Sankhyā : The Indian Journal of Statistics
Special issue in memory of D. Basu
2002, Volume 64, Series A, Pt. 3, pp 532-544

IN THE FOOTSTEPS OF BASU: THE PREDICTIVE MODELLING APPROACH TO SAMPLING FROM FINITE POPULATION

By SHELEMYAHU ZACKS
Binghamton University, New York, USA

SUMMARY. D.Basu's approach to (survey) sampling from finite populations is well known. Basu adhered to the likelihood principle and to the conditionality principle. He gradually became ardently Bayesian, and eventually rejected randomization in the data analysis stage. He accepted the predictive modelling approach as (empirical) Bayesian procedures. The present article reviews recent developments, specially in the model based approach, in light of Basu's philosophy.

1. Introduction

Much has already been written on Basu's approach to sampling from finite populations. In particular, see the review paper of Meeden (1992) and the collection of Basu's papers with illuminating discussions, which were edited by Ghosh (1988). The question is: What new can be said on the subject? The present paper highlights some modern approaches, which are compatible with Basu's philosophy on sampling from finite populations.

During the academic years 1968-1970 I was on the faculty of the Department of Mathematics and Statistics at the University of New Mexico. I was fortunate to associate closely with D.Basu, who was visiting the department during that period. We spent much time together, discussing the theory of sampling from finite populations. My paper on Bayesian sequential sampling from finite populations, Zacks (1969), was influenced by these discussions. At the time Basu was working on his fundamental and revolutionary papers, Basu (1969, 1971).

Paper received February 2002.
AMS (1991) *subject classification.* 62A15, 62D05.
Keywords and phrases. Finite populations; survey sampling; design based approach; model based approach; superpopulations; Bayesian analysis; robustness.

Basu had major difficulty to accept the prevalent approach of probability designs for sampling, which attempted to provide the framework for statistical inference by probability (random) sampling procedures, independent of the values in the population (the parameters). In particular, Basu could not accept estimates of population quantities, which incorporate functions of the sampling rule, like the inclusion probabilities. Such data analysis contradicted the *principle of conditionality* and the *likelihood principle*, to which he adhered. Thus, he rejected the celebrated Horvitz-Thompson estimator of the population total (see Des Raj, 1969, Hedayat and Sinha, 1991). Basu always illustrated his points by striking examples. To show that the Horvitz-Thompson estimator could lead to ridiculous estimators, Basu created the "elephants" example, see Basu (1971), in which the problem was to estimate the total weight of $N = 50$ elephants by weighting just one elephant. We will discuss this example in section 3, in the context of super-population models and also in section 4. Basu (1978) listed ten questions, which critical reviewers had raised on his approach, and answered each question in a way summarizing his philosophy.

The questions deal with the following problems:

1. Can we consider the values of the units in the population as parameters?

2. Is the predictive approach (see section 3) essentially different from the Bayesian one?

3. Can a probability design for choosing samples at random, together with known covariates, yield information on the unknown values in the population in addition to the information given by the sample values?

4. How can the likelihood principle be valid in the context of sampling survey?

5. How should we define the sample space?

6. Do we have to distinguish between different null sets?

7. What kind of probability interpretation should we adopt? Subjective? Model-Predictive? Randomization (Design-Based)?

8. How can we interpret the flatness of the likelihood function?

9. Should one consider an MLE in the sampling set-up?

10. How can one be Bayesian in a typical large scale survey? The dimension of the unknown parametric vector may run into hundreds of thousands. Can one be a fully fledged Bayesian in such situations?

Basu tried to answer the first nine questions in his theoretical formulation, which will be outlined in section 2. On the last question he wrote:

"The Bayesian as a surveyor must make all kind of compromises... He may even agree to introduce an element of randomization into his plan... I can not put this enormous speculative process into a jacket of a theory. I happen to believe that data analysis is more than a scientific method..." (see Ghosh, 1988, p.270).

