mahalanobis, P.C. (1938). A statistical report on the experimental crop census in 1937. *Indian Central Jute Committee*, 13-83

Mahalanobis, P.C. (1940). A sample survey of the acreage under jute in Bengal. *Sankhya*. **4**, 511-530.

Mahalanobis, P.C. (1943). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Statist. Soc., A,* **109**, 3425-378.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18**, 379-390.

Murthy, M.N. (1963). Some recent advances in sampling theory. *J. Amer. Statist. Assoc.*, **58**, 735-755.

Murthy, M.N. (1967). *Sampling Theory and Methods,* Statistical Publishing Society, Calcutta.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Jour. of Indian Soc. of Agril. Statist.*, **3**, 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.,* **97**, 555-606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.*, **33**, 101-116.

O'Muircheartaigh, C. and Wong, T. (1981): The impact of sampling theory on survey sampling practice. *Bull. Int. Statist. Inst.* **49**, Book 1, 465-493.

Panse, V.G. and Sukhatme, P.V. (1948) Crop Surveys in India-I. *Journal of Indian Society of Agricultural. Statistics.*, **1**, 34-58

Panse, V.G. and Sukhatme, P.V. (1951) Crop Surveys in India-II. *Journal of Indian Society of Agricultural. Statistics.*, **3**, 98-168.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *J. Roy. Statist. Soc.* B, **12**, 241-255.

Rao, J.N.K. (1999). Some current trends in sample survey theory and methods. *Sankhya*, Special issue on Sample Surveys. **61**, B, Part 1, 1-57.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

Rao, J.N.K. and Bellhouse, D.R. (1990). History and developments of the theoretical foundations of survey based estimation and analysis. *Survey Methodology,* **16**, 1-29.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). A simple procedure of unequal probability sampling without replacement, *J. Roy. Statist. Soc,* B**24**, 482-491.

Royall, R.M. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.*, **63**, 1269-1279.

Royall, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.

Royall, R. (1971). Linear regression models in finite population sampling theory. In V. P. Godambe and D. A Sprott, Eds. *Foundations of statistical inference*. Toronto: Holt, Rinchart and Winston. 259-274.

Royall, R.M. and Herson, J. (1973a). Robust estimation in finite populations – I. *J. Amer. Statist. Assoc.*, **68**, 880-889.

Royall, R.M. and Herson, J. (1973b). Robust estimation in finite populations – II. *J. Amer. Statist. Assoc.*, **68**, 890-893.

Sampford, M.R. (1967). On Sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Sarndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*, Springer Series in Statistics.

Smith, T.M.F. (1976). The foundations of survey sampling. *J. Roy. Statist. Soc. A*, 139, Part **2**, 183-204.

Srivastava, A.K. (2009). Some aspects of estimating poverty at small area levels. *Journal of Indian Society of Agricultural. Statistics.,* **63**(1), 1-23.

Sukhatme, P.V. (1953). *Introduction to Sampling Theory with Applications to Agriculture*, Oliver and Boyd, London

Sukhatme, P.V. (1959). Major developments in the theory and applications of sampling during the last twenty five years. *Estadistica*, **17**, 652-679.