In section 3 we will review the predictive approach, following the superpopulation models outlined in Bolfarine and Zacks (1992). In section 4 we will discuss the role of the design, within the framework of the predictive approach. We show that the optimality of a sample design (the one which minimizes the predictive risk) strongly depends on the predictive model. Randomization might play a role in *robust* designs. There are no new results in the present paper. It only reviews some modern approaches based on the foundations laid by Basu.

2. Basu's Theoretical Framework

Let $\mathrm{U} = (u_1, u_2, \ldots, u_N)$ be a list of a finite number, N, of units ordered in some fashion. The list U, and its length N are known. A sample is a collection of units from U. We will consider only samples of distinct units from U with indices $i_1, i_2, \ldots$ (sampling without replacement). Let $S = (i_1, i_2, \ldots, i_{n(S)})$ be the ordered set of indices of units in the sample, i.e., $1 \leq i_1 < i_2 < \ldots < i_{n(S)}$, where $n(S) = |S|$ is the sample size. We will assume that all the $n(S)!$ permutations of the indices in S represent equivalent samples (this is the case when the order of sampling is immaterial).

Let $y_1, \ldots, y_N$ be values of interest (real or vector valued) of the units in U. For simplicity, we will consider real y-values. The values $y_1, \ldots, y_N$ are not entirely known. Only the values of the observed units in the sample S are known. All other values of y are unknown. Basu considered the y values as *parameters*, and the vector $Y = (y_1, \ldots, y_N)'$ as a *point* in a parameter space Θ. The *information* in the sample S on the parameter point θ is given by the statistic $Z = (S, y_S)$, where $y_S = (y_{i_1}, y_{i_2}, \ldots, y_{i_{n(S)}})$ is the vector of y-values of the units in S. It is assumed that observations on y are without errors. Generally, in statistical theory a parameter θ is not directly observable. In sampling surveys, if $S = \mathrm{U}$ (complete census) then θ is completely observable. If the sample size $n(S)$ is smaller than N, θ is

only partially observable. Let $\mathrm{S}_n = \{(i_1, \ldots, i_n) : 1 \leq i_1 < \ldots < i_n \leq N\}$ be the sample space of all (reduced) samples of fixed size $n,\ n = 1, \ldots, N$. On $\mathrm{S}_n \times \Theta$ define the function $Y_n(S, \theta) = (y_{i_1}, \ldots, y_{i_n})$, i.e., $Y_n : S_n \times \Theta \to R^n$. For $n = 1, \ldots, N$, let $Z_n = S_n \times R^n$. The sample space is thus $Z = \cup_{n=1}^{N} Z_n$. For a Bayesian framework, let B be the Borel σ-field on Θ, and (Θ, B, ξ) a probability space. ξ is a prior probability measure on (Θ, B). The Bayesian model assumes that the corresponding prior distribution on Θ belongs to a family of prior distributions, H.

One needs now a likelihood function to serve as a kernel of the transformation from the prior distribution $H(\theta)$ to the posterior distribution $H(\theta \,|\, Z)$. Basu considered the normalized function $L(\theta, Z)$ which assumes a constant value, 1, for all θ compatible with the statistic Z. More formally, for $Z \in \mathrm{Z}_n,\ n = 1, \ldots, N$, let $Z = \{(i_1, \ldots, i_n), (y_1, \ldots, y_n)\}$ and

$$\Theta_Z^n \quad = \quad \{\theta : \theta \in \Theta, \theta_{i_j} = y_j \forall j = 1, \ldots n\} \tag{2.1}$$

Thus, for $Z \in \mathrm{Z}_n$

$$L(\theta; Z) \quad = \quad \begin{cases} 1\ , \text{ if } \theta \in \Theta_Z^n \\ 0\ , \text{ otherwise} \end{cases} \tag{2.2}$$

Assume that the family of prior distributions H is dominated, and let $h(\theta_1, \ldots, \theta_N)$ be a prior density of an element of H. For a given $S = (i_1, \ldots, i_n)$, let $g_S(y_1, \ldots, y_n)$ be the *prior marginal* density of the component of θ with indices in S.

Thus, given a sample statistic $Z = (S, y_S)$, with $y_S = (y_1, \ldots, y_n)$, the *posterior density* of θ given Z is

$$h(\theta \,|\, (S, y_S)) \quad = \quad \frac{h(\theta_1, \ldots, \theta_N) L(\theta; (S, y_S))}{g_S(y_1, \ldots, y_n)} \tag{2.3}$$

Notice that $h(\theta \,|\, (S, y_S)) = 0$ if $\theta \notin \Theta_Z^n$. Bayesian analysis can proceed, using this posterior density.

This definition of posterior density was used extensively by Zacks (1969), in the study of Bayes sequential sampling from finite populations. See also Ghosh and Meeden (1977, p.11) for an example of exchangeable posterior density of θ given Z. The posterior density (2.3) does not depend on the sampling rule. Thus, the Bayesian inference does not depend on the sampling design. However, as shown in Zacks (1969), the posterior risk depends on $Z = (S, y_S)$. The expected posterior risk, depends on S and on the prior distribution. Thus, one can discuss *optimal* designs from a Bayesian point of view.

Suppose we wish to estimate the population total $T = \theta' 1_N = \sum_{i=1}^{N} y_i$. Given $Z = (S, y_S)$ the Bayesian estimator of T, for the squared error loss function is

$$\hat{T}_B = E\left\{\sum_{i=1}^{N} y_i \,|\, (S, y_S)\right\} \tag{2.4}$$

$$= \sum_{i \in S} y_i + \sum_{i \notin S} E\left\{y_i \,|\, (S, y_S)\right\} \tag{2.5}$$

The Bayesian framework provides the tools to estimate (predict) functions of the unobserved components of θ, using the values of the observed sample (S, y_S).

We conclude this section with the remark that the above purely Bayesian approach was developed by Basu gradually. In Basu (1971) he had still used the (randomized) designs framework to define probability distributions of Z given θ. But in the same article he argued that, in the stage of data analysis, the randomization probability distribution $p(S)$ on S should not play any role, since in non-sequential procedures $p(S)$ does not depend on θ.

In Basu (1978) he turned completely to the Bayesian approach. He wrote there (see Ghosh, 1988, p.269):

"I strongly suspect that the Bayesian guideline in terms of subjective probabilities (models of prior distributions) is more reliable than the objective randomization probabilities."

3. The Model Based Approach: Predictive Modelling

The model-based approach considers the values $y_1, \ldots, y_N$ in the population as a realization of some random process. This is like assuming that there is a hypothetical infinite superpopulation of random θ vectors. The modelling of this superpopulation requires to formulate the joint distribution of the θ vectors. This is similar to the usual practice in other areas of statistical inference, Bayesian or non-Bayesian. The model is generally a family of distributions, parametric or non-parametric, pertaining to the observed results of the "experiment" under consideration. The inference is generally on some characteristic (functional) of the true distribution. In the superpopulation model the complete N-dimensional vector θ is realized according to one of the distributions in the assumed family. From the observed values $Z = (S, y_S)$, inference is done on the distribution (or its parameters) of θ, and the value of some function of the unobserved part of θ is *predicted*.

As an example, for the population total T, after observing Z we have to predict the quantity $\sum_{i \notin S} y_i$. For reading on superpopulation models see Thompson (1988).

Although used sporadically in various articles for over seventy years, the model-based approach received a big boost with the celebrated papers of R.M. Royall (1970, 1971, 1976). The reader is referred to Cassel et al. (1977), and Bolfarine and Zacks (1992). A few models and basic tools are discussed below. Before continuing, however, it is interesting to quote Basu's opinion on this approach. This is taken from Basu's response to Royall's discussion on the (1971) paper (see Ghosh, 1988, p.243).

"To me such a (superpopulation) model looks exactly like a Bayesian formalization of the surveyor's background knowledge and information. Certainly there is nothing objective about the model. Indeed, is any probability model objective?"

It is often the case that information is available on all the N units in the population, which is related to the variable of interest, Y. Let $x_{i,1}, \ldots, x_{i,k}$ be k such values of units i, which are covariates of y_i $i = 1, \ldots, N$. Let X be the $N \times k$ matrix of these covariates. If we denote here by y the N-dimensional population vector, one can consider the *linear model*

$$y = X\beta + e \tag{3.1}$$

where e is a random vector with $E\{e\} = 0$ and covariance matrix V. We designate by $\psi = (\beta, V)$ the super parameters of the model. Various cases of the linear model (3.1) are studied in the literature. These are summarized in Bolfarine and Zacks (1992, p.9), as models SM1–SM6. We present later in more details only models SM1 and SM2, to illustrate certain properties.

Partition y to y_S and y_R, where y_R consists of all components of y which are not in the observed sample vector y_S. Let n be the size of the sample S. The dimension of y_R is $N - n$. Since the indices of the components of y which are in y_S are specified in S, we will assume (without loss of generality) that $y' = (y'_S, y'_R)$. We partition X and V respectively to

$$X = \begin{bmatrix} X_S \\ S_R \end{bmatrix} \quad \text{and}$$

$$V = \begin{bmatrix} V_S & V_{SR} \\ V_{RS} & V_R \end{bmatrix}$$

The problem is to predict y_R, given the observed sample y_S and the information in the model. From a purely Bayesian point of view the model (3.1) is incomplete. One has to specify the actual multivariate distribution

of y. If one assumes the normal multivariate model, i.e. $y \sim N(X\beta, V)$, with *known* parameter $\psi = (\beta, V)$, then the prediction of y_R, is the linear predictor

$$E\{y_R \,|\, y_S\} \quad = \quad X_R\beta + V_{RS}V_S^{-1}(y - X\beta) \tag{3.2}$$

If β is *unknown*, V is *known* and the normal model applied, one could use a hierarchical Bayesian approach, by assigning a hyper-prior distribution for β. In Bolfarine and Zacks (1992, ch.3) Bayesian predictors of the population total T, and the population variance $S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$, where $\bar{y}_N = T/N$, are derived.

Let $\omega(y)$ be a function of y. A *predictor* of $\omega(y)$ is a statistic $\hat{\omega}(S, y_S)$. In the set-up of the present section we can write $\hat{\omega}(y_S)$. A predictor $\hat{\omega}(y_S)$ is called model unbiased if

$$E_\psi\{\hat{\omega}(y_S) - \omega(y)\} = 0 \quad , \forall\ \psi \tag{3.3}$$

The expectation in (3.3) is according to the specified model.

The prediction model mean-squared-error (MSE) of $\hat{\omega}$ is

$$\text{MSE}_{\,\psi}\{\hat{\omega}\} \quad = \quad E_\psi\{[\hat{\omega}(y_S) - \omega(y)]^2\} \tag{3.4}$$

The following linear predictor of the population total, T, is best unbiased in terms of minimum prediction variance for the linear model ψ (see Bolfarine and Zacks, 1992, p.25).

$$\hat{T}_{BLU} \quad = \quad 1_S'y_S + 1_R'[X_R\hat{\beta}_S + V_{RS}V_S^{-1}(y_S - X_S\hat{\beta}_S)] \tag{3.5}$$

where 1_S is an n-dimensional vector of 1's, 1_R is an $N-n$-dimensional vector of 1's, and $\hat{\beta}_S$ is the (weighted) least squares estimator of β, given by

$$\hat{\beta}_S \quad = \quad (X_S'V_S^{-1}X_S)^{-1}X_S'V_S^{-1}y_S. \tag{3.6}$$

The prediction variance of $\hat{T}_{BLU}$ is given in Bolfarine and Zacks (1992, p.25). To derive the BLU predictor, $\hat{T}_{BLU}$, one does not have to specify the multivariate distribution of y. This is a great advantage. In models discussed below the covariance matrix V is specified as a function of the covariates matrix X.

For the sake of illustration, we consider the first two simple models, which are special cases of the linear model (3.1).

3.1 *Model SM1*

$$y \;=\; \beta 1 + e \;\;,\;\; V \;=\; \sigma^2\, I. \tag{3.7}$$

If a sample S was drawn from this population, and $\bar{y}_S = 1'_S y_S/n$ is the sample mean, then the best unbiased linear predictor of T is

$$\hat{T}_1 \;=\; N\bar{Y}_S. \tag{3.8}$$

This is known as the "expansion" estimator (see Cochran, 1977). The prediction variance of this predictor, according to SM1, is

$$E_1\{(\hat{T}_1 - T)^2\} \;=\; \sigma^2 \frac{N(N-n)}{n}. \tag{3.9}$$

This prediction variance is independent of S. Thus, under SM1 any sample is optimal for estimating T.

3.2 *Model SM2*

$$y \;=\; \beta x + e \tag{3.10}$$

with

$$V \;=\; \sigma^2 \mathrm{diag}\,\{x_i^g, i = 1, \ldots, N\}\;,\; g = 0, 1, 2.$$

and $x = (x_1, \ldots, x_N)'$, $x_i > 0$, $i = 1, \ldots N$. diag $\{\ \}$ designates a diagonal matrix, with x_i^g as diagonal elements. The following predictors $\hat{T}_{BLU}$ are obtained:

1. $g = 0$

$$\hat{T}_{2,0} \;=\; \sum_{i\in S} y_i + \frac{\sum_{i\in S} y_i x_i}{\sum_{i\in S} x_i^2} T_R(x) \tag{3.11}$$

where

$$T_R(x) \;=\; \sum_{i\in R} x_i \tag{3.12}$$

2. $g = 1$

$$\begin{aligned} \hat{T}_{2,1} \;&=\; \sum_{i\in S} y_i + \frac{\sum_{i\in S} y_i}{\sum_{i\in S} x_i} T_R(x) \\ &=\; \frac{\bar{y}_S}{\bar{x}_S} \left(\sum_{i=1}^{N} x_i \right) \end{aligned} \tag{3.13}$$

where $\bar{y}_S$ and $\bar{x}_S$ are the sample averages of the y and x values, respectively.

3. $g = 2$

$$\hat{T}_{2,2} = \sum_{i \in S} y_i + \frac{T_R(x)}{n} \sum_{i \in S} \frac{y_i}{x_i}. \tag{3.14}$$

The prediction variances of these predictors, for the cases of g, are:

1. $g = 0$

$$E\{(\hat{T}_{2,0} - T)^2\} = \sigma^2 \left[N - n + \frac{T_R^2(x)}{\sum_{i \in S} x_i^2} \right]. \tag{3.15}$$

2. $g = 1$

$$E\{(\hat{T}_{2,1} - T)^2\} = \sigma^2 \frac{T_R(x)}{\sum_{i \in S} x_i} \sum_{i=1}^{N} x_i. \tag{3.16}$$

3. $g = 2$

$$E\{(\hat{T}_{2,2} - T)^2\} = \sigma^2 \left[\sum_{j \in R} x_j^2 + \frac{\bar{x}_s}{n} T_R^2(x) \right]. \tag{3.17}$$

A choice of sample S under model SM2 is called *optimal* if it minimizes the prediction variance. From (3.15)-(3.16) it is clear that the optimal sample under SM2, for predicting T when $g = 0$ or $g = 1$, is the one with units having the *largest* x-values. This might be true also when $g = 2$. Notice that if $n = 1$ then $\hat{T}_{2,0} = \hat{T}_{2,1} = \hat{T}_{2,2}$.

In Basu's elephants example (see Basu, 1971), let $y_1, \ldots, y_{50}$ be the present weights of the elephants. Let $x_1, \ldots, x_{50}$ be their past weights (known). If at present we can weigh only one elephant then according to model SM2 ($g = 0, 1$) the optimal procedure is to weigh the largest elephant, and to predict the total weight of the herd by $\hat{T} = R \sum_{i=1}^{50} x_i$, where $R = \bar{y}_S / \bar{x}_S$. Basu implied that the acceptable procedure is to weight the "average" size elephant and use $\hat{T}_1 = \bar{y}_S N$. It is hard to believe that Basu considered model SM1 to be correct. However, as will be shown in the next section, Basu's procedure is *robust* if one uses the predictor $\hat{T}_1$ under model SM2 ($g = 1$).

4. Robust Design for Predictive Models

We have seen in the previous section that optimal sampling designs for predictive models depend strongly on the model, and whether the model assumed by the statistician is correct. For each model there is an optimal predictor and an optimal design, in terms of the predictor risk. For linear models of the type given by (3.1), the question is, what happens if the model considered does not include some important covariates. In other words, suppose that the *assumed* model is

$$y \;=\; X\beta + e \quad , \quad \psi \;=\; (\beta, V) \tag{4.1}$$

but the *correct* model is

$$y \;=\; \left[\begin{array}{cc} X & X^* \end{array}\right] \left[\begin{array}{c} \beta \\ \beta^* \end{array}\right] + e \quad , \quad \psi \;=\; \left(\left[\begin{array}{c} \beta \\ \beta^* \end{array}\right] , V^*\right). \tag{4.2}$$

This is a well known problem in statistics, of model misspecification. The statistician often sacrifices optimality for *robustness*. We illustrate the problem with the simple models SM1 and SM2, discussed in the previous section. If the statistician adopts the model SM1, his BLU predictor of T is $\hat{T}_1$ given by (3.8), with prediction risk (3.9), which is independent of S. Thus any strategy $(S, \hat{T}_1)$ is optimal if the model SM1 is the true one. Suppose that the statistician adopts SM1, but the correct model is SM2, with $g = 1$. The statistician predicts T with $\hat{T}_1$, while the optimal predictor is $\hat{T}_{2,1}$ (3.13). The covariate x is not taken into account.

The prediction MSE of $\hat{T}_1$ under SM2 with $g = 1$ is

$$\begin{aligned} E_{\psi_{2,S}}\{(\hat{T}_1 - T)^2\} \;&=\; \\ \sigma^2(N-n)\left[\bar{x}_R + \frac{N-n}{n}\bar{x}_S\right] &+ \beta^2(N-n)^2(\bar{x}_R - \bar{x}_S)^2 \end{aligned} \tag{4.3}$$

where $\bar{x}_R = T_R(x)/(N-n)$ is the mean of the x values of the units not in S. $\bar{x}_S$ is the mean of the x values of the units in S. If SM2 is correct $\beta^2 > 0$. If the statistician chooses a sample S^0, for which

$$\bar{x}_S \;=\; \bar{x}_R \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \;=\; \bar{x}_N$$

then (4.3) reduces to

$$E_{\psi_{2,S^0}}\{(\hat{T}_1 - T)^2\} \;=\; \sigma^2 \frac{(N-n)N}{n}\bar{x}_N. \tag{4.4}$$

Since the prediction risk of $(S^0, \hat{T}_1)$ is, under SM2, independent of β, the strategy $(S^0, \hat{T}_1)$ is a *robust minimax* strategy against model misspecification. A sample S^0 such that $\bar{x}_{S^0} = \bar{x}_N$ is called a *balanced* sample.

Let $\hat{T}_{BLU}(\psi)$ and $\hat{T}_{BLU}(\psi^*)$ be the BLU predictors of T under the two models. Algebraic conditions, in terms of the matrices X, X^*, V and V^* for the predictors $\hat{T}_{BLU}(\psi)$ and $\hat{T}_{BLU}(\psi^*)$ to be equal are given in chapter 6 of Bolfarine and Zacks (1992). As shown there, robustness requires choice of balanced sample, with respect to the various covariates. This is generally difficult to achieve in non-trivial populations. One could attempt to get a sample having mean $\bar{x}_S$ close to the population mean $\bar{x}_N$ by random sampling. Such a combination of the design approach and the modelling approach was studied by Cassel et al. (1977), Godambe (1982), Brewer (1979), Tain (1988), Wright (1983), Rodriguez and Bolfarine (1989). In particular see Bolfarine and Zacks (1992, ch. 9). Notice that the expected predictive MSE of $\hat{T}_1$ under simple random sampling (SRS) without replacement of fixed size n, under SM2, is

$$E_P\, E_{\psi_{2,S}}\{(\hat{T}_1 - T)^2\} = \sigma^2 \frac{(N-n)N}{n}\left(\bar{x}_N + \beta^2 S_N^2(x)(1 - \frac{1}{N})\right), \qquad (4.5)$$

where $S_N^2(x)$ is the population variance of the x-values. The effect of β^2 on the prediction MSE is not cancelled by taking a random sample. However, with respect to bias, under SRS, $\hat{T}_1$ becomes a design-model-unbiased predictor of T. Indeed,

$$E_P\, E_{\psi_{2,S}}\{(\hat{T}_1 - T)\} \;=\; (N-n)\beta E_P\{(\bar{x}_R - \bar{x}_S)\} \;=\; 0. \qquad (4.6)$$

Godambe (1982) considered the problem of choosing a predictor $\hat{T}$ and a sampling design (S, P) such that the design-model expected prediction risk is minimized. Here P is a probability function on the sample space S. Godambe restricted attention to the class of probability functions for which the inclusion probabilities

$$\pi_i \;=\; \sum_S I_i(S)P(S) > 0 \;\;,\; i = 1, \ldots, N.$$

Here $I_i(S) = 1$ if $i \in S$ and $I_i(S) = 0$ otherwise. Godambe studied the family of predictive models, in which

$$y_i \;=\; \alpha_i + \beta x_i + e_i \;\;,\; i = 1, \ldots, N,$$

where $x = (x_1, \ldots, x_N)'$ is a vector of *known* covariates, $E(e_i) = 0$, $\forall\ i = 1, \ldots, N$ and $V\{e_i\} = \sigma^2 v_i$, $Cov(e_i, e_j) = 0$, for all $i \neq j$, and $v_i = f(x_i)$.

Let $\psi = (\alpha, \beta, \sigma^2)$, $\alpha = (\alpha_1, \ldots, \alpha_N)'$. Godambe (1982) proved that for procedures of fixed sample size n, if P is such that $\pi_i = n\sqrt{v_i}/\sum_{i=1}^N \sqrt{v_i}$, and if $\hat{T} = \sum_{i\in S}(y_i/\pi_i)$ (the Horvitz-Thompson estimator, HTE) then $E_P\, E_\psi\{(\hat{T} - T)^2\}$ is minimized. This "optimal" strategy is independent of (possibly unknown parameters) ψ and was therefore labelled a *robust* procedure. We remark first, that this type of robustness is different from the robustness against model misspecification. Second, the class of strategies considered by Godambe *excludes* purposive sampling, in which a specified sample S^0 is chosen with $P(S^0) = 1$ and $P(S) = 0$ for all S which are not equivalent to S^0. In such purposive samples, $\pi_i = 1$ for $i \in S^0$ and $\pi_i = 0$ for all $i \notin S^0$.

Consider model SM2 with $g = 1$. This is a special case of Godambe's model, in which $\alpha = 0$, β and σ^2 are unknown. As we have shown earlier, the optimal strategy is $(P^0, \hat{T}_{2,1})$, where $P^0(S^0) = 1$ if S^0 is the sample with the n largest x-values. Indeed, $\hat{T}_{2,1}$ is BLU predictor, and therefore

$$\begin{aligned} E_{P^0}\, E_\psi\{(\hat{T}_{2,1} - T)^2\} &= \inf_{S\in S_n} \sigma^2 \frac{N(N-n)}{n} \frac{\bar{x}_R \bar{x}_N}{\bar{x}_S} \\ = \inf_P E_P\, E_\psi\{(\hat{T}_{2,1} - T)^2\} &= \inf_P \inf_{T\in C_{LU}} E_P\, E_\psi\{(\hat{T}_{2,1} - T)^2\} \end{aligned}$$

where C_{LU} is the class of linear unbiased predictors. If the surveyor *knows* that SM2 is true, the optimal sampling is a *purposive* one. This is in line with Basu's philosophy.

References

Basu, D. (1969). Role of sufficiency and likelihood principles in survey theory, *Sankhyā, Ser. A*, **31**, 441-453.

Basu, D. (1971). An essay on the logical foundations of survey sampling, in *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, eds., Holt, Rinehart and Winston, Toronto.

Basu, D. (1978). On the relevance of randomization in data analysis, in *Survey sampling and Measurement*, N.K. Namboodiri, ed., Academic Press, New York, 267-339.

Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*, Springer Series in Statistics, Springer-Verlag, New York.

Brewer, K.R.W. (1979). A class of robust sampling designs for large scale surveys, *J. Amer. Statist. Assoc.*, **74**, 911-915.

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

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, Wiley, New York.

Des Raj (1972). *The Design of Sampling Surveys*, McGraw-Hill, New York.

Ghosh, J.K. (1988). Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu, *Lecture Notes in Statistics*, **45**, Springer-Verlag, New York.

Ghosh M. and Meeden G. (1997). *Bayesian Methods for Finite Populations Sampling*, Chapman and Hall, New York.

Godambe V.P. (1982). Estimation in survey sampling: robustness and optimality, *J. Amer. Statist. Assoc.*, **77**, 393-406.

Hedayat A.S. and Sinha B.K. (1991). *Design and Inference in Finite Population Sampling*, Wiley, New York.

Meeden G. (1992). Basu's contributions to the foundations of sample survey. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, M. Ghosh and P.K. Pathak, eds., Lecture Notes- Monograph Series, Institute of Mathematical Statistics, **17**, Hayward, CA, 178-186.

Rodriguez J. and Bolfarine H. (1989). *A note on asymptotically design unbiased designs in survey sampling*, Technical Report, University of Sao Paulo, Brazil.

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

Royall R.M. (1971). Linear regression models in finite populations sampling theory. In *Foundations of Statistical Inference*, V.P. Godambe and D. A. Sprott, Eds. Holt, Rinehart and Winston, Toronto.

Royall, R.M. (1976). The linear least squares prediction approach to two stage sampling, *J. Amer. Statist. Assoc.*, **71**, 657-664.

Tallis, G.M. (1978). Note on robust estimation in finite populations, *Sankhyā, Ser. C*, **40**, 136-138.

Tam, S.M. (1988). Asymptotically design-unbiased predictors in survey sampling, *Biometrika*, **75**, 175-177.

Thompson, M.E. (1988). Superpopulations models. In *Encyclopedia of Statistical Sciences*, Kotz, S. and Johnson, N.L. **9**, 93-98.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information, *J. Amer. Statist. Assoc.*, **74**, 879-884.

Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations, *J. Amer. Statist. Assoc.*, **64**, 1342-1349.

Shelemyahu Zacks

Department of Mathematical Sciences

Binghamton University, NY 13902-6000

USA

E-mail: shelly@math.binghamton.edu